DOCUMENT RESUME

ED 385 244 IR 017 257

TITLE	From Desktop to Teraflop: Exploiting the U.S. Lead in High Performance Computing. NSF Blue Ribbon Panel on High Performance Computing.
INSTITUTION	National Science Foundation, Washington, D.C.
REPORT NO	NSB-93-205
PUB DATE	Aug 93
NOTE	65p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Computers; *Computer Science; Economic Progress; *Engineering; Futures (of Society); Investment; Research and Development; *Technological Advancement
IDENTIFIERS	Barriers to Change; *High Performance Computing; National Science Board; National Science Foundation; Supercomputers

ABSTRACT

This report addresses an opportunity to accelerate progress in virtually every branch of science and engineering concurrently, while also boosting the American economy as business firms also learn to exploit these new capabilities. The successful rapid advancement in both science and technology creates its own challenges, four of which are outlined here for the National Science Board. Four sets of interdependent recommendations are made in response to the challenges. The first implements a balanced pyramid of computing environments. Each element in the pyramid supports the others; whatever resources are applied to the whole, the balance in the pyramid should be sustained. The second set addresses the essential research investments and other steps to remove the obstacles to realizing the technologies in the pyramid and the barriers to the effective use of these environments. The third set addresses the institutional structure for delivery of the HPC capabilities, and consists itself of a pyramid. At the base of the institutional pyramid is the diverse array of investigators in their universities and other settings, who use all the facilities at all levels of the pyramid, followed by departments and research groups devoted to specific areas of computer science and engineering, and the National Science Foundation (NSF) high performance computing (HPC) Centers. At the apex is the national teraflop-class society, which is recommended as a multi-agency facility pushing the frontiers of high performance into the next decade. A final recommendation addresses the NSF role at the national level and its relationship with the states in HPC. Concepts are illustrated with two figures and two tables. Appendices include: a list of the membership of the Blue Ribbon Panel on High Performance Computing; information on the history and origin of this study on the NSF and HPC; a discussion of technology trends and barriers to further progress; four figures illustrating supercomputer data; and a review and prospectus of computational and computer science and engineering with personal statements by panel members. (MAS)

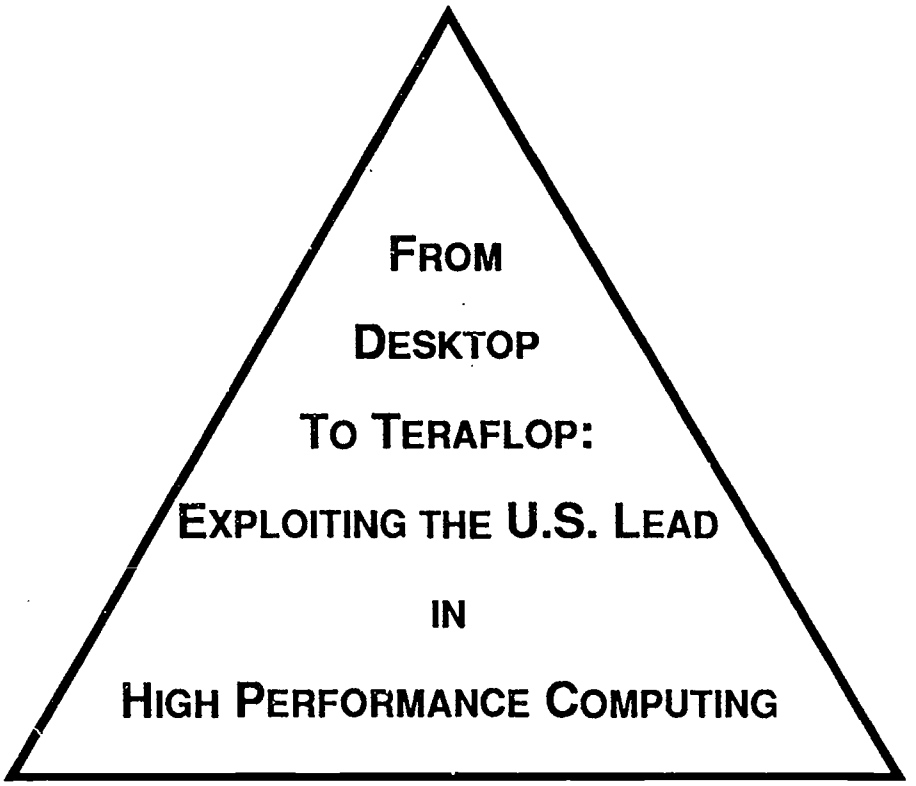
FROM DESKTOP TO TERAFLOP: EXPLOITING THE U.S. LEAD IN HIGH PERFORMANCE COMPUTING

NSF Blue Ribbon Panel on
High Performance Computing

August 1993

Dedication

This report is dedicated to one of the nation's most distinguished computer scientists, a builder of important academic institutions, and a devoted and effective public servant; Professor Nico Habermann. Dr. Habermann took responsibility in organizing this Panel's work and saw it through to completion, but passed away just a few days before it was presented to the National Science Board. The members of the panel deeply feel the loss of his creativity, wisdom, and friendship.



NSF Blue Ribbon Panel on High Performance Computing

August 1993

Lewis Branscomb (Chairman)
Theodore Belytschko
Peter Bridenbaugh
Teresa Chay
Jeff Dozier
Gary S. Grest
Edward F. Hayes
Barry Honig
Neal Lane (resigned from Panel, July 1993)
William Lester, Jr.
Gregory J. McRae
James A. Sethian
Burton Smith
Mary Vernon

"It is easier to invent the future than to predict it" - Alan Kay

EXECUTIVE SUMMARY

An Introductory Remark: Many reports are prepared for the National Science Board and the National Science Foundation that make an eloquent case for more resources for one discipline or another. This is not such a report. This report addresses an opportunity to accelerate progress in virtually every branch of science and engineering concurrently, while also giving a shot in the arm to the entire American economy as business firms also learn to exploit these new capabilities. The way much of science and engineering are practiced will be transformed, if our recommendations are implemented.

The National Science Board can take pride in the Foundation's accomplishments in the decade since it implemented the recommendations of the Peter Lax Report on high performance computing (HPC). The Foundation's High Performance Computing Centers continue to play a central role in this successful strategy, creating an enthusiastic and demanding set of sophisticated users, who have acquired the specialized computational skills required to use the fast advancing but still immature high performance computing technology.

Stimulated by this growing user community, the HPC industry finds itself in a state of excitement and transition. The very success of the NSF program, together with those of sister agencies, has given rise to a growing variety of new experimental computing environments, from massively parallel systems to networks of coupled workstations, that could, with the right research investments, produce entirely new levels of computing power, economy, and usability. The U.S. enjoys a substantial lead in computational science and in the emerging technology; it is urgent that the NSF capitalize on this lead, which not only offers scientific preeminence but also the industrial lead in a growing world market.

The vision of the rapid advances in both science and technology that the new generation of supercomputers could make possible has been shown to be realistic. This very success, measured in terms of new discoveries, the thousands of researchers and engineers who have gained experience in HPC, and the extraordinary technical progress in realizing new computing environments, creates its own challenges. We invite the Board to consider four such challenges:

Challenge 1: How can NSF, as the nation's premier agency funding basic research, remove existing barriers to the rapid evolution of high performance computing, making it truly usable by all the nation's scientists and engineers? These barriers are of two kinds: technological barriers (primarily to realizing the promise of highly parallel machines, workstations, and networks) and implementation barriers (new mathematical methods and new ways to formulate science and engineering problems for efficient and effective computation). An aggressive commitment by NSF to leadership in research and prototype development, in both computer science and in computational science, will be required.

Challenge 2: How can NSF provide scalable access to a pyramid of computing resources, from the high performance workstations needed by most scientists to the critically needed teraflop-and-beyond capability required for solving Grand Challenge problems? What balance of among high performance desktop workstations, vs. mid-range or mini-supercomputer, vs. networks of workstations, vs. remote, shared supercomputers of very high performance should NSF anticipate and encourage?

Challenge 3: The third challenge is to encourage the continued broadening of the base of participation in HPC, both in terms of institutions and in terms of skill levels and disciplines. This calls for expanded education and training, and participation by state-based and other HPC institutions.

Challenge 4: How can NSF best create the intellectual and management leadership for the future of high performance computing in the U.S.? What role should NSF play within the scope of the nationally coordinated HPCC program? What

relationships should NSF's activities in HPC have to the activities of other federal agencies?

This report recommends significant expansion in NSF investments, both in accelerating progress in high performance computing through computer and computational science research and in providing the balanced pyramid of computing facilities to the science and engineering communities. The cost estimates are only approximate, but in total they do not exceed the Administration's stated intent to double the investments in HPCC during the next 5 years. We believe these investments are not only justified but are compatible with stated national plans, both in absolute amount and in their distribution.

RECOMMENDATIONS:

We have four sets of interdependent recommendations. The first implements a balanced pyramid of computing environments (see Figure A following this Summary). Each element in the pyramid supports the others; whatever resources are applied to the whole, the balance in the pyramid should be sustained. The second set addresses the essential research investments and other steps to remove the obstacles to realizing the technologies in the pyramid and the barriers to the effective use of these environments.

The third set addresses the institutional structure for delivery of HPC capabilities, and consists itself of a pyramid (see Figure B following this Summary), of which the NSF Centers are an important part. At the base of the institutional pyramid is the diverse array of investigators in their universities and other settings, who use all the facilities at all levels of the pyramid. At the next level are departments and research groups devoted to specific areas of computer science or computational science and engineering. At the next level are the NSF HPC Centers, which must continue to be providers of shared high capability computing systems and to provide aggregations of specialized capability for all aspects of use and advance of high performance computing. At the apex is the national teraflop-class facility, which we recommend as a multi-agency facility pushing the frontiers of high performance into the next decade.

A final recommendation addresses the NSF's role at the national level and its relationship with the states in HPC.

A. CENTRAL GOAL FOR NSF HPC POLICY

Recommendation A-1: The National Science Board should take the lead, under OSTP guidance and in collaboration with ARPA, DoE and other agencies, to expand access to all levels of the dynamically evolving pyramid of high performance computing capability for all sectors of the whole nation. The realization of this pyramid depends, of course, on rapid progress in the pyramid's technologies. The computational capability we envision includes not only the research capability for which NSF has special stewardship, but also includes a rapid expansion of capability in business and industry to use HPC profitably, and many operational uses of HPC in commercial and military activities.

VISION OF THE HPC PYRAMID

Recommendation A-2: At the apex of the pyramid is the need for a national capability at the highest level of computing power the industry can support with both efficient software and hardware. A reasonable goal would be the design, development, and realization of a national teraflop-class capability, subject to the successful development of software and computational tools for such a large machine (recommendation B-1). NSF should initiate, through OSTP, an interagency plan to make this investment, anticipating multi-agency funding and usage.

Recommendation A-3: Over a period of 5 years the research universities should be assisted to acquire mid-range machines. These mid-sized machines are the underfunded element of the pyramid today — about 10% of NSF's FY92 HPC budget is devoted to their acquisition. They are needed for both demanding science and engineering problems that do not require the very maxi-

mum in computing capacity, and for use by the computer science and computational mathematics community in addressing the architectural, software, and algorithmic issues that are the primary barriers to progress with massively parallel processor architectures.

Recommendation A-4: We recommend that NSF double the current annual level of investment ($22 million) providing scientific and engineering workstations to its 20,000 principal investigators. Within 4 or 5 years workstations delivering up to 400 megaflops costing no more than $15,000 to $20,000 should be widely available. For education and a large fraction of the computational needs of science and engineering, these facilities will be adequate.

Recommendation A-5: We recommend that the NSF expand its New Technologies program to support expanded testing of the new parallel configurations for HPC applications. For example, the use of Gigabit local area networks to link workstations may meet a significant segment of mid-range HPC science and engineering applications. A significant supplement to HPC applications research capacity can be had with minimal additional cost if such collections of workstations prove practical and efficient.

B. RECOMMENDATIONS TO IMPLEMENT THESE GOALS

REMOVING BARRIERS TO HPC TECHNICAL PROGRESS AND HPC USAGE

Recommendation B-1: To accelerate progress in developing the HPC technology needed by users, NSF should create, in the Directorate for Computer and Information Science and Engineering, a challenge program in computer science with grant size and equipment access sufficient to support the systems and algorithm research needed for more rapid progress in HPC capability. The Centers, in collaboration with hardware and software vendors, can provide test platforms for much of this work, and recommendation A-3 provides the hardware support required for initial development of prototypes.

Recommendation B-2: A significant barrier to rapid progress in HPC application lies in the formulation of the computational strategy for solving a scientific or engineering problem. In response to Challenge 1, the NSF should focus attention, both through CISE and through its disciplinary program offices, on support for the design and development of computational techniques, algorithmic methodology, and mathematical, physical and engineering models to make efficient use of the machines.

BALANCING THE PYRAMID OF HPC ACCESS

Recommendation B-3: We recommend NSF set up a task force to develop a way to ameliorate the imbalance in the HPC "pyramid" — the under-investment in the emerging mid-range scalable, parallel computers and the inequality of access to stand-alone (but potentially networked) workstations in the disciplines. This implementation plan should involve a combination of funding by disciplinary program offices and some form of more centralized allocation of NSF resources.

C. THE NSF HPC CENTERS

Recommendation C-1 : The Centers should be retained and their missions should be reaffirmed. However, the NSF HPC effort now embraces a variety of institutions and programs — HPC Centers, Engineering Research Centers, and Science & Technology Centers devoted to HPC research, and disciplinary investments in computer and computational science and applied mathematics — all of which are essential elements of the HPC effort needed for the next decade. Furthermore, HPC institutions outside the NSF orbit also contribute to the goals for which the NSF Centers are chartered. Thus we ask the Board to recognize that the overall structure of the HPC program at NSF will have more institutional diversity, more flexibility, and more interdependence with

other agencies and private institutions than was possible in the early years of the HPC initiative.

The NSF should continue its current practice of encouraging HPC Center collaboration, both with one another and with other entities engaged in HPC work. The division of the support budget into one component committed to the centers and another for multi-center activities is a useful management tool, even though it may have the effect of reducing competition among centers. The National Consortium for HPC (NCHPC), formed by NSF and ARPA is a welcome measure as well.

Recommendation C-2 : The current situation in HPC is both more exciting, more turbulent, and more filled with promise of really big benefits to the nation than at any time since the Lax report; this is not the time to "sunset" a successful, changing venture, of which the Centers remain an important part. Furthermore, we also recommend against re-competition of the four Centers at this time, favoring periodic performance evaluation and competition for some elements of their activities, both among Centers and when appropriate with other HPC Centers such as those operated by states (see Recommendation D-1).

Recommendation C-3 : The mission of the Centers is to foster rapid progress in the use of HPC by scientists and engineers, to accelerate progress in usability and economy of HPC and to diffuse HPC capability throughout the technical community, including industry. Provision to scientists and engineers of access to leading edge supercomputer resources will contine to be a primary purpose of the Centers. The following additional components of the Center missions should be affirmed:

- Supporting computational science, by research and demonstration in the solution of significant science and engineering problems.
- Fostering interdisciplinary collaboration — across sciences and between sciences and computational science and computer science — as in the Grand Challenge programs.
- Prototyping and evaluating software, new architectures, and the uses of high speed data communications in collaboration with: computer and computational scientists, disciplinary scientists exploiting HPC resources, the HPC industry, and business firms exploring expanded use of HPC.
- Training and education, from post-docs and faculty specialists to introduction of less experienced researchers to HPC methods, to collaboration with state and regional HPC centers working with high schools and community colleges.

ALLOCATION OF CENTER HPC RESOURCES TO INVESTIGATORS

Recommendation C-4: The NSF should continue to monitor the administrative procedures used to allocate Center resources, and the relationship of this process to the initial funding of the research by the disciplinary program offices, to ensure that the burden on scientists applying for research support is minimized. NSF should continue to provide HPC resources to the research community through allocation committees that evaluate competitively proposals for use of Center resources.

EDUCATION AND TRAINING

Recommendation C-5: The NSF should give strong emphasis to its education mission in HPC, and should actively seek collaboration with state-sponsored and other HPC centers not supported primarily on NSF funding. Supercomputing regional affiliates should be candidates for NSF support, with education as a key role. HPC will also figure in the Administration's industrial extension program, in which the states have the primary operational role.

D. NSF AND THE NATIONAL HPC EFFORT; RELATIONSHIPS WITH THE STATES

Recommendation D-1: We recommend that NSF urge OSTP to establish an advisory committee representing the states, HPC users, NSF Centers, computer manufacturers, computer and computational scientists (similar to the Federal Networking Council's Advisory Committee), which should report to HPCCIT. A particularly important role for this body would be to facilitate state-federal planning related to high performance computing.

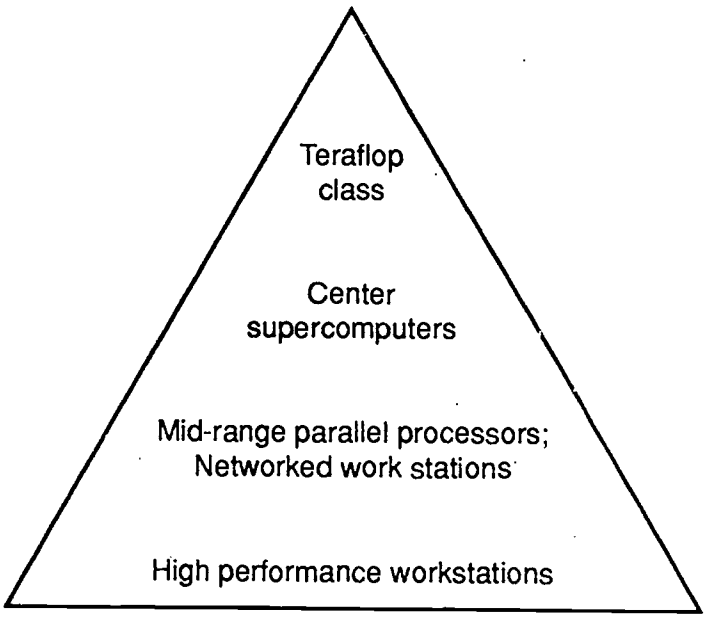


Figure A
PYRAMID OF HIGH PERFORMANCE COMPUTING ENVIRONMENTS

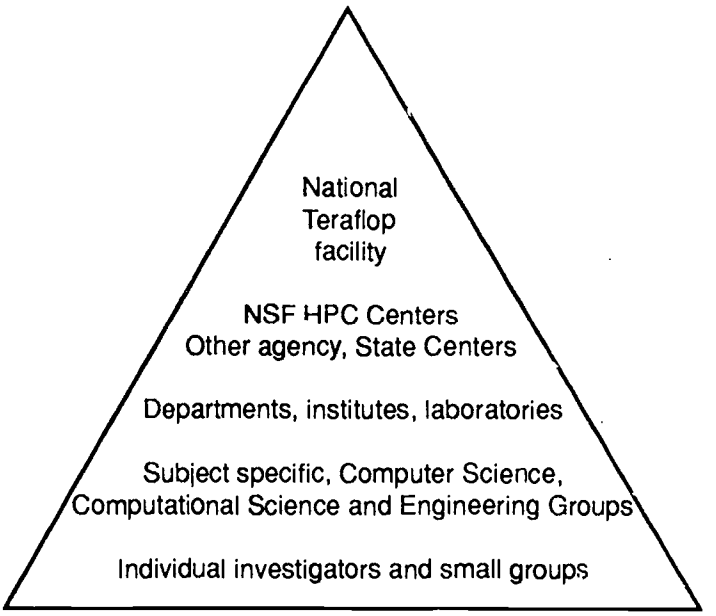


Figure B
PYRAMID OF HIGH PERFORMANCE COMPUTING INSTITUTIONS

INTRODUCTION AND BACKGROUND

A revolution is underway in the practice of science and engineering, arising from advances in computational science and new models for scientific phenomena, and made possible by advances in computer science and technology. The importance of this revolution is not yet fully appreciated because of the limited fraction of the technical community that has developed the skills required and has access to high performance computational resources. These skill and access barriers can be dramatically lowered, and if they are, a new level of creativity and progress in science and engineering may be realized which will be quite different from that known in the past. This report is about that opportunity for all of science and engineering; it is not about the needs of one or two specialized disciplines.

A little over a decade ago, the National Science Board convened a panel chaired by Prof. Peter Lax to explore what should NSF do to exploit the potential for science and industry of the rapid advances in high performance computing.[1] The actions taken by the NSF with the encouragement of the Board to implement the "Large Scale Computing in Science and Engineering" Report of 1982 have helped computing foster a revolution in science and engineering research and practice, in academic institutions and to a lesser extent in industrial applications. At the time, centralized facilities were the only way to provide access to high performance computing, which compelled the Lax panel to recommend the establishment of NSF Supercomputer Centers interconnected by a high speed network. The new revolution is characterized both by advances in the power of supercomputers and by the diffusion throughout the nation of access to and experience with using high performance computing.[2] This success has opened up a vast set of new research and applications problems amenable to solution through high levels of computational power and better computational tools.

The key features of the new capabilities include:

- The power of the big, multiprocessing vector supercomputers, today's workhorse of supercomputing, has increased by a factor of 100 to 200 since the Lax Report.[3]

- An exciting array of massively parallel processors (MPP) have appeared in the market, offering three possibilities: an acceleration in the rate of advance of peak processing power, an improvement in the ratio of performance to cost, and the option to grow the power of an installation incrementally as the need arises.[4]

- Switched networks based on high speed digital communications are extending access to major computational facilities, permitting the dynamic redeployment of computing power to suit the users' needs, and improving connectivity among collaborating users.

---

[1]Report of the Panel on Large Scale Computing in Science and Engineering, Peter Lax, chairman, commissioned by the National Science Board in cooperation with the U.S. Department of Defense, Department of Energy, and the National Aeronautics and Space Administration, December 26, 1982.

[2]With every new generation of computing machines, the capability associated with "high performance computing" changes. High performance computing (HPC) may be defined as "a computation and communications capability that allows individuals and groups to extend their ability to solve research, design, and modelling problems substantially beyond that available to them before." This definition recognizes that HPC is a relative and changing concept. For the PC user a scientific workstation is high performance computing. For the technical people with specialized skill in computational science and access to high performance facilities, a reasonable level for 1992-1993 might be 1 Gflop for a vector machine and 2 Gflops for a MPP system.

[3]As noted in Appendix C, the clock speed of a single vector processor has only increased by a factor of 5 to 6 since 1976, but a 16-way Cray C-90 with one additional vector pipe multiplies the effective spedd by the estimated factor of a hundred or more.

[4]The promise (not yet realized) of massively parallel systems is a much higher degree of installed capacity expandability with minimal disruption to the user's programming.

• Technical progress in computer science and microelectronics have transformed yesterday's supercomputers into today's emerging desktop workstations. These workstations offer more flexible tradeoffs between ease of access and inherent computing power and can be coupled to the largest supercomputers over a national network, used in locally-networked clusters, or as standalone processors.

• Advances in computer architectures, computational mathematics, algorithmic modeling, and software, along with new computer architectures, are solving some of the most intractable but important scientific, technical, and economic problems facing our society.

To address these changes, the National Science Board charged this panel with taking a fresh look at the current situation and new directions that might be required. (See Appendix A for institutional identification of the panel membership and Appendix B for historical background leading to the present study and the Charge to the Panel.)

To provide both direction and potential to exploit these advances, a leadership role for the NSF continues to be required. The goal of this report is to suggest how NSF should evolve its role in high performance computing. Our belief that NSF can and should continue to exert influence in these fields is based in part on its past successes achieved through the NSF Program in High Performance Computing and Communications.

Achievements Since the Lax Report

In the past 10 years, the NSF Program in High Performance Computing and Communications has:

• Facilitated many new scientific discoveries and new industrial processes, and supported fundamental work which has led to advances in architectures, tools and algorithms for computational science.

> In Appendix E of this report several panel members describe examples of those accomplishments and suggest their personal visions for what may be even more dramatic progress in the future.

• Supported fundamental work in computer science and engineering which has led to advances in architectures, tools, and algorithms for computational science.

• Initiated collaborations with many companies to help them realize the economic and technological benefits of high performance computing.

> Caterpillar Inc. uses supercomputing to model diesel engines in an attempt to reduce emissions.
>
> Dow Chemical Company simulates and visualizes fluid flow in chemical processes to ensure complete mixing.
>
> USX has turned to supercomputing to improve the hot rolling process-control systems used in steel manufacturing.
>
> Solar Turbine, Inc. applies computational finite-element methods to the design of very complex mechanical systems.

• Opened up supercomputer access to a wide range of researchers and industrial scientists and engineers.

> This was one of the key recommendations of the Lax Report. The establishment of the four NSF Supercomputer Centers (in addition to NCAR) has been extraordinarily successful. By providing network access, through the NSFNET and Internet linkages, NSF has put these computing resources at the fingertips of scientists, engineers, mathematicians and other professionals all over the nation. Users seldom need to go personally to these Centers; in fact, the distribution of computational cycles by the four NSF Supercomputer Centers shows surprisingly little geographic bias. This extension of compute power, away from dedicated, on-site facilities and towards a seamless national computing environment has been instrumental in creating the conditions required for advances on a broad front in science, engineering, and the tools of computational science. There seems to be a lack

of geographic bias in users - Figure 1 in Appendix D shows users widely distributed across the United States.

• Educated literally thousands of scientists, engineers and students, as well as a new generation of researchers who now use computational science equally with theory and experiment.

At the time of the Lax Report access to the most advanced facilities was restricted to a relatively small set of users. Furthermore supercomputing was regarded by many scientists as either an inaccessible tool or as an inelegantly brute force approach to science. The NSF program successfully inoculated virtually all of the disciplines with the realization that HPC is both a powerful and a practical tool for many purposes. These NSF initiatives have not only pushed the technology and computational science ahead in sophistication and power, they have helped bring high performance computing to a large fraction of the technical community. There has been a 5-fold increase in number of NSF funded scientists using HPC and a 5-fold increase in ratio of graduate students to faculty using HPC through the NSF Supercomputer Centers. (See Figure 2 of Appendix D)

• Provided the HPC industry a committed, enthusiastic, and dedicated class of expert users who share their experience and ideas with vendors, accelerating the evolutionary improvement in the technology and its software.

One of the problems in the migration of new technologies from experimental environments to production modes are the inherent risks in committing substantial resources towards converting existing codes and developing software tools. The NSF Supercomputer Centers have provided a proving ground for these new technologies; various industrial players have entered into partnerships with the Centers aimed at accelerating this migration while maintaining solid and reliable underpinnings.

• Encouraged the Supercomputer Centers to leverage their relationship with HPC producers to reduce the cost of bringing innovation to the scientific and engineering communities.

In recognition of Center activity in improving early versions of hardware and software for high performance computing systems, the computer industry has provided equipment at favorable prices and important technical support. This has allowed researchers earlier and more useful access to HPC facilities than might have been the case under commercial terms.

• Joined into successful partnerships with other agencies to make coordinated contributions to the U.S. capability in HPC.

A decade ago the United States enjoyed a world-wide commercial lead in vector systems. In part as the result of more recent development and procurement actions of the Advanced Projects Agency, the Department of Energy, and the National Science Foundation, the U.S. now has the dominant lead in providing new Massively Parallel Processing (MPP) systems.[5] As an example, the NSF has enabled NSF Supercomputer Center acquisitions of scalable parallel systems first developed under seed money provided by ARPA, and thus has been instrumental in leveraging ARPA projects into the

[5]Massively parallel computers are constructed from large numbers of separate processors linked by high speed communications providing access to each other and to shared I/O devices and/or computer memory. There are many different architectural forms of MPP machines, but they have in common economies of scale from the use of microprocessors produced at high volumes and the ability to combine them at many levels of aggregation. The challenge in using such machines is to formulate the problem so that it can be decomposed and run efficiently on most or all of the processors concurrently. Some scientific problems lend themselves to parallel computation much more easily than others, suggesting that improved utility of MPP machines will not be availed in all fields of science at once.

mainstream.[6] (Figure 3 of Appendix D shows data on the uptake of advanced computing by sector across the world).

The Lax Report

All of these accomplishments have, in a large part, arisen from the response by NSF to the recommendations of the 1982 Lax report "Large Scale Computing in Science and Engineering". These recommendations included:

- Increase access to regularly upgraded supercomputing facilities via high bandwidth networks.
- Increase research in computational mathematics, software, and algorithms.
- Train people in scientific computing.
- Invest in research on new supercomputer systems.

For several reasons, NSF's investment in computational research and training has been a startling success. First, there has been a widespread acceptance of computational science as a vital component and tool in scientific and technological understanding. Second, there have been revolutionary advances in computing technology in the past decade. And third, the demonstrated ability to solve key critical problems has advanced the progress of mathematics, science and engineering in many important ways, and has created great demand for additional HPC resources.

The New Opportunities in Science and in Industry

As discussed in detail in the Appendix E essays, the prospects are for dramatic progress in science and engineering and for rapid adoption of computational science in industry. The next major HPC revolution may well be in industry, which is still seriously under-utilizing HPC (with some exceptions such as aerospace, automotive, and microelectronics).

The success of the chemical industry in designing and simulating pilot plants, of the aircraft industry in simulating wind tunnels and performing dynamic design evaluation, and in the electronics industry in designing integrated circuits and modelling the performance of computers and networks suggests the scale of available opportunities. The most important requirements are (a) improving the usability and efficiency problems of high performance machines, and (b) training in HPC for people going into industry. The Supercomputer Centers have demonstrated they can introduce the commercial sector to HPC at little cost, and with high potential benefits to economy (productivity of industry and stimulation of markets for U.S. HPC vendors). Success in stimulating HPC usage in industry will also accelerate need for HPC education and technology, thus exploiting the benefits of collaboration with universities and vendors. The Centers' role can be a catalytic one, but often rises to the level of a true collaborative partnership with industry, to the mutual advantage of the firm and the NSF Centers. As industrial uses of HPC grow, the scientists, mathematicians, and engineers benefit from the falling costs and rising usability of the new equipment. In addition the technological uses of HPC spur new and interesting problems in science. The following chart indicates the increasing importance of advanced computing in industry.

Cray Research Inc. supercomputer sales

Era	Percent to government	Percent to industry	Percent to Universities
Early 1980s	70	25	5
Late 1980s	60	25	15
Today	40	40	20

[6]Scalable parallel machines are those in which the number of processor nodes can be expanded over a wide range without substantial changes in either the shared hardware or the application interfaces of the operating system.

The New Technology

Most HPC production work being done today uses big vector machines in single processor (or loosely coupled multiprocessor) mode. Vectorizing Fortran compilers and other software tools are well tested and many people have been trained in their use. These big shared memory machines will continue to be the mainstay of high performance computing, at least for the next 5 years or so, and perhaps beyond if the promise of massively parallel supercomputing is delayed longer than many expect.

New desk top computers have made extraordinary gains in cost-performance (driven by competition-driven commodity microprocessor production). Justin Rattner of Intel estimated that in 1996 microprocessors with clock speeds of 200 MHz may power an 800 Mflops peak speed workstation.[7] He, and others from the industry, predicted the convergence of the clock speeds of microprocessor chips and the large vector machines such as the Cray C90, perhaps as soon as 1995. They held out the likelihood that in 1997 microprocessors may be available at 1 gigaflop; a desktop PC might be available with this speed for $10,000 or less. Mid-range workstations will also show great growth in capacity; Today one can purchase a mid-range workstation with a clock speed of 200MHz for an entry price of $40,000 to $50,000.

Thus a technical transition is underway from the world in which uniprocessor supercomputers were distinguished from desktop machines by having much faster cycle times, to a world in which cycle times converge and the highest levels of computer power will be delivered through parallelism, memory size and bandwidth, and I/O speed. The widespread availability of scientific workstations will accelerate the introduction of more scientists and engineers to high performance computing, resulting in a further acceleration of the need for higher performance machines. Early exploration of message-passing distributed operating systems gives promise of loosely-coupled arrays of workstations being used to process large problems in the background and when the workstations are unused at night, as well as coupling the workstations (on which problems are initially designed and tested) to the supercomputers located at remote facilities.

Of course, the faster microprocessors also make possible new MPP machines of ever increasing peak processing speed. MPP is catching on fast, as researchers with sufficient expertise (and diligence) in computational science are solving a growing number of applications that lend themselves to highly parallel architectures. In some cases those investigators are realizing a ratio of theoretical to peak performance approaching that achieved by vector machines, with significant cost-performance advantages. Efficient use of MPP on the broad range of scientific and engineering problems is still beyond the reach of most investigators, however, because of the expertise and effort required. Thus the first speculative phase of MPP HPC is coming to an end, but its ultimate potential is still uncertain and largely unrealized.

Limiting progress in all three of these technologies is a set of architecture and software issues that are discussed below in Recommendations B. Principal among them is the evolution of a programming model that can allow portability of applications software across architectures.

These technical issues are discussed at greater length in Appendix C.

---

[7]The instruction execution speeds of scientific computers are generally reckoned in the number of floating point instructions that can be executed in one second. Thus a 1 Megaflop machine executes 1 million floating point instructions per second, a Gigaflop would be one billion instructions per second, and a Teraflop 10^{12} floating point instructions per second. Since different computer architectures may have quite different instruction sets one "flop" may not be the same as another, either in application power or in the number of machine cycles required. To avoid such difficulties, those who want to compare machines of different architecture generally use a benchmark suite of test cases to measure overall performance on each machine.

FOUR CHALLENGES FOR NSF

High performance computing is changing very fast, and NSF policy must chase a moving target. For that reason, the strategy adopted must be agile and flexible in order to capitalize on past investments and adapt to the emerging opportunities. The Board and the Foundation face four central challenges, on which we will make specific recommendations for policy and action. These challenges are:

- Removing barriers to the rapid evolution of HPC capability
- Providing scalable access to all levels of HPC capability
- Finding the right incentives to promote access to all three levels of the computational power pyramid
- Creating NSF's intellectual and management leadership for the future of high performance computing in the U.S.

CHALLENGE NO. 1: Removing barriers to the rapid evolution of HPC capability

How can NSF, as the nation's premier agency funding basic research, remove existing barriers to the rapid evolution of High Performance Computing? These barriers are of two kinds: technological barriers (primarily to realizing the promise of highly parallel machines, workstations, and networks) and exploitation barriers (new mathematical methods and new ways to formulate science and engineering problems for efficient and effective computation). An aggressive commitment by NSF to leadership in research and prototype development, in both computer science and computational science, will be required. Indeed, NSF's position as the leading provider of HPC capability to the nation's scientists and engineers will be strengthened if it commands a leadership role in technical advances in both areas, which will contribute to the nation's economic position as well as its position as a world leader in research.

Computer Science and Engineering. The first challenge is to accelerate the development of the technology underlying high performance computing. Among the largest barriers to effective use of the emerging HPC technologies are parallel architectures from which it is easy to extract peak performance, system software (operating systems, databases of massive size, compilers, and programming models) to take advantage of these architectures and provide portability of end-user applications, parallel algorithms, and advances in visualization techniques to aid in the interpretation of results. The technical barriers to progress are discussed in Appendix C. What steps will most effectively reduce these barriers?

Computational Tools for Advancing Science and Engineering. Research in the development of computational models, the design of algorithmic techniques, and their accompanying mathematical and numerical analysis, is required in order to ensure the continued evolution of efficient and accurate computational algorithms designed to make optimal use of these emerging technologies.

In the past ten years, exciting developments in computer architectures, hardware and software have come in tandem with stunning breakthroughs in computational techniques, mathematical analysis, and scientific models. For example, the potential of parallel machines has been realized in part through new versions of numerical linear algebra routines and multigrid techniques; rethinking and reformulating algorithms for computational physics within the domain of parallel machines has posed significant and challenging research questions. Advances in such areas of N-body solvers, fast special function techniques, wavelets, high resolution fluid solvers, adaptive mesh techniques, and approximation theory have generated highly sophisticated algorithms to handle complex problems. At the same time, important theoretical advances in the modelling of underlying physical and engineering problems have led to new, efficient and accurate discretization techniques. Indeed, in the evolution to scalable

computing across a range of levels, designing appropriate numerical and computational techniques is of paramount importance. The challenge facing NSF is to weave together existing work in these areas, as well as fostering new bridges between pure, applied and computational techniques, engaging the talents of disciplinary scientists, engineers, and mathematicians.

CHALLENGE NO. 2: Providing scalable access to all levels of HPC capability

How can NSF provide scalable access to computing resources, from the high performance workstations needed by most scientists to the critically needed teraflop-and-beyond capability required for solving Grand Challenge problems?[8] What balance should NSF anticipate and encourage among high performance desktop workstations, mid-range or mini-supercomputers, networks of workstations, and remote, shared supercomputers of very high performance?

Flexible strategy. NSF must ensure that adequate additional computational capacity is available to a steadily growing user community to solve the next generation of more complex science and engineering problems. A flexible and responsive strategy that can support the large number of evolving options for HPC and can adapt to the outcomes of major current development efforts (for example in MPP systems and in networked workstations) is required.

A pyramid of computational capability. There will continue to be an available spectrum spanning almost five orders of magnitude of computer capabilities and prices.[9] NSF, as a leader in the national effort in high performance computing, should support a "pyramid" of computing capability. At the apex of the pyramid is the highest performance systems that affordable technology permits, established at national facilities. At the next level, every major research university should have access to one, or a few, intermediate-scale high-performance systems and/or aggregated workstation clusters.[10] At the lowest level are workstations with visualization capabilities in sufficient numbers to support computational scientists and engineers.

Mid-range computational requirements. Over the next five years, the middle range of scientific computing and computational engineering will be handled by an amazing variety of moderately parallel systems. In some cases, these will be scaled-down versions of the highest performance systems available; in other cases, they will be systems targeted at the midrange computing market. The architecture will vary from shared memory at one end of the spectrum to workstation networks at the other, depending on the types of parallelism in the local spectrum of applications. Loosely coupled networks of workstations will compete with mid-range systems for performance of production HPC work. At the same time autonomous mid-range systems are needed to support the development of next-generation architectures and software by computer science groups.

The panel perceives that there are imbalances in access to the pyramid of HPC resources (see the following table). The disciplinary NSF program offices have not been uniformly effective in responding to the need for a desktop environ-

[8] By scalable access we mean the ability to develop a problem on a workstation or intermediate sized machine and migrate the problem with relative efficiency to larger machines as increased complexity requires it. Scalable access implies scalable architectures and software.

[9] A Paragon machine of 300 Gigaflops peak performance would be five orders of magnitude faster than a 3 megaflop entry workstation. Effective performance in most science applications would, however, be perhaps a factor of ten lower.

[10] As discussed in the recommendations, dedicated mid-range systems are required not only for science and engineering applications but also for research to improve HPC hardware and software, and for interactive usage. For science and science and engineering batch applications, networks of workstations will likely develop into an alternative.

ment for their supported researchers, and there is serious under-investment in the mid-sized machines. The distribution of investment tends to be bimodal, to the disadvantage of mid-range systems. The incentive structures internal to the Foundation do not address this distortion. NSF's HPCC coordinating mechanism needs to address this distortion in a more direct manner.

Computational Infrastructure at NSF
(FY92 $, M)

	Other NSF	ASC
Workstations . . .	20.1	3.2
Small Parallel . . .	2.1	0.5
Large Parallel . . .	9.4	3.2
Mainframe	9.1	16.3
Total	40.8	23.2

CHALLENGE NO. 3: The right incentives to promote access to all three levels of the computational institution pyramid

The third challenge is to encourage the continued broadening of the base of participation in HPC, both in terms of institutions and in terms of skill levels and disciplines.

Lax Report incentives. At the time of the Lax report, relatively few people were interested in HPC; even fewer had access to supercomputers. Some users were fortunate to have contacts with someone at one of a few select government laboratories where computer resources were available. Most, however, were less fortunate and were forced to carry out their research on small departmental machines. This severely limited the research that could be carried out to problems that would "fit" into available resources. NSF addressed this problem by concentrating supercomputer resources in Centers; by this means those in the academic community most prepared and motivated were provided with access to machine cycles.

Need for expanded scope of access. Now that these resources are available on a peer review basis to everyone no matter where they work, it is clear the research community cannot accept a return to the previous mode of operation. The high performance computing community has grown to depend on NSF to make the necessary resources available to continually upgrade the Supercomputer Centers in support of their computational science and engineering applications. NSF needs to broaden the base of participation in HPC through NSF program offices as well as through the Supercomputer Centers. There is no question that HPC has broken out of its original narrow group of privileged HPC specialists. The SuperQuest competition for high school students already demonstrates how quickly young people can master the effective use of HPC facilities. Other agencies, states, and private HPC centers are springing up, making major contributions not only to science but to K-12 education and to regional economies. NSF's policies on expanding access and training must take advantage of the leverage these Supercomputer Centers can provide.

Allocation of HPC resources. There remains the question of the best way to allocate HPC resources. Should Supercomputer Centers continue to be funded to allocate HPC cycles competitively, or should NSF depend on the "market" of funded investigators for allocation of HPC resources? This question gets at two other issues: (a) the future role of the Centers and (b) the best means for insuring adequate funding of workstations and other means of HPC access throughout the NSF. The Centers have peer review committees which allocate HPC resources on the basis of competitive project selection. The Panel believes these allocations are fairly made and reflect solid professional evaluation of computational merit. The only remaining issue is whether there continues to be a need for protected funding for HPC access in NSF, including access to shared Supercomputer Centers facilities? We believe strongly that there is such a need. The panel does have suggestions for broadening the support for the remainder of the HPC pyramid;

these are articulated in the recommendations below.

Education and training. A major requirement for education and training continues to exist. Even though most disciplines have been inoculated with successful uses of HPC (see Appendix D essays), and even though graduate student and postdoctoral uses of HPC resources is rising faster than faculty usage, only a minority of scientists have the training to allow them to overcome the initial barrier to proficiency, especially in the use of MPP machines which require a high level of computational sophistication for most problems.

CHALLENGE NO. 4: How can NSF best create the intellectual and management leadership for the future of high performance computing in the U.S.?

What relationships should NSF's activities in HPC have to the activities of other federal agencies? NSF is a major player. What role should NSF play within the scope of the nationally coordinated HPCC program and budget, as indicated in the following chart?

HPCC Agency Budgets

Agency	FY92 Funding ($, M)
ARPA	232.2
NSF	200.9
DOE	92.3
NASA	71.2
HHS/NIH	41.3
DOC/NOAA	9.8
EPA	5.0
DOC/NIST	2.1

NSF leadership in HPCC. The voice of HPCC users needs to be more effectively felt in the national program; NSF has the best contact with this community. NSF has played, and continues to play, a leadership role in the NREN program and the evolution of the Internet. Its initiative in creating the "meta-center" concept establishes an NSF role in the sharing and coordination of resources (not only in NSF but in other cooperating agencies as well), and the concept can be usefully extended to cooperating facilities at state level and in private firms. The question is, does the current structure in CISE, the HPCC coordination office, the Supercomputer Centers, and the science and engineering directorates constitute the most favorable arrangement for that leadership? The panel does not attempt to suggest the best ways to manage the relationships among these important functions, but asks the NSF leadership to assure the level of attention and coordination required to implement the broad goals of this report.

Networking. The third barrier is the need for network access with adequate bandwidth. For wide area networks, this is addressed in the NSF HPCC NREN strategy. In the future, NSF will focus its network subsidies on HPC applications and their supporting infrastructure, while support for basic Internet connectivity shifts to the research and education institutions.[11]

[11]NREN is the National Research and Education Network, envisioned in the High Performance Computing Act of 1991. NREN is not a network so much as it is a program of activities including the evolution of the Internet to serve the needs of HPC as well as other information activities.

RECOMMENDATIONS

We have four sets of interdependent recommendations for the National Science Board and the Foundation. The first implements a balanced pyramid of computing environments; each element supports the others, and as priorities are applied the balance in the pyramid should be sustained. The second set addresses the essential research investments and other steps to remove the obstacles to realizing the technologies of the pyramid and the barriers to the effective use of these environments. The third set addresses the institutional structure for the delivery of HPC capabilities, and consists itself of a pyramid. At the base of the institutional pyramid is the diverse array of investigators in their universities and other settings who use all the facilities at all levels of the pyramid. At the next level are departments and research groups devoted to specific areas of computer science or computational science and engineering. Continuing upward are the NSF HPC Centers, which must continue to play a very important role, both as providers of the major resources of high capability computing systems and as aggregations of specialized capability for all aspects of use and advance of high performance computing. At the apex is the national teraflop facility, which we recommend as a multi-agency facility pushing the frontiers of high performance into the next decade. A final recommendation addresses the NSF's role at the national level and its relationship with the states in HPC.

This report recommends significant expansion in NSF investments, both in accelerating progress in high performance computing through computer and computational science research and in providing the balanced pyramid of computing facilities to the science and engineering communities, but in total they do not exceed the Administration's stated intent to double the investments in HPCC during the next 5 years. We believe these investments are not only justified, but are compatible with stated national plans, both in absolute amount and in their distribution.

A. CENTRAL GOAL FOR NSF HPC POLICY

Recommendation A-1: We strongly recommend that NSF build on its success in helping the U.S. achieve its preeminent world position in high performance computing by taking the lead, under OSTP guidance and in collaboration with ARPA, DoE and other agencies, to expand access to all levels of the rapidly evolving pyramid of high performance computing for all sectors of the nation. The realization of this pyramid depends, of course, on rapid progress in the pyramid's technologies.

High performance computing is essential to the leading edge of U.S. research and development. It will provide the intelligence and power that justifies the breadth of connectivity and access promised by the NREN and the National Information Infrastructure. The computational capability we envision includes not only the research capability for which NSF has special stewardship, but also includes a rapid expansion of capability in business and industry to use HPC profitably and the many operational uses of HPC in commercial and military activities.

The panel is concerned that if the government fails to implement the planned HPCC investments to support the National Information Infrastructure, the momentum of the U.S. industry, which blossomed in the first phase of the national effort, will be lost. Supercomputers are only a $2 billion industry, but an industry that provides critical tools for innovation across all areas of U.S. competitiveness, including pharmaceuticals, oil, aerospace, automotive, and others. The administration's planned new investment of $250 million in HPCC is fully justified. Japanese competitors could easily close the gap in the HPC sectors in which the U.S. enjoys that lead; they are continuing to invest and could capture much of the market the U.S. government has been helping to create.

VISION OF THE HPC PYRAMID

Recommendation A-2: At the apex of the HPC pyramid is a need for a national capability at the highest level of computing power the industry can support with both efficient software and hardware.

A reasonable goal for the next 2-3 years would be the design, development, and realization of a national teraflop-class capability, subject to the effective implementation of Recommendation B-1 and the development of effective software and computational tools for such a large machine.[12] Such a capability would provide a significant stimulus to commercial development of a prototype high-end commercial HPC system of the future. We believe the importance of NSF's mission in HPC justifies NSF initiating an interagency plan to make this investment, and further that NSF should propose to operate the facility in support of national goals in science and technology. For budgetary and interagency collaboration reasons OSTP should invoke a FCCSET project to establish such a capability on a government-wide basis with multi-agency funding and usage.

If development begins in 1995 or 1996, a reasonable guess at the cost of a teraflop machine is $50/megaflop for delivery in 1997 to 1998. If so, $50 million a year might buy one such machine per year.[13] Development cost would be substantial, perhaps in excess of the production cost of one machine; although it is not clear to what extent government support would be required, this is a further reason to suggest a multi-agency program.[14] Support costs would also be additional, but one can assume that one or more of the NSF Supercomputer Centers could host such a facility with something like the current staff.

Such a nationally shared machine, or machines, must be open to competitive merit-evaluated proposals for science and engineering computation, although it could share this mission of responding to the community's research priorities with mission-directed work of the sponsoring agencies. The investment is justified by (a) the existence of problems whose solution awaits a teraflop machine, (b) the importance of driving the HPC industry's innovation rate, (c) the need for early and concrete experience with the scalability of software environments to higher speeds and larger arrays of processors, since software development time is the limiting factor to hardware acceptance in the market.

---

[12]Some panel members have reservations about the urgency of this recommendation, are pessimistic about the likelihood of realizing the effective performance in applications, or are concerned about the possible opportunity cost to NSF of such a large project. The majority notes that the recommendation is intended to drive solutions to those architectural and software problems. Intel's Paragon machine is on the market today with 0.3 Teraflops peak speed, but without the support to deliver that speed in most applications. The panel also recommends a multi-agency federal effort. NSF's share of cost and role in managing such a project are left to a proposed FCCSET review.

[13]The cost estimates in this report cannot be much more than informed guesses. We have assumed a cost of $50/megaflop for purchase of a one teraflop machine in 1997 or 1998. We suspect that this cost might be reached earlier, say in 1995 or 1996 in a mid-range machine, because a tightly-coupled massively parallel machine may have costs rising more than linearly with the number of processors, overcoming the scale economies that might make the cost rise less than linearly. The cost estimates in recommendations A2-4 are intended to indicate that scale of investment we recommend is not incompatible with the published pllans of the administration for investment in HPCC in the next 5 years, and further that roughly equal levels of incremental expenditures in the three levels of the HPC pyramid could produce the balance among these levels that we recommend.

[14]The Departments of Energy and Defense and NASA might share a major portion of the development cost and might also acquire such machines in the future as well.

Recommendation A-3: Over a period of 5 years the research universities should be assisted to acquire mid-range machines.

This will bring a rapid expansion in access to very robust capability, reducing pressure on the Supercomputer Centers' largest facilities, and allowing the variety of vendor solutions to be exercised extensively. If the new MPP architectures prove robust, usable, and scalable, these institutions will be able to grow the capacity of such system in proportion to need and with whatever incremental resources are available. This capability is also needed to provide testbeds for computer and computational science research and testing.

These mid-sized machines are the underfunded element today — less than 5% of NSF's FY92 HPC budget is devoted to their acquisition. They are needed for both demanding science and engineering problems that do not require the very maximum in computing capacity, and importantly for use by the computer science and computational mathematics community in addressing the architectural, software, and algorithmic issues that are the primary barriers to progress with MPP architectures.[15]

Engineering is also a key candidate for their use. There are 1050 University-Industry Research Centers in the U.S. Those UIRCs that are properly equipped with computational facilities can increase the coupling with industrial computation, adding greatly to what the NSF HPC Supercomputer Centers are doing. Many engineering applications, such as robotics research, require "real time" interactive computation which is incompatible with the batch environment on the highest performance machines.

If we assume a cost in three or four years of $50/megaflop for mid-sized MPP machines, an annual expenditure of $10 million would fund the annual acquisition of one hundred 2 Gigaflop (peak) computers. Support costs for users would be additional.

Recommendation A-4: We recommend that NSF double the current annual level of investment ($22 million) in scientific and engineering workstations for its 20,000 principal investigators.

Many researchers strongly prefer the new high performance workstations that are under their control and find them adequate to meet many of their initial needs. Those without access to the new workstations may apply to use remote access to a supercomputer in a Center, but often they do not need all the I/O and other capabilities of the large shared facilities. NSF needs a strategy to off-load work not requiring the highest level machines in the Centers. The justification is not economy of scale, but economy of talent and time.

When the Lax report was written a 160 Mflop peak Cray 1 was a high performance supercomputer. Within 4 or 5 years workstations delivering up to 400 megaflops costing no more than $15,000 to $20,000 should be widely available. For education and a large fraction of the computational needs of science and engineering, these facilities will be adequate. However, once visualization of computational output becomes routinely required they will be ubiquitously needed. With the rapid pace of improvement, the useful lifetimes of workstations are decreasing rapidly; they often cannot cope with the latest software. Researchers face escalating costs to upgrade their computers. NSF supports perhaps some 20,000 principal investigators. Equipping an additional 10 percent of this number each year (2,000 machines) at $20,000 each requires an incremental $20 million. Recommendation B-3 addresses how this investment might be managed.

Recommendation A-5: We recommend that NSF expand its New Technologies program to support expanded testing of the practicality of new parallel

[15]The development of prototypes of architectures and operating systems for parallel computation requires access to a machine whose hardware and software can be experimentally modified. This research often cannot be done on machines dedicated to full time production.

configurations for HPC applications.[16] For example, networks of workstations may meet a significant part of midrange HPC science and engineering applications. As progress is made in the development of this and other technologies, experimental use of the new configurations should be encouraged. A significant supplement to HPC applications research capacity can be had with minimal additional cost if such collections of workstations prove practical and efficient.

There have already been sufficient experiments with use of distributed file systems and loosely coupled workstations to encourage the belief than many compute-intensive problems are amenable to this approach. For those problems that do not suffer from the latency inherent in this approach the incremental costs can be very low indeed, for the problems run in background and at times the workstations are otherwise unengaged. There are those who strongly believe that in combination with object-oriented programming this approach can create a revolution in software and algorithm sharing as well as more economical machine cycles.[17]

B. RECOMMENDATIONS TO IMPLEMENT THESE GOALS

REMOVING BARRIERS TO HPC TECHNICAL PROGRESS AND HPC USAGE

Recommendation B-1: To accelerate progress in developing the HPC technology needed by users, NSF should create, in CISE, a challenge program in computer science with grant size and equipment access sufficient to support the systems and algorithm research needed for more rapid progress in HPC. The Supercomputer Centers, in collaboration with hardware and software vendors, can provide test platforms for much of this work. Recommendation A-3 provides the hardware support required for initial development of prototypes.

There is consensus that the absence of sufficient funding for systems and algorithms work which is not mission-oriented is the primary barrier to lower cost, more widely accessible, and more usable massively parallel systems. This work, including bringing the most promising ideas to prototype stage for effective transfer to the HPC industry, would address the most significant barriers to the ultimate penetration of parallel architectures in workstations. Advances on the horizon that could be accelerated include more advanced network interface architectures and operating systems technologies to provide low overhead communications in collections of workstations, and advances in algorithms and software for distributed databases of massive size. Computer science has made, and continues to make, important contributions to both hard and soft parallel machine technology, and has effectively transferred these ideas to the industry.

Two problems impede the full contribution of computer science to rapid advance in MPP development; grant sizes in the discipline are typically too small to allow enough concentrated effort to build and test prototypes, and too few computer science departments have access to a mid-sized machine on which systems development can be done.

The Board should ask for a proposal from CISE to effectively mobilize the best computer science and computational mathematics talent to addressing the solution of these problems in the areas of both improved operating systems, architectures, compilers, and algorithms for existing systems as well as research in next-generation systems. We recommend establishing a number of major projects, with higher levels of annual funding than is typical in Computer Science, and assured duration of

[16]Today NSF CISE has a "new technologies" program that co-funds with disciplinary program offices perhaps 50 projects/yr. This program is in the division that funds the Centers, but is focused on projects which can ultimately benefit all users of parallel systems. This program funds perhaps 15 methods and tools projects annually, in addition to those co-funded with science programs.

[17]MITRE Corporation, among others, is pursuing this vision.

up to five years, for a total annual incremental investment of $10 million. We recommend that this challenge fund be managed by CISE, and be accessible to all disciplinary program offices who wish to forward team proposals for add-on funding in response to specific proposals from the community.

Recommendation B-2: A significant barrier to rapid progress in the application of HPC lies in formulating a computational strategy to solve a problem. In response to Challenge 1 above, NSF should focus attention, both through CISE and through its disciplinary program offices, on support for the design and development of computational techniques, algorithmic methodology, and mathematical, physical and engineering models to make efficient use of the machines.

Without such work in both theoretical and applied areas of numerical analysis, applied mathematics, and computational algorithms, the full benefit of advances in architecture and systems software will not be realized. In particular, significantly increased funding of collaborative and individual state-of-the-art methodology is warranted, and is crucial to the success of high performance computing. Some of this can be done through the individual directorates with funds supplemented by HPCC funds; the Grand Challenge Applications Group awards are a good first step.

Recommendation B-3: We recommend NSF set up an agency-wide task force to develop a way to ameliorate the imbalance in the HPC pyramid - the under-investment in the emerging mid-range scalable, parallel computers and the inequality of access to stand-alone (but potentially networked) workstations in the disciplines. This implementation plan should involve a combination of funding by disciplinary program offices and some form of more centralized allocation of NSF resources.

Some directorates have "infrastructure" programs; others do not. Still others fund workstations until they reach the "target" set by the HPCC coordination office. We believe that individual disciplinary program managers should consider it their responsibility to fund purchase of workstations out of their equipment funds. But we recognize that these funds need to be supplemented by HPCC funds. CISE has an office which co-funds interdisciplinary applications of HPC workstations. We believe this office may require more budgetary authority than it now enjoys, to ensure the proper balance of program and CISE budgets for workstations.

Scientific value must be a primary criterion for resource allocation. It would be unwise to support mediocre projects just because they require supercomputers. The strategy of application approval will depend very heavily on funding scenarios. If sufficient HPCC funds are made available to individual programs for computer usage, then the Supercomputer Centers should be reserved for applications that cannot be carried out elsewhere, with particular priority to novel applications. If individual science programs continue to be underfunded relative to large centers, the Supercomputer Centers may be forced into a role of supporting less novel or demanding computing applications. Under these circumstances, less stringent funding criteria should be applied.

C. THE NSF SUPERCOMPUTER CENTERS

Recommendation C-1: The Supercomputer Centers should be retained and their missions, as they have evolved since the Lax Report, should be reaffirmed. However, the NSF HPC effort now embraces a variety of institutions and programs - HPC Centers, Engineering Research Centers (ERC) and Science and Technology Centers (STC) devoted to HPC research, and disciplinary investments in computer and computational science and applied mathematics - all of which are essential elements of the HPC effort needed for the next decade. NSF plays a primary but not necessarily dominant role in each of them (see Figure 4 of Appendix D). Furthermore, HPC institutions outside the NSF orbit also contribute to the goals for which the NSF Supercomputer Centers are chartered. Thus we ask the Board to recognize that the overall structure of the HPC program at

NSF will have more institutional diversity, more flexibility, and more interdependence with other agencies and private institutions than in the early years of the HPC initiative.

We anticipate an evolution, which has already begun, in which the NSF Supercomputer Centers increasingly broaden their base of support, and NSF expands its support in collaboration with other institutional settings for HPC. Center-like groups, especially NSF S&T Centers, are an important instrument for focusing on solving barriers to HPC, although they do not provide HPC resources to users. An excellent example is the multi-institutional Center for Research in Parallel Computation at Rice University, which is supported at about $4M/yr, with additional support from ARPA. Another example is the Center for Computer Graphics and Scientific Visualization, an S&T Center award to University of Utah with participation of University of N. Carolina, Brown, Caltech, and Cornell. Still another example is the Discrete Mathematics and Computational Science Center (DIMACS) at Rutgers and Princeton. These centers fill important roles today, and the ERC and S&T Center structures provide a necessary addition to the Supercomputer Centers for institutionalizing the programmatic work required for HPC.

The NSF should continue its current practice of encouraging HPC Center collaboration, both with one another and with other entities engaged in HPC work. The division of the support budget into one component committed to the Supercomputer Centers and another for multi-center activities is a useful management tool, even though it may have the effect of reducing competition among Supercomputer Centers. The National Consortium for HPC (NCHPC), formed by NSF and ARPA is a welcome measure as well.

Recommendation C-2: The current situation in HPC is more exciting, more turbulent, and more filled with promise of really big benefits to the nation than at any time since the Lax report; this is not the time to "sunset" a successful, changing venture, of which Supercomputer Centers remain an essential part. Furthermore, we also recommend against an open recompetition of the four Supercomputer Centers at this time, favoring instead periodic performance evaluation and competition for some elements of their activities, both among the Centers themselves and when appropriate with other HPC Centers such as those operated by states (see Recommendation D-1).

Continuing evaluation of each Center's performance, as well as the performance of the overall program, is, of course, an essential part of good management of the Supercomputer Centers program. Such evaluations must take place on a regular basis in order to develop a sound basis for adjustments in support levels, to provide incentives for quality performance and to recognize the need to encourage other institutions such as S&T Centers that are attacking HPC barriers and state-based centers with attractive programs in education and training. While recompetition of existing Supercomputer Centers does not appear to be appropriate at this time, if regular review of the Centers and the Centers program identifies shortcomings in a Center or the total program, a recompetition of that element of the program should be initiated.

Supercomputer Centers are highly leveraged through investments by industry, vendors, and states. This diversification of support impedes unilateral action by NSF, since the Centers' other sponsors must be consulted before decisions important to the Center are made.[18] It also suggests that the issue of recompetition may, in future, become moot as the formal designation "NSF HPC Center" erodes in sig-

[18]Each year each center gets a cooperative agreement level which is negotiated. Each center gets about $14M; about 15% is flexible. NSF centers have also received help from ARPA to buy new MPP machines. Most of the Centers have important outside sources of support, which imply obligations NSF must respect, such as the Cornell Center relationship with IBM and the San Diego Center's activities with the State of California.

nificance. There is a form of recompetition already in place; the Centers compete for support for new machine acquisition and for roles in multi-center projects.

Recommendation C-3: The NSF should continue to provide funding to support the Supercomputer Centers' HPC capacity. Any distortion in the uses of the computing pyramid that result from this dedicated funding are best offset by the recommendations we make for other elements in the pyramid. Provision to scientists and engineers of access to leading edge supercomputer resources will continue to be a primary purpose of the Centers, but it is a means to a broader mission; to foster rapid progress in the use of HPC by scientists and engineers, to accelerate progress in usability and economy of HPC and to diffuse HPC capability throughout the technical community, including industry. The following additional components of the Center missions should be affirmed:

- Supporting computational science, by research and demonstration in the solution of significant science and engineering problems.

- Fostering interdisciplinary collaboration - across sciences and between sciences and computational science and computer science - as in the Grand Challenge programs.

- Prototyping and evaluating software, new architectures, and the uses of high speed data communications in collaboration with three groups: computer and computational scientists, disciplinary scientists exploiting HPC resources, the HPC industry, and business firms exploring expanded use of HPC.

- Training and education, from post-docs and faculty specialists to introduction of less experienced researchers to HPC methods, to collaboration with state and regional HPC centers working with high schools, community colleges, colleges, and universities.

The role of a Supercomputer Center should, therefore, continue to be primarily one of a facilitator, pursuing the goals just listed by making the hardware and human resources available to computational scientists, who themselves are intellectual leaders. In this way the Centers will participate in leadership but will not necessarily be its primary source. With certain notable exceptions, intellectual leadership in computational science has come from scientists around the country who have at times used the resources available at the Centers. This situation is unlikely to change nor should it change. It would be unrealistic to place this type of demand on the Supercomputer Centers and it would certainly not be in the successful tradition of American science.

The Supercomputer Centers facilitate interdisciplinary collaborations because they support users from a variety of disciplines, and are aware of their particular strengths. The Centers have been deeply involved in nucleating Grand Challenge teams, and particularly in reaching out to bring computer scientists together with computational scientists. Visualization, for example, is no longer just in the realm of the computational scientist; experimentalists use the same tools for designing and simulating experiments in advance of actual data generation. This common ground should not be separated from the enabling technologies which have made this work possible. Rather high performance computing and the new science it has enabled have seeded advances that would not have happened any other way.

ALLOCATION OF CENTER HPC RESOURCES TO INVESTIGATORS

Recommendation C-4: The NSF should review the administrative procedures used to allocate Center resources, and the relationship of this process to the initial funding of the research by the disciplinary program offices, to ensure that the burden on scientists applying for research support is minimized, when that research also requires access to the facilities of the Centers, or perhaps access to other elements of the HPC pyramid that will be established pursuant to our recommendations. How-

ever we believe the NSF should continue to provide HPC resources to the research community through allocation committees that evaluate competitively proposals for use of Center resources.[19]

At the present time, the allocation of resources in the Supercomputer Centers for all users is handled by requiring principal investigators to submit annual proposals to a specified Center for access to specific equipment. The NSF should not require a duplicate peer review of the substantive scientific merit of the proposed scientific investigation, first by disciplinary program offices, and then again by the Center Allocation Committees. For this reason, it is proposed that the allocation of supercomputer time be combined with the allocation of research funds to the investigator.

Although this panel is not in a position to give administrative details of such a procedure, it is suggested that requests for computer time be attached to the original regular NSF proposal, with (a) experts in computational science included among peer reviewers, or, (b) that portion of the proposal be reviewed in parallel by a peer review established by the Centers. In either case only one set of peer reviewers should evaluate scientific merits, and only one set of reviewers should determine that the research task is being formulated properly for use of HPC resources.

Second, we recommend that the Centers collectively establish the review and allocation mechanism, so that while investigators might express a preference for a particular computer or Center for their work, all Centers facilities would be in the pool from which each investigator receives allocations.

We recognize, of course, that the specific allocation of machine time often cannot be made at the time of the original proposal for NSF research support, since in some cases the work has not progressed to the point that the mathematical approach, algorithms, etc., are available for Center experts to evaluate and translate into estimates of machine time. Nor is the demand function for facilities known at that time.

EDUCATION AND TRAINING

Recommendation C-5: The NSF should give strong emphasis to its education mission in HPC, and should actively seek collaboration with state-sponsored and other HPC centers not supported primarily on NSF funding. Supercomputing regional affiliates should be candidates for NSF support, with education as a key role. HPC will also figure in the Administration's industrial extension program, in which the states have the primary operational role.

The serious difficulties associated with the use of parallel computers pose a new training burden. In the past it was expected that individual investigators would port their code to new computers and this could usually be done with limited effort. This is no longer the case. The Supercomputer Centers should see their future mission as providing direct aid to the rewriting of code for parallel processors.

Computational science is proving to be an effective way to generate new knowledge. As part of its basic mission, NSF needs to teach scientists, engineers, mathematicians, and even computer scientists how high performance computing can be used to produce new scientific results. The role of the Supercomputer Centers is critical to such a mission since the Centers have expertise on existing hardware and software systems, modelling, and algorithms, as well as knowledge of useful high performance comput-

[19]For NSF funded investigators, allocation committees at Supercomputer Centers should evaluate requests for HPC resources only on the appropriateness of the computational plans, choice of machine, and amount of resource requested. Centers should rely on disciplinary program office determinations of scientific merit, based on their peer review. In this way a two level review of the merits of the science is avoided. A further simplification might be for the application for computer time at the Centers to be included in the original disciplinary proposal, and forwarded to the Centers when the proposal is approved. For non-NSF funded investigators an alternative form of peer review of the research is required.

ing application packages, awareness of trends in high performance computing and requisite staff.

D. NSF AND THE NATIONAL HPC EFFORT; RELATIONSHIPS WITH STATES

Recommendation D-1: We recommend that the National Science Board urge OSTP to establish an advisory committee representing the states, HPC users, NSF Supercomputer Centers, computer manufacturers, computer and computational scientists (similar to the Federal Networking Council's Advisory Committee), which should report to HPCCIT. A particularly important role for this body would be to facilitate state-federal planning related to high performance computing.

Congress required advisory committee reporting to the PMES, but the committee has not yet been implemented. The committee we propose would provide policy level advice and coordination with the states. The main components of HPCC are networking and HPC, although the networks seem to be receiving priority attention. The Panel believes it is important to continue to emphasize the importance of ensuring adequate compute power in the network to support the National Information Infrastructure applications. We also believe that as participation in HPC continues to broaden through initiatives by the states and by industry, the NSF (and other federal agencies) should encourage their collaboration in the national effort.

The Coalition of Academic Supercomputer Centers (CASC) was founded in 1989 to provide a forum to encourage support for high performance computing and networking. Unlike the FCCSET task force, CASC is dependent on others to bring the money to support high performance computing - usually their own State government or university. The result is a valuable discussion group for exchanging information and developing a common agenda and CASC should be encouraged. However, CASC is not a substitute for a more formal federal advisory body.

This recommendation is consistent with a recent Carnegie Commission Report entitled "Science, Technology and the States in America's Third Century," which recommends the creation of a system of joint advisory and consultative bodies to foster federal-state exchanges and to create a partnership in policy development, especially for construction of national information infrastructure and provision of services based on it. Because of the importance of high performance computing to future economic development, we need a new balance of cooperation between federal and state government in this area, as in a number of others.

Appendix A

MEMBERSHIP OF THE BLUE RIBBON PANEL ON HIGH PERFORMANCE COMPUTING

Lewis Branscomb, John F. Kennedy School of Government, Harvard University (Chairman)

Lewis Branscomb is a physicist, formerly chairman of the National Science Board (1980-1984) and Chief Scientist of IBM Corp. (1972-1986).

Theodore Belytschko, Department of Civil Engineering, Northwestern University

Ted Belytschko's research interests are in computational mechanics, particularly in the modeling of nonlinear problems, such as failure, crashworthiness, and manufacturing processes.

Peter R. Bridenbaugh, Executive Vice President - Science, Engineering, Environment, Safety & Health, Aluminum Company of America

Peter Bridenbaugh serves on a number of university advisory boards, and is a member of the National Academy of Engineering's Industrial Ecology Committee. He also serves on the NSF Task Force 1994 Budget Committee and is a Fellow of ASM International.

Theresa Chay, Professor, Department of Biological Sciences, University of Pittsburgh

Teresa Chay's research interests are in modelling biological phenomena such as nonlinear dynamics and chaos theory in excitable cells, cardiac arrhythmias by bifurcation analysis, mathematical modeling for electrical activity of insulin secreting pancreatic B-cells and agonist-induced cytosolic calcium oscillations, and elucidation of the kinetic properties of ion channels.

Jeff Dozier, Center for Remote Sensing, University of California, Santa Barbara

Jeff Dozier, University of California, Santa Barbara, is a hydrologist and remote sensing specialist. From 1990-1992 he was Senior Project Scientist on NASA's Earth Observing System.

Gary Grest, Exxon Corporate Research Science Laboratory

Gary Grest's research interest are in the areas of computational physics and material science, recently emphasizing the modeling the properties of polymers and complex fluids.

Edward Hayes, Vice President for Research, Ohio State University

Edward F. Hayes is a computational chemist, formerly Controller and Division Director for Chemistry at NSF.

Barry Honig, Department of Biochemistry and Molecular Biology, Columbia University

Barry Honig's research interests are in theoretical and computational studies of biological macromolecules. He is an associate editor of the Journal of Molecular Biology and is a former president of the Biophysical Society (1990-1991).

Neal Lane, Provost, Rice University (resigned from the Panel July 1993)

William A. Lester, Jr., Professor and Associate Dean, Department of Chemistry, University of California, Berkeley

William A. Lester, Jr., is a theoretical chemist, formerly Director of the National Resource for Computation in Chemistry (1978-81) and Chairman of the NSF Joint Advisory Committees for Advanced Scientific Computing and Networking and Communications Research and Infrastructure (1987).

Gregory McRae,Professor, Department of Chemical Engineering, MIT

James Sethian, Professor, University of California at Berkeley

James Sethian is an applied mathematician in the Mathematics Department at the University of California at Berkeley and in the Physics Division of the Lawrence Berkeley Laboratory.

Burton Smith, Tera Computer Company

Burton Smith is Chairman and Chief Scientist of Tera Computer Company, a manufacturer of high performance computer systems.

Mary Vernon, Department of Computer Science, University of Wisconsin

Mary Vernon is a computer scientist who has received the NSF Presidential Young Investigator Award and the NSF Faculty Award for Women Scientists and Engineers in recognition of her research in parallel computer architectures and their performance.

Appendix B

NSF AND HIGH PERFORMANCE COMPUTING: HISTORY AND ORIGIN OF THIS STUDY

Introduction

This report of the Blue Ribbon Panel on High Performance Computing follows a number of separate, but related, activities in this area by the NSF, the computational science community, and the Federal Government in general acting in concert through the Federal Coordinating Committee on Science, Engineering, and Technology. The Panel's findings and recommendations must be viewed within this broad context of HPC. This section provides a description of the way in which the panel has conducted its work and a brief overview of the preceding accomplishments which were used as the starting point for the Panel's deliberations.

The Origin of the Present Panel and Charter

Following the renewal of four of the five NSF Supercomputer Centers in 1990, the National Science Board (NSB) maintained an interest in the Centers' operations and activities. Given the national scope of the Centers, and the possible implications for them contained in the HPCC Act of 1992, the NSB commissioned the formation of a blue ribbon panel to investigate the future changes in the overall scientific environment due the rapid advances occurring in the field of computers and scientific computing. The panel was instructed to investigate the way science will be practiced in the next decade, and recommend an appropriate role for NSF to enable research in the overall computing environment of the future. The panel consists of representatives from a wide spectrum of the computer and computational science communities in industry and academia. The role expected of the Panel is reflected by its Charter :

A. Assess the contributions of high performance computing to scientific and engineering research and education, including ancillary benefits, such as the stimulus to the pace of innovation in U.S. industries and the public sector.

B. Project what hardware, software and communication resources may be available in the next five to ten years to further these advances and identify elements that may be particularly important to the development of HPC.

C. Assess the variety of institutional forms through which access to high performance computing may be gained including funding of equipment acquisition, shared access through local centers, and shared access through broad band telecommunications.

D. Project sources, other than NSF, for support of such capabilities, and potential cooperative relationships with: states, private sector, other federal agencies, and international programs.

E. Identify barriers to the development of more efficient, usable, and powerful means for applying high performance computing, and means for overcoming them.

F. Provide recommendations to help guide the development of NSF's participation in supercomputing and its relation to the federal interagency High Performance Computing and Communications Program.

G. Recommend policies and managerial structures needed to achieve NSF program goals, including clarification of the peer review procedures and suggesting appropriate processes and mechanisms to assess program effectiveness necessary for insuring the highest quality science and engineering research.

At its first meeting in January 1993, the panel approved its Charter, and established a scope of work which would allow a final report to be

presented to the NSB in Summer 1993. A large number of questions were raised amplifying the Charter's directions. Prior to its second meeting in March 1993 the Panel solicited input from the national research community; a response to the following four questions was requested.

- How would you project the emerging high performance computing environment and market forces over the next five years and the implications for change in the way scientists and engineers will conduct R&D, design and production modeling?
- What do you see as the largest barriers to the effective use of these emergent technologies by scientists and engineers and what efforts will be needed to remove these barriers? What is the proper role of government, and, in particular, the NSF to foster progress?
- To what extent do you believe there is a future role for government-supported supercomputer centers? What role should NSF play in this spectrum of capabilities?
- To what extent should NSF use its resources to encourage use of high performance computing in commercial industrial applications through collaboration between high performance computing centers, academic users and industrial groups?

Over fifty responses were received and were considered and discussed by the Panel at its March meeting. The Panel also received presentations, based on these questions, from vendors of high performance computing equipment and representatives from non-NSF supercomputer centers.

NSF's Early Participation in High Performance Computing

Although the National Science Foundation is now a major partner in the nation's high performance computing effort, this was not always the case. In the early 1970s the NSF ceased its support of campus computing centers, and by the mid-1970s there were no "supercomputers" on any campus available to the academic community. Certainly computers of this capability were available through other government agency (DoE and NASA) laboratories, but NSF did not play a role, and hence many of its academic researchers did not have the ability to perform computational research on anything other than a departmental minicomputer, thereby limiting the scope of their research.

This lack of NSF participation in the high performance computing environment began to be noted in the early 1980s with the publication of a growing number of reports on the subject. A report to the NSF Division of Physics Advisory Committee in March 1981 entitled "Prospectus for Computational Physics", edited by W. Press, identified a "crisis" in computational physics, and recommended support for facilities. Subsequent to this report a joint agency study, "Large Scale Computing in Science and Engineering", edited by P. Lax, appeared in December 1982 and acted as the catalyst for NSF's reemergence in the support of high performance computing. The Lax Report presented four recommendations for a government-wide program:

- Increased access to regularly upgraded supercomputing facilities via high bandwidth networks
- Increased research in computational mathematics, software, and algorithms
- Training of personnel in scientific computing
- R&D of new supercomputer systems

The key suggestions contained in the Lax Report were studied by an internal NSF working group, and the findings were issued in July 1983 as "A National Computing Environment for Academic Research", a report edited by M. Bardon and K. Curtis. The report studied NSF supported scientists' needs for academic computing, and validated the conclusions of the Lax Report for the NSF supported research community. The findings of Bardon/Curtis reformulated the four recommendations of the Lax Report into a six point implementation plan for the NSF. Part of this action plan was a recommendation to establish ten academic supercomputer centers.

The immediate NSF response was to set up a means for academic researchers to have access, at existing sites, to the most powerful computers of the day. This was an interim step prior to a solicitation for the formation of academic supercomputer centers directly supported by the NSF. By 1987, five NSF Supercomputer Centers had been established, and all had completed at least one year of operation.

During this phase the Centers were essentially isolated "islands of supercomputing" whose role was to provide supercomputer access to the academic community. This aspect of the Centers' activities has changed considerably. The NSF concept of the Centers' activities was mandated to be much broader, as indicated by the Center's original objectives:

- Access to state of the art supercomputers
- Training of computational scientists and engineers
- Stimulate the U.S. supercomputer industry
- Nurture computational science and engineering
- Encourage collaboration among researchers in academia, industry and government

In 1988-1989 NSF conducted a review to determine whether support was justified beyond 1990. In developing proposals, the Centers were advised to increase their scope of responsibilities. Quoting from the solicitation:

> "To insure the long term health and value of a supercomputer center, an intellectual environment, as well as first class service, is necessary. Centers should identify an intellectual component and research agenda".

In 1989 NSF approved continuation through 1995 of the Cornell Theory Center, the National Center for Supercomputing Applications, the Pittsburgh Supercomputing Center, and the San Diego Supercomputer Center. Support for the John von Neumann Center was not continued.

The Federal High Performance Computing and Communications Initiative

At the same time the NSF Supercomputer Centers were beginning the early phases of their operations the Federal Coordinating Committee for Science, Engineering, and Technology began a study in 1987 on the status and direction of high performance computing, and its relationship to federal research and development. The results were "A Research and Development Strategy for High Performance Computing" issued by the Office of Science and Technology Policy (OSTP) in November 1987, followed in September 1989 by another OSTP document "The Federal High Performance Computing Program". These two reports set the framework for the inter-governmental agency cooperation on high performance computing which led to the High Performance Computing and Communications (HPCC) Act of 1991.

HPCC focuses on four integrated components* of computer research and applications which very closely echo the Lax Report conclusions:

- High Performance Computing Systems - technology development for scalable parallel systems to achieve teraflop speed
- Advanced Software Technology and Algorithms - generic software and algorithm development to support Grand Challenge projects, including early access to production scalable systems
- National Research and Education Network - to further develop the national network and networking tools, and to support the research and development of gigabit networks
- Basic Research and Human Resources - to support individual investigator research on fundamental and novel science and to initiate activities to significantly increase the pool of trained personnel

*At the time of writing this Report, a fifth component, entitled, Information Infrastructure, Technology, and Applications is being defined for inclusion in the HPCC Program

With this common structure across all the participating agencies, the Program outlines each agency's roles and responsibilities. NSF is the lead agency in the National Research and Education Network, and has major roles in Advanced Software Technology and Algorithms, and in Basic Research and Human Resources.

The Sugar Report

After the renewal of the four NSF Supercomputer Centers the NSF Division of Advanced Scientific Computing recognized that the computing environment within the nation had changed considerably from that which existed at the inception of the Centers Program. The Division's Advisory Committee was asked to survey the future possibilities for high performance computing, and report back to the Division. Two workshops were held in the Fall of 1991 and Spring of 1992. Thirty one participants with expertise in computational science, computer science and the operation of major supercomputer centers were involved.

The final report, edited by R. Sugar of the U. of California at Santa Barbara, recommended future directions for the Supercomputer Centers Program which would "enable it to take advantage of these (HPCC) opportunities and to meet its responsibilities to the national research community". The committee's recommendations can be summarized as:

- Decisions and planning by the Division need to be made in a programmatic way, rather than on an individual Center by Center basis - the meta-center concept provides a vehicle for this management capability which goes beyond the existing Centers.
- Access to stable computing platforms (currently vector supercomputers) needs to be augmented by access to state of the art technology (currently massively parallel computers) - but, the former cannot be sacrificed to provide the latter
- The Supercomputer Centers can be focal points for enabling collaborative efforts across many communities - computational and computer science, private sector and academia, vendors and academia.

Appendix C

TECHNOLOGY TRENDS and BARRIERS to FURTHER PROGRESS

BACKGROUND:

What is the state of the HPC industry here and abroad? What is its prognosis?

The high performance computer industry is in a state of turmoil, excitement and opportunity. On the one hand, the vector multiprocessors manufactured by many firms, large and small, have continued to improve in capability over the years. These systems are now quite mature, as measured by the fact that delivered performance is a significant fraction of the theoretical peak performance of the hardware, and are still the preferred platform for many computational scientists and engineers. They are the workhorses of high performance computing today and will continue in that role even as alternatives mature.

On the other hand, dramatic improvements in microprocessor performance and advances in microprocessor-based parallel architectures have resulted in "massively parallel" systems that offer the potential for high performance at lower cost.[20] For example, $10 million in 1993 buys over 40 gigaflops peak processing power in a multicomputer but only 5 gigaflops in a vector multiprocessor. As a result, increasing numbers of computational scientists and engineers are turning to the highly parallel systems manufactured by companies such as Cray Research Inc., IBM, Intel, Kendall Square, Thinking Machines Inc., MasPar, and nCUBE.

Microprocessor performance has increased by 4X every three years, matching the rate of integrated circuit logic density improvement as predicted by Moore's law. For example, the microprocessors of 1993 are around 200 times faster than those of 1981. By contrast, the clock rates of vector processors have improved much more slowly; today's fastest vector processors are only five or six times faster than 1976's Cray 1. Thus, the performance gap between these two technologies is quickly disappearing in spite of other performance improvements in vector processor architecture.

Although microprocessor-based massively parallel systems hold considerable promise for the future, they have not yet reached maturity in terms of ease of programming and ability to deliver high performance routinely to large classes of applications. Unfortunately, the programming technology that has evolved for the vector multiprocessors does not directly transfer to highly parallel systems. New mechanisms must be devised for high performance communication and coordination among the processors. These mechanisms must be efficiently supported in the hardware and effectively embodied in programming models.

Currently, vendors are providing a variety of systems based on different approaches, each of which has the potential to evolve into the method of choice. Vector multiprocessors support a simple shared memory model which demands no particular attention to data arrangement in memory. Many of the currently available highly parallel architectures are based on the "multicomputer" architecture which provides only a message-passing interface for inter-processor communication. Emerging architectures, including the Kendall Square KSR-1 and systems being developed by Convex, Cray Research, and Silicon Graphics, have shared address spaces with varying degrees

---

[20]Note that the higher cost of vector machines is partly caused by their extensive use of static memory chips for main memory and the interconnection networks they use for high shared-memory bandwidth. These attributes contribute to increased programmability and the realization of a high fraction of peak performance on user applications. Realized performance on MPP machines is still uncertain. A comparison of today's vector machines versus MPP systems based on realized performance per dollar reveals much less difference in cost-performance than comparisons based on peak performance.

of hardware support and different refinements of the shared memory programming model. These computers represent a compromise in that they offer much of the programming simplicity of shared memory yet still (at least so far) require careful data arrangement to achieve good performance. (The data parallel language on the CM-5 has similar properties.) A true shared memory parallel architecture, based on mechanisms that hide memory access latency, is under development at Tera Computer.

The size of the high performance computer market worldwide is about $2 billion (excluding sales of the IBM add-on vector hardware), with Cray Research accounting for roughly $800 million of it. IBM and Fujitsu are also significant contributors to this total, but most companies engaged in this business have sales of $100 million or less. Some companies engaged in high speed computing have other, larger sources of revenue (IBM, Fujitsu, Intel, NEC, Hitachi); other companies both large (Cray Research) and small (Thinking Machines, Kendall Square, Meiko, Tera Computer) are high performance computer manufacturers exclusively. There are certainly more companies in the business than can possibly be successful, and no doubt there are new competitors that will appear. Helping to sustain this high level of competitive innovation should be an important objective for NSB policy in HPC.

FINDINGS

Where is the hardware going to be in 5 years? What will be the performance and cost of the most powerful machines, the best workstations, the mid-range computers?

The next five years will continue to see improvements in hardware price/performance ratios. Since microprocessor speeds now closely approach those of vector processors, it is unclear whether microprocessor performance improvement can maintain its current pace. Still, as long as integrated circuits continue to quadruple in density but only double in cost every three years we can probably expect a fourfold price/performance improvement in both processors and memory by 1998. Estimating in constant 1993 dollars, the most powerful machines ($50 million) will have peak performance of nearly a teraflop[21]; mini-supercomputers ($1 million) will advertise 20 gigaflops peak performance; workstations ($50,000) will approach 1 gigaflops, and personal computers ($10,000) will approach 200 megaflops.[22]

During this period, parallel architectures will continue to emerge and evolve. Just as the CM-5 represented a convergence between SIMD and MIMD parallel architectures and brought about a generalization of the data-parallel programming model, it is likely that the architectures will continue to converge and better user-level programming models will continue to emerge. These developments will improve software portability and reduce the variety of architectures that are required for computational science and engineering research, although there will likely still be some diversity of approaches at the end of this 5-year horizon. Questions that may be resolved by 1998 include:

- Which varieties of shared memory architecture provide the most effective tradeoff between hardware simplicity, system performance, and programming convenience? and
- What special synchronization mechanisms for processor coordination should be supported in the hardware?

Most current systems are evolving in these directions, and answers to the issues will provide a more stable base for software efforts. Furthermore, much of the current computer science research in shared memory architectures is looking for cost-effective hardware support that can be im-

[21]One teraflop is 1000 gigaflops or 10^{12} floating point instructions per second.

[22]Spokesman from Intel, Convex and Silicon Graphics in addressing the panel all made even higher estimates than this.

plemented in multiprocessor workstations that are interconnected by general-purpose local area networks. Thus, technology from high performance parallel systems may be expected to migrate to workstation networks, further improving the capabilities of the ... systems to deliver high-performance computing to particular applications. It is possible that in the end the only substantial difference between the supercomputers of tomorrow and the workstation networks of tomorrow will be the installed network bandwidth.[23]

Where is the software/programmability going to be in 5 years? What new programming models will emerge for the new technology? How transparent will parallel computers be to users?

While the architectural issues are being resolved, parallel languages and their compilers will need to continue to improve the programmability of new high performance computer systems. Implementations of "data parallel" language dialects like High Performance Fortran, Fortran D, and High Performance C will steadily improve in quality over the next five years and will simplify programming of both multi-computers and shared address systems for many applications. For the applications that are not helped by these languages, new languages and programming models will emerge, although at a slower pace. Despite strong efforts addressing the problem from the language research community, the general purpose parallel programming language is an elusive and difficult quarry, especially if the existing Fortran software base must be accommodated, because of difficulties with the correct and efficient use of shared variables.

Support tools for software development have also been making progress, with emphasis on visualization of a program's communication and synchronization behavior. Vendors are increasingly recognizing the need for sophisticated performance tuning tools, with most now developing or beginning to develop such tools for their machines. The increasing number of computer scientists who are also using these tools could lead to even more rapid improvement in the quality and usability of these support tools. Operating systems for high performance computers are increasingly ill suited to the demands placed on them. Virtualization of processors and memory often leads to poor performance, whereas relatively fixed resource partitioning produces inefficiency, especially when parallelism within the application varies. High performance I/O is another area of shortfall in many systems, especially the multicomputers. Research is needed in nearly every aspect of operating systems for highly parallel computers.

What market forces or technology investments drive HPC technologies and products?

Future high performance systems will continue to be built using technologies and components built for the rest of the computer industry. Since integrated circuit fabrication facilities now represent billion dollar capital investments, integrated circuits benefit from very large scale economies; accordingly it has been predicted that only mass-market microprocessors will prove to have acceptable costs in future high performance systems. Certainly current use of workstation microprocessors such as Sparc, Alpha and the RS-6000 chips suggests this trend. Even so the cost of memory chips is likely to be a major factor in the costs of massively parallel systems, which require massive amounts of fast memory. Thus the integrated circuit technology available for both custom designs and industry standard processors will increasingly be driven by the requirements of much larger markets, including consumer electronics.

The health of the HPC vendors and the structure of their products will be heavily influenced by demand from industrial customers. Business applications represents the most rapidly growing

[23]While parallel architectures mature, vector multiprocessors will continue to evolve. Scaling to larger numbers of processors ultimately involves solving the same issues as for the microprocessor-based systems.

market for HPC products; they have much higher potential growth than government or academic uses. Quite apart from NSF's obligation to contribute to the nation's economic health through its research activities, this fact motivates the importance of cooperation with industry users in expanding HPC usage. This reality means that NSF should be attentive to the value of throughput as a figure of merit in HPC systems (in contrast with turnaround time which academic researchers usually favor), as well as the speed with which large volumes of data can be accessed. Industry won't put up with a stand-alone, idiosyncratic environment.

How practical will be the loose coupling of desktop workstations to aggregate their unused compute power?

Networks of workstations will become an important resource for the many computations that perform well on them. The probable success of these loosely coupled system will inevitably raise the standard for communication capabilities in the multicomputer arena. Many observers believe that competition from workstation networks on one side and shared address space systems on the other will drive multi-computers from the scene entirely; in any event, the network bandwidth and latency of multi-computers must improve to differentiate them from workstation networks. Many large institutions have 1000 or more workstations already installed; the utilization rate of their processors on a 24 hour basis is probably only a few percent. An efficient way to use the power of such heterogeneous networks would be more financially attractive. It will, however, raise serious question about security, control, virus-prevention, and accounting programs.

Are there some emerging HPC technologies of interest other than parallel processing? What is their significance?

Neural networks have recently become popular and have been successfully applied to many pattern recognition and classification problems. Fuzzy logic has enjoyed an analogous renaissance. Technologies of this sort are both interesting and important in a broad engineering context and also are having impact on computational science and engineering. Machine learning approaches, such as neural networks, are most appropriate in scientific disciplines where there is insufficient theory to support accurate computer modeling and simulation.

How important are simulation and visualization capabilities?

Simulation will play an ever increasing role in science and engineering. Much of this work will be able to be carried out on workstations or intermediate-scale systems, but it will continue to be appropriate to share the highest performance systems (and the expertise in using them) on a national scale, to accomplish large simulations within human time scales. Smaller configurations of these machines should be provided to individual research universities for application software development and research that involves modifying the operating system and/or hardware.

Personal computer capabilities will improve, and visualization on the desktop will become more routine. Scientists and engineers in increasing numbers will need to be equipped with visualization capabilities. The usefulness of high performance computing relies on these systems because printed lists of numbers (or printed sheaves of pictures, for that matter) are increasingly unsatisfactory as an output medium, even for moderately sized simulations.

BARRIERS TO CONTINUED RAPID PROGRESS

What software and/or hardware "inventions" are needed? Who will address meeting these needs?

The most important impediment to the use of new highly parallel systems has been the difficulty of programming these machines and the wide variation that exists in communication capabilities across generations of machines as well as among

the machines in a given generation. Application software developers are understandably reluctant to re-implement their large scale production codes on multi-computers, when significant effort is required to port the codes across parallel systems as they evolve. In theory, any programming model can be implemented (by appropriate compilers) on any machine. However, the inefficiency of certain models on certain architectures is so great as to render them impractical.[24] What is needed in high performance computing is an architectural consensus and a simple model to summarize and abstract the machine interface to allow compilers to be ported more easily across systems, facilitating the portability of application programs. Ideally, the consensus interface should efficiently support existing programming models (even the multi-computers have created their own dusty decks), as well as more powerful models. Considerable research in the computer science community is currently devoted to these issues. It is unlikely that the diversity of programming models will decrease within the next five years, but it is likely that models will become more portable.

How important will be access to data, data management?

Besides needing high performance I/O, some fields of computational science need widely distributed access to data bases that are extremely large and constantly growing. The need is particularly felt in the earth and planetary sciences, although the requirements are also great in cellular biology, high energy physics, and other disciplines. Large scale storage hierarchies and the software to manage them must be developed, and means to distribute the data nationally and internationally are also required. Although this area of high performance computing has been relatively neglected in the past, these problems are now receiving significantly more attention.

[24]For example, it is not practical to implement data-parallel compliers on the Intel iPSC/860.

ROLES FOR GOVERNMENT AGENCIES

What should government agencies (NSF, DoD, DoE) do to advance HPC beyond today's state of the art? What more might they be doing?

The National Science Foundation plays several critical roles in advancing high performance computing. First, NSF's support of basic research and human resources in all areas of science and engineering (and particularly in mathematics, computer science and engineering) has been responsible for many of the advances in our ability to successfully tackle grand challenge problems. The Supercomputer Centers and the NSFnet have been essential to the growth of high performance computing as a basic paradigm in science and engineering. These efforts have been successful and should be continued.

However, NSF has done too little in supporting computational engineering in the computer science community. For example, the NSF Supercomputer Centers were slow in providing experimental parallel computing facilities and are currently not responding adequately to integrating emerging technologies from the computer science community. Although this situation is gradually changing, the pace of the change should be accelerated.

Many advances in high performance computer systems have been funded and encouraged by the Advanced Research Projects Agency (ARPA), the major supporter of large scale projects in computer science and engineering research and development in the US. ARPA has been charged by Congress to champion "dual use" technology; in so doing it is addressing many of the needs of computational science and engineering, even in the mathematical software arena, that are common to defense and commercial applications, and the science that underlies both.

The Department of Energy has traditionally provided substantial support to computer science and engineering research within its national laboratories and at universities with strong impetus being provided by national defense require-

ments and resources. More recently, the focus has shifted to the high performance computing and communications needs of the unclassified Energy Research programs within DoE. The National Energy Research Supercomputer Center (NERSC) and the Energy Sciences Network (ESnet) provide production services similar to the NSF supercomputer centers and the NSFnet. Under the DoE HPCC component, "grand challenge" applications are supported at NERSC and also at two High Performance Computing Research Centers (HPCRCs) which offer selected access for grand challenge applications to leading edge parallel computing machines. DoE also sponsors a variety of graduate fellowships in the computational sciences. The computational science infrastructure and traditions of DoE remain sound; however, the ability of the Department to advance the state-of-the-art in high performance computing systems will be paced by its share of the funding available through the Federal High Performance Computing Initiative or through Defense conversion funds.

The Department of Commerce has not been a significant source of funds for computer system research and development since the very early days of the computer industry when the National Bureau of Standards built one of the first digital computers. NBS has been an important factor in supporting standards development, particularly for the Federal Information Processing Standards issued by GSA. The expanded role of the National Institute for Standards and Technology (as NBS is now called) under the Clinton administration may include this kind of activity, especially when industrial participation is a desired component.

NASA is embarked on a number of projects of potential importance, especially in the development of a shared data system for the global climate change program, which will generate massive amounts of data from the Earth Observing Satellite Program.

What is role of NSF computer science and applied mathematics research program? Is it relevant to the availability of HPC resources in a five year time span?

Investments in mathematics and computer science research provide the foundation for attacking today's problems in high performance computing and must continue. NSF continues to be the primary U.S. source of funds for mathematics and computer science research within the scope of what one or two investigators and several graduate assistants can do. Many fundamental advances in algorithms, programming languages, operating systems, and computer architecture have been NSF funded. This mission has been just as vital as ARPA's and is complementary to it.

Among the largest barriers to the effective use of emergent computing technologies are parallel architectures from which it is relatively easy to extract peak performance, system software (operating systems, databases, compilers, programming models) to take advantage of these architectures, parallel algorithms, mathematical modeling, and efficient and high order numerical techniques. These are core computational mathematics and computer science/engineering research issues, many of which are best tackled through NSF's traditional peer-reviewed model. NSF should increase its support of this work. Increased investment in basic research and human resources in mathematics and computer science/engineering could significantly accelerate the pace of HPC technology development. In addition, technology transfer can be increased by supporting a new scale of research not currently being funded by any agency: small teams with annual budgets in the $250K - $1M range. These projects were once supported by DoE, and also indirectly by ARPA at the former "block grant" universities.[25] These

[25]ARPA made an enormous contribution to the maturing of computer science as a discipline in U.S. universities by consistently funding research at about $1 million per year at MIT, Carnegie Mellon, Stanford University and Berkeley. At this level of consistent support these universities could build up a critical mass of faculty and trained the next generation of faculty leadership for departments being set up at every substantial research university. This targeted investment played a role in computer science not unlike what NSF has done in computational science at the four Centers.

enterprises are now generally too small for ARPA and seem to be too big for current NSF budget levels in computer science. A project of this scale could develop and release an innovative piece of software to the high performance computing community at large, or build a modest hardware prototype as a stepping stone to more significant funding. A project of this scale would also allow multi-disciplinary collaboration, either within mathematics and computer science/engineering (architecture, operating systems, compilers, algorithms). and related disciplines (astronomers or chemists working with computer scientists and mathematicians interested in innovative programming or architectural support for that problem domain).

Appendix D

Figure 1
Supercomputer Usage by State: Fiscal Year 1992
Four National Supercomputer Centers

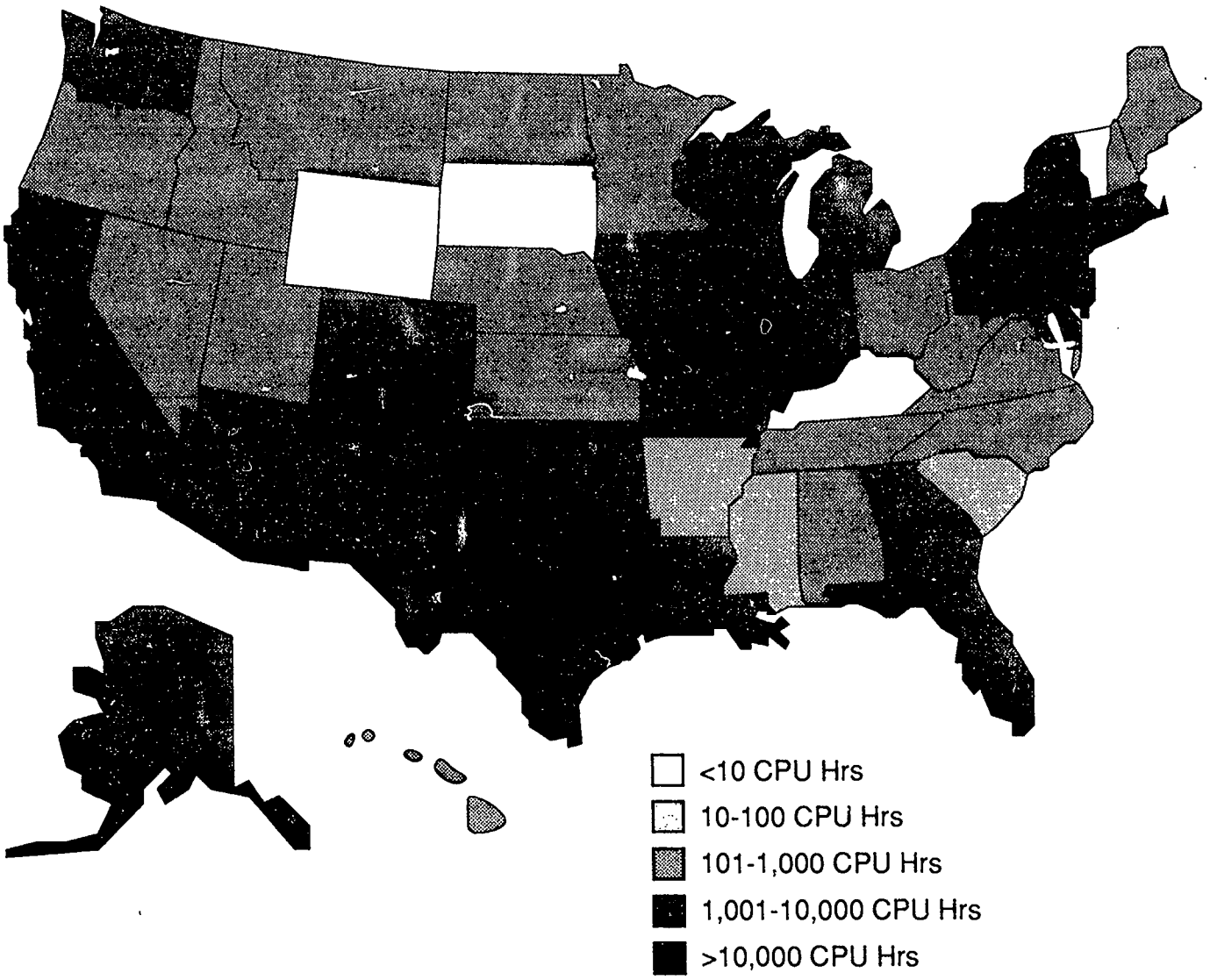


Figure 2
Trends in user status at four National Supercomputer Centers
(Data from use of vector supercomputers only)

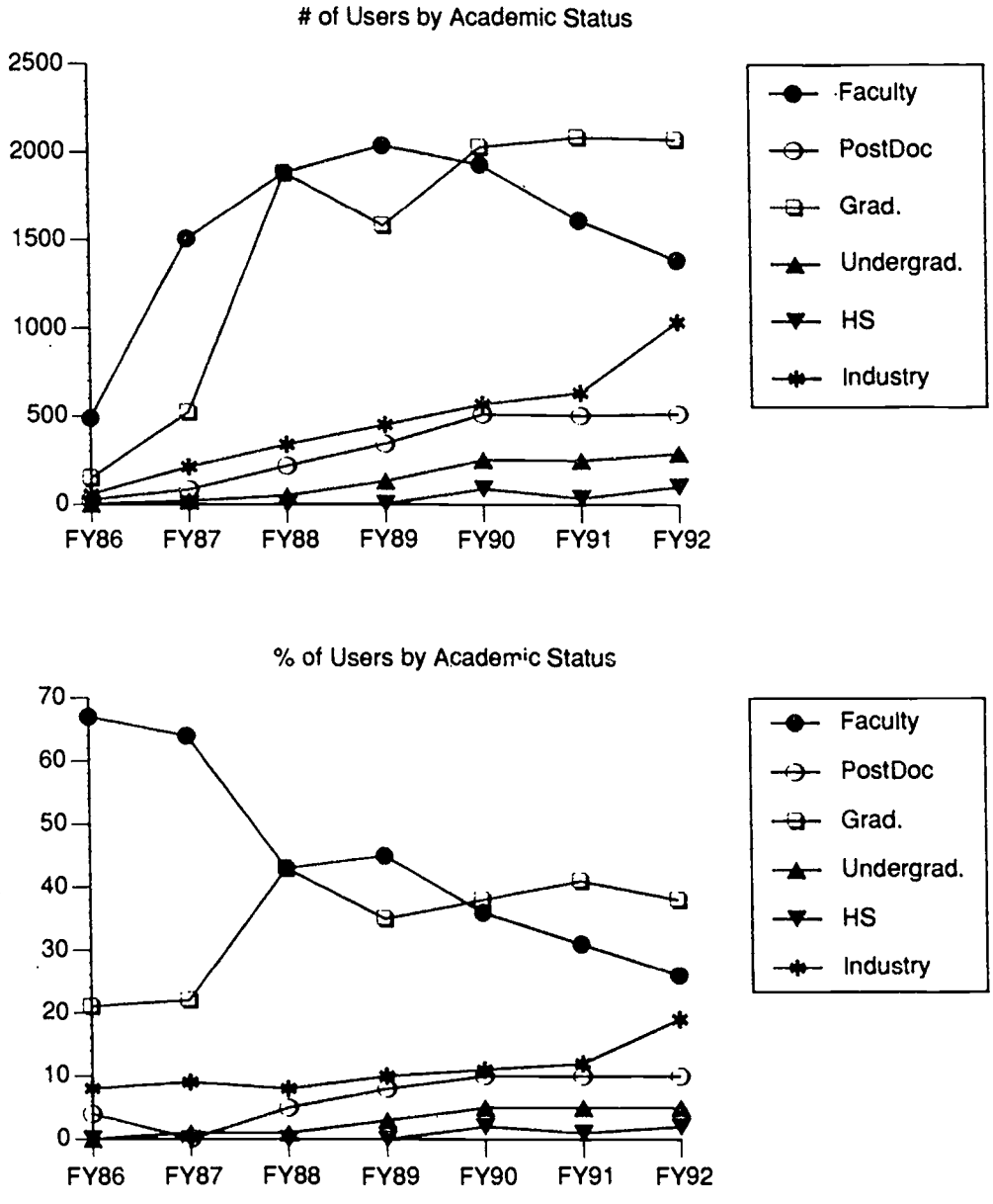
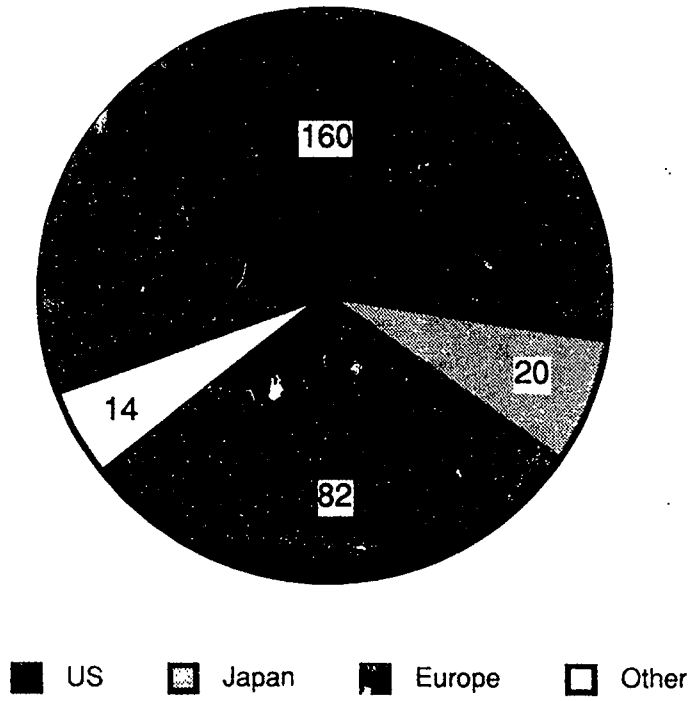


Figure 3
Installed Supercomputer Base
(U.S. Vendors Only)



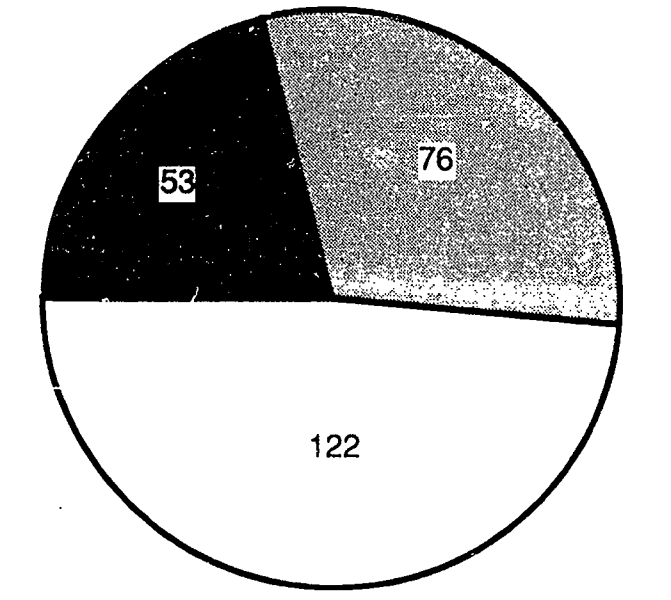
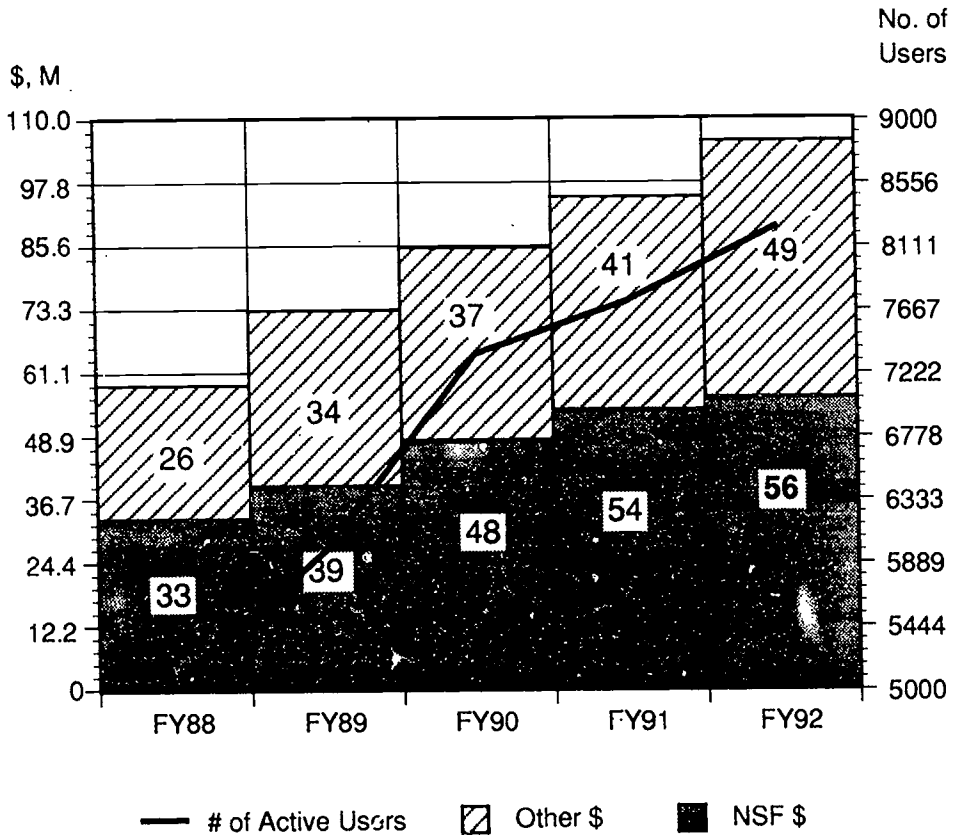


Figure 4
Trend in NSF Supercomputer Center Leverage and Growth in number of users



Appendix E

REVIEW AND PROSPECTUS OF COMPUTATIONAL AND COMPUTER SCIENCE AND ENGINEERING

Personal Statements by Panel Members

Computational Mechanics and Structural Analysis

by Theodore Belytschko

High performance computing has had a dramatic impact on structural analysis and computational mechanics, with significant benefits for various industries. The finite element method, which was developed at aerospace companies such as Boeing in the late 1950's and subsequently at the universities, has become one of the key tools for the mechanical design of almost all industrial products, including aircraft, automobiles, power plants, packaging, etc. The original applications of finite element methods were primarily in linear analysis, which are useful for determining the behavior of engineering products in normal operating conditions. Most linear finite element analyses are today performed on workstations , except for problems with the order of 1 million unknowns.

Supercomputers are used primarily for nonlinear analysis, where they replace prototype testing. One rapidly developing area has been automobile crashworthiness analysis, where models of automobiles are used to design for occupant safety and for features such as accelerometer placement for a air bag deployment. The models which are currently used are generally on the order of 100,000 to 250,000 unknowns, and even on the latest supercomputers such as the CRAY C90 require on the order of 10 to 20 hours of computer time. Nevertheless these models are still often too coarse to permit true prediction and hence they must be tuned by tests.

Such models have had a tremendous impact on reducing the design time for automobiles, since they eliminate the need for building numerous prototypes. Almost all major automobiles manufacturers have undertaken extensive programs in crashworthiness simulations by computer on high performance machines, and many manufacturers have bought supercomputers almost expressly for crashworthiness simulation.

Because of the increasing concern with safety among manufacturers of many other products, nonlinear analysis are also emerging in many other industries: the manufacturer of trucks and construction equipment, where the product must be certified for safety in various accidents such as overturning or impact due to falling construction equipment; railroad car safety; the safety of aircraft, where recently the FAA have undertaken programs to simulate the response of aircraft to small weapons so that damage from such explosives can be minimized. Techniques of this type are also being used the analysis the safety of jet engines due to bird impact, the containment of fragments in case of jet engine failure, and bird impact on aircraft canopies. In several cases, NSF Supercomputer Centers have introduced industry to the potentials of this type of simulation. In all of these, highly nonlinear analysis which require on the order of 10x floating point operations for a simulation must be made: such simulations even on today's supercomputers are still often so time consuming that decisions cannot be reached fast enough. Therefore an urgent need exists for increasing the speed with which such simulations can be made.

Nonlinear finite element analysis is also becoming increasingly important in the simulation manufacturing processes. For example, tremendous im-

provements can be made in processes such as sheet metal forming, extrusion, and machining processes if these are carefully designed through nonlinear finite element simulation. These simulations offer large cost reductions and reduce design time. Also the design of materials can be improved if computers are first used to examine how these materials fail and then to design the material so that failure is either decreased or so that the material fails in a less catastrophic manner. Such simulations require great resolution, and at the tips of cracks phenomenon at the atomic scale must be considered.

Most of the calculations mentioned above are not made with sufficient resolution because of limitations in computational power and speed. Also, important physical phenomena are omitted for reasons of expediency, and their computational modeling is not well understood. Therefore, the availability of more computational power will increase our understanding of modeling nonlinear structural response and provide industry with more effective tools for design.

Cellular and Systemic Biology

by Teresa Chay

HPC has made a great impact on a variety of biological disciplines, such as physiology, biological macromolecules, and genetics. I will discuss below three vital organs in our body where our understanding has greatly benefitted from high-performance comuting and will continue to do so in the future.

Computer Models For Vital Organs In Our Body

Although the heart, brain, and pancreas function differently in our body (i.e., the heart circulates the blood, the brain stores and transfers information, and the pancreas secretes vital hormones such as insulin), the mechanisms underlying their functioning are quite similar - "excitable" cells that are coupled electrically and chemically, forming a network.

Ion channels in the cell membranes are involved in information transfer. The ion channels receive stimuli from neighboring cells and from cells in other organs. Upon receiving stimuli some ion channels open while others close. When these channels are open, they pass ions into or out of the cells, creating an electrical difference (membrane potential) between the outside and the inside of the cell. Some of these ion channels are sensitive to the voltage (i.e., membrane potential) and others are responsive to chemical substances (e.g., neurotransmitters/ hormones).

Opening of ion channels creates the "action potential" which spreads from cell to cell, either directly or via chemical mediators. Electrical transmission and chemical transmission are interdependent in that chemical substances can influence the ionic currents and visa versa. For example, the arrival of the action potential at a presynaptic terminal may cause a release of chemical substances; in turn these chemicals can open/close the ion channels in the postsynaptic cell.

Why is high-performance computing needed? How the signals are passed from cell to cell is a nonlinear dynamical problem and can be treated mathematically by solving simultaneous differential equations. These equations involve voltage, conductance of the ionic current, and concentration of those chemical substances that influence conductances. Depending on the model, each network can be represented by a set of several million differential equations. The need for parallel processors is obvious — the organs process information just the same way as the most powerful parallel supercomputers do. Since the mechanisms

involved in these three organs are essentially the same, algorithms developed for one can be easily modified to solve for another.

Three specific areas in which high-performance computing is central are cardiac research, neural networks, and insulin secretion. These are detailed below.

Cardiac Research

It would be a great benefit to cardiac research if a realistic computer model of the heart, its valves and the nearby major vessels were to become available. With such a model the general public would be able to see how the heart generates its rhythm, how this rhythm leads to contraction, and how the contraction leads to blood circulation. Scientists, on the other hand, could study normal and diseased heart function without the limitations of using human and animal subjects. With future HPC and parallel processing, it may be possible to build a model heart without consuming too many hours of computer time. As a step toward achieving this goal, the scientists in computational cardiology have thus far accomplished the following three objectives.

1. A computer model of blood flow in the heart: Researchers have used supercomputers to design an artificial mitral valve (the valve that controls blood flow between the left atrium and left ventricle) which is less likely to induce clots. The computer simulated mitral valve has been patented and licensed to a corporation developing it for clinical use. With parallel processors, this technique is now expanded in order to construct a realistic three dimensional heart.
2. Constructing an accurate map of the electrical potential of the heart surface (epicardium): Arrhythmia in the heart is caused by a breakdown in the normal pattern of cardiac electrical activity. Many arrhythmias occur because of an abnormal tissue inside the heart. Bioengineers have been developing a technique with which to obtain the epicardial potential map from the coarse information of it that can be recorded on the surface of the body via electrocardiogram. With such a map, clinicians can accurately locate the problem tissue and remove it with a relatively simple surgical procedure instead of with drastic open-heart surgery.
3. Controlling sudden cardiac death: Sudden cardiac death is triggered by an extra heart beat. Such a beat is believed to initiate spiral waves (i.e., reentrant arrhythmias) on the main pumping muscle of the heart known as the ventricular myocardium. With HPC it is possible to simulate how this part of the heart can generate reentrant arrhythmia upon receiving a premature pulse. Computer modelling of reentrant arrhythmia is very important clinically since it can be used as a tool to predict the onset of this type of deadly arrhythmia and find a means to cure it by properly administrating antiarrhythmia drugs (instead of actually carrying out experiments on animals).

Parallel computing and development of better software will soon enable the researchers to extend their simulations to a more realistic three-dimensional system which includes the detailed geometry of ventrical muscle.

Neural Networks

Learning how the brain works is a grand challenge in science and engineering. Artificial neural nets are based largely on their connection patterns, but they have very simple processing elements or nodes (either ON or OFF). That is, a simple network consists of a layered network with an input layer (sensory receptors), one or more "hidden" layers (representing the interneurons which allow animals to have complex behaviors), and an output layer (motor neurons). Each unit in a neural net receives inputs, both excitatory and inhibitory, from a number of other units and, if the strength of the signal exceeds a given threshold, the "on-unit" sends signals to other units.

The real nervous system, however, is a complex organ that cannot be viewed simply as an artificial neural net. Neural nets are not hard-wired but are made of neurons which are connected by synap-

ses. There are at least 10 billion active neurons in the brain. There are thousands of synapses per neuron, and hundreds of active chemicals which can modify the properties of ion channels in the membrane.

With HPC and massive parallel computing, neuroscientists are moving into a new phase of investigation which focuses on biological neural nets, incorporating features of real neurons and the connectivity of real neural nets. Some of these models are capable of simulating patterns of electrical activity, which can be compared to actual neuronal activity observed in experiments. With the biological neural nets, we begin to understand the operation of the nervous system in terms of the structure, function and synaptic connectivity of the individual neurons.

Insulin Secretion

Insulin is secreted from the beta cells in the pancreas. To cure diabetes it is essential to understand how beta cells release insulin. The beta cells are located in a part of pancreas known as the islet of Langerhans. In islets, beta cells are coupled by a special channel (gap junctional channel) which connects one cell to the next. Gap junctional channels allow small ions such as calcium ions pass through from cell to cell. In the plasma of beta cells, there are ion channels whose properties change when the content of calcium ions changes. There are other types of cells in the islet which secrete hormones. These hormones in turn influence insulin secretion by altering the properties of the receptors bound in the membrane of a beta cell. Thus the study on how beta cells release insulin involves very complex non-linear dynamics. With a supercomputer it is possible to construct a model of the islet of Langerhans. With this model, researchers would learn how beta cells release insulin in response to the external signals such as glucose, neurotransmitters and hormones. They would also learn the roles of other cell types in the islet of Langerhans and how they influence the functional properties of beta cells. A model in which beta cells function as a cluster has been already constructed.

Material Science and Condensed Matter Physics

by Gary S. Grest

The impact of high performance computing on material science and condensed matter physics has been enormous. Major developments in the sixties and seventies set the stage for the establishment of computational material science as a third discipline, equal, yet distinct from analytic theory and experiment. These developments include the introduction of molecular dynamics and Monte Carlo methods to simulate the properties of liquids and solids under a variety of conditions. Density functional theory was developed to model the electron-electron interactions and pseudopotential methods to model the electron-ion interactions. These methods were crucial in computing the electronic structure for a wide variety of solids. Later, the development of path integral and Green's function Monte Carlo methods allowed one to begin to simulate quantum many-body problems. Quantum molecular dynamics which combine well-established electronic methods based on local density theory with molecular dynamics for atoms have recently been introduced. On a more macroscopic scale, computational mechanics which was discussed above by T. Belyschko was developed to study structural properties relations.

Current usage of high performance computing in material science and condensed matter physics can be broadly classified as Classical Many-Body and Statistical Mechanics, Electronic Structure and Quantum Molecular Dynamics , and Quantum Many-Body, which are discussed below.

Classical Many-Body and Statistical Mechanics

Classical statistical mechanics, where one treats a huge number of atoms collectively date back to Boltzmann and Gibbs. In these systems, quantum mechanics plays only a subsidiary role. While it is needed to determine the interaction between atoms, in practice these interactions are often replaced by phenomenologically determined pairwise forces between the atoms. This allows one to treat large ensembles of atoms, by molecular dynamics and Monte Carlo methods. Successes of this approach include insight into the properties of liquids, phase transitions and critical phenomena, crystallization of solids and compositional ordering in alloys. For systems where one needs a quantitative comparison to experiment, embedded atom methods have been developed in which empirically determined functions are employed to evaluate the energy and forces. Although the details of the electronic structure are lost, these empirical methods have been successful in giving reasonable descriptions of the physical processes in many systems in which directional bonding is not important. Theoretical work in the mid-70's on renormalization group methods, showed that a wide variety of different kinds of phase transitions could be classified according to the symmetry of the order parameter and the range of the interaction and did not depend on the details of the interaction potential. This allowed one to use relatively simply models, usually on a lattice, to study critical phenomena and phase behavior.

While the basic computational techniques used in classical many-body theory are now well established, there remain a large number of important problems in material science which only be addressed with these techniques. At present, with Cray YMP class computers, one can typically handle thousands of atoms for hundreds of picoseconds. With the next generator of massively parallel machines, this can be extended to millions of particles for microseconds. While not all problems require this large number of particles or long times, many do. Problems which will benefit from the faster computational speed typically involve either large lengths and/or long time scales. Examples include polymers and macromolecular liquids, where typical lengths scales of each molecule can be hundreds angstroms and relaxation time scales extend from microseconds and longer, liquids near their glass transition where relaxation times diverge exponentially, nucleation and phase separation which requires both large systems and long times and effects of shear. Macromolecular liquids typically contain objects of very different sizes. For example, in most colloidal suspensions, the colloid particles are hundreds of angstroms in size while the solvent is only a few angstroms. At present, the solvent molecules must be treated as a continuum background. While this allows one to study the static properties of the system, the dynamics are incorrect. Faster computer will allow us to study flocculation, sedimentation and the effects of shear on order. Non-equilibrium molecular dynamics methods have been developed to simulate particles under shear. However due to the lack of adequate computation power, simulations at present can only be carried out at unphysically high shear rates. Access to HPC will enable one to understand the origins of shear thinning and thickening in a variety of technologically important systems. While molecular dynamics simulations are inherently difficult to vectorize, recent efforts to run them on parallel computers have been very encouraging, with increases in speed of nearly a factor of 30 in comparison to the Cray YMP.

Monte Carlo simulations on a lattice remain a very powerful computational technique. Simulations of this type have been very successful in understanding critical phenomena, phase separation, growth kinetics and disordered magnetic systems. Successes include accurate determination of universal critical exponents, both static and dynamic, and evidence for the existence of a phase transition in spin glasses. Future work using massively parallel computers will be essential to understand wetting and surface critical exponents as well as systems with complex order parameters. Direct numerical integration of a set of Langevin equations that describe the nonlinear fluctuating hydrodynamics can be solved in two-dimensions on a Cray YMP class supercomputer but the exten-

sion to three dimensions requires HPC. Finally, cell automata solutions of Navier-Stokes and Boltzmann equations are a powerful method for studying hydrodynamics. All of these methods, because of their inherent locality, run very efficiently on parallel computers.

Electronic Structure and Quantum Molecular Dynamics

The ability of quantum mechanics to predict the total energy of a system of electrons and nuclei enables ones to reap tremendous benefits from quantum-mechanical calculations. Since many physical properties can be related to the total energy of a system or to differences in total energy, tremendous theoretical effort has gone into developing accurate local density functional total energy techniques. These methods have been very successful in predicting with accuracy equilibrium constants, bulk moduli, phonons, piezoelectric constants and phase-transition pressures and temperatures for a variety of materials. These methods have recently been applied to study the structural, vibrational, mechanical and other ground state properties of systems containing up to several hundred atoms. Some recent successes include the unraveling of the normal state properties of high T_c superconducting oxides, predictions of new phases of materials under high pressure, predictions of superhard materials, determination of the structure and properties of surfaces, interfaces and clusters, and calculations of properties of fullerenes and fullerites.

Particularly important are the developments of the past few years which make it possible to carry out "first principles" computations of complex atomic arrangements in materials starting from nothing more than the identities of the atoms and the rules of quantum mechanics. Recent developments in new iterative diagonalization algorithms coupled with increases in the computational efficiency of modern high performance computers have made it possible to use quantum mechanical calculations of the dynamics of systems in the solid, liquid and gaseous state. The basic idea of these methods which are known as *ab initio* methods is to minimize the total energy of the system by allowing both the electronic and the ionic degrees of freedom to relax towards equilibrium simultaneously. While ab initio methods have been around for more than a decade, only recently have they been applied to systems of more than a few atoms. Now, however, this method can be used to model a few hundred atom system and this number will increase by at least a factor of 10 within the next five years. The method has already lead to new insights into the structure of amorphous materials, finite temperature simulations of the new C60 solid, computation of the atomic and electronic structures of 7×7 reconstruction of Si(111) surface, melting of carbon and studies of step geometries on semiconductor surfaces. In the future, it will be possible to address many important materials phenomena including phase transformations, grain boundaries, dislocations, disorder and melting.

The problem of understanding and improving the methods of growth of complicated materials, such as multi-component heterostructures which are produced by epitaxial growth using molecular beam or chemical vapor deposition techniques, stands out as one very important technological application of this method. Although a "brute-force" simulation of atomic deposition on experimental time scales will not possible for sometime, one can learn a great deal from studying the mechanisms of reactive film growth. Combining atomic calculations for the structure of an interface with continuum theories of elasticity and plastic deformation is also an important area for the future.

One of the most obvious areas for future applications are biological systems, where key reaction sequences would be simulated in ab initio fashion. These calculations would not replace existing molecular mechanics approaches, but rather supplement them in those areas where they are not sufficiently reliable. This includes enzymatic reactions involving transition metal centers and other multi-center bond-reforming processes. A related area is catalysis, where the various proposed reaction mechanisms could be explicitly evaluated. Short-time finite temperature simulations can also be explored to search for unforeseen reaction pat-

terns. The potential for new discoveries in these areas is high.

Important progress has also been made in understanding the excitation properties of solids, in particular the predictions of band offsets and optical properties. This requires the evaluation of the electron self-energy and is computationally much heavier than the local density approaches discussed above. This first principles quasiparticle approach has allowed for the first time the ab initio calculation of electron excitation energies in solids valid for quantitative interpretation of spectroscopic measurements. The excitation of systems as complex as C_60 fullerites have been computed. Although the quasiparticle calculations have yet to be implemented on massively parallel machines, it is doable and the gain in efficiency and power is expected to be similar to the ab initio molecular dynamics types of calculations.

Much effort has been devoted in the past several years to algorithm development to extend the applicability of these new methods to ever larger systems. The ab initio molecular dynamics have been successfully implemented on massively parallel machines for systems as large as 700 atoms. Tight binding molecular dynamics methods are an accurate, empirical way to include the electronic degrees of freedom which are important for covalently bonded materials, at speeds of 300 time faster than ab initio methods. This method has already been used to simulate 2000 atoms and with the new massively parallel machines, this number will easily increase to 10,000 within a year. Another very exciting recent development in this area is work on the so-called order N methods for electronic structure calculations. At present, quantum mechanical calculations scale at least as N^3 in the large N limit, where N is the number of atoms in a unit cell. Significant progress has been made recently by several groups in developing methods which would scale as N. The success of these approaches would further enhance our ability to study very large molecular and materials systems including systems with perhaps thousands of atoms in the near future.

Quantum Many-Body

The quantum many body problem lies at the core of understanding many properties of materials. Over the last decade much of the classical methodology discussed above has been extended into the quantum regime in particular with the development of the path integral and Green's function Monte Carlo methods. Early calculations of the correlation energy of the electron gas are extensively used in local density theory to estimate correlation energy in solids. The low temperature properties of liquid and solid helium, three and four, the simplest strongly correlated many-body quantum system, are now well understood thanks in large part to computer simulations. These quantum simulations have required thousands of hours on Cray-YMP class computers.

While there still remain very difficult algorithmic issues, exact fermion methods and quantum dynamical methods to name two, the progress in the next decade should parallel the previous developments in classical statistical mechanics. Computer simulations of quantum many-body systems will become a ubiquitous tool, integrated into theory and experiment. The software and hardware has reached a state where much larger, complex and realistic systems can be studied. Some particular examples are electrons in transition metals, in restricted geometries, at high temperatures and pressures or in strong magnetic fields. Mean field theory is unreliable in many of these situations. However, these applications, if they are to become routine and widely distributed in the materials science community will require high performance hardware. Quantum simulations are naturally parallel and are likely to be among the first applications using massively parallel computers.

Thanks to J. Bernholc, D. Ceperley, J. Joannopoulous, B. Harmon and S. Louie for their help in preparing this subsection.

Computational Molecular Biology/Chemistry/Biochemistry

by Barry Honig

Background

There have been a number of revolutionary developments in molecular biology that have greatly expanded the need for high performance computing within the biological community. First, there has been an exponential growth in the number of gene sequences that have been determined, and no end is in sight. Second, there has been a parallel (although slower) growth in the number of proteins whose structures have been determined from x-ray crystallography and, increasingly, from multidimensional NMR. This literal explosion in new information has led to developments in areas such as statistical analysis of gene sequences and of three dimensional structural data, new databases for sequence and structural information, molecular modeling of proteins and nucleic acids, and three-dimensional pattern recognition. The recognition of grand challenge problems such as protein folding or drug design has resulted in large part from these developments. Moreover, the continuing interest both in sequencing the human genome and in the field of structural biology guarantees that computational requirements will continue to grow rapidly in the coming decade.

To illustrate the type of problems that can arise, consider the case where a new gene has been isolated and its sequence is known. In order to fully exploit this information it is necessary to first obtain maximum information about the protein this gene encodes. This can be accomplished by searching a nucleic acid sequence data base for structurally or functionally related proteins, and/or by detecting sequence patterns characteristic of the three dimensional fold of a particular class of proteins. There are numerous complexities that arise in such searches and the computational demands imposed by the increasingly sophisticated statistical techniques that are being used can be imposing.

A variety of methods, all of them requiring vast computational resources, are currently being applied to the protein folding problem (predicting three dimensional structure from amino acid sequence). Methods include statistical analyses (including neural nets) of homologies to known structures, approaches based on physical chemical principles and simplified lattice models of the type used in polymer physics. A major problem in understanding the physical principles of protein and nucleic conformation is the treatment of the surrounding solvent. Molecular dynamics techniques are widely used to model the solvent but their accuracy depends on the potential functions that are used as well as the number of solvent molecules that can be included in a simulation. Thus, the technique is limited by the available computational power. Continuum solvent models offer an alternative approach but these too are highly computer intensive.

Even assuming a reliable method to evaluate free energies, the problem of conformational search is daunting. There are a large number of possible conformations available to a macromolecule and it is necessary to develop methods, such as Monte Carlo techniques with simulated annealing, to ensure that the correct one has been included in the generated set of possibilities. A similar set of problems arises, for example, in the problem of structure-based drug design. In this case one may know the three-dimensional structure of a protein and it is necessary to design a molecule that binds tightly to a particular site on the surface. Efficient conformational search, energy evaluation and pattern recognition are requirements of this problem, all requiring significant computational power.

Significant progress has been made in these and many other related areas. Ten years ago most calculations were made without including the effects of solvent. This situation has changed dramatically due to scientific progress that has been potentiated by the availability of significant computational power. Some of this has been provided by Supercomputer centers while some has been made available by increasingly powerful workstations. Fast

computers have also been crucial in the very process of three dimensional structure determination. Both x-ray and NMR data analysis have exploited methods such as molecular dynamics and simulated annealing to yield atomic coordinates of macromolecules. More generally, the new discipline of structural biology, which involves the structure determination and analysis of biological macromolecules, has been able to evolve due the increased availability of high performance computing.

Future

Despite the enormous progress that has been made the field is just beginning to take off. Gene sequence analysis will continue to become more effective as the available data continue to grow and as increasingly sophisticated data analysis techniques are applied. It will be necessary to make state-of-the-art sequence analysis available to individual investigators, presumably through distributed workstations and through access to centralized resources. This will require a significant training effort as well as the development of user-friendly programs for the biological community.

There is enormous potential in the area of three dimensional structure analysis. There is certain to be major progress in understanding the physical chemical basis of biological structure and function. Improved energy functionals resulting from progress in quantum mechanics will become available. Indeed a combination of quantum mechanics and reaction field methods will make it possible to obtain accurate descriptions of molecules in the condensed phase. The impact of such work will be felt in chemistry as well as in biology. Improved descriptions of the solvent through a combination of continuum treatments and detailed molecular dynamics simulations at the atomic level will lead to truly level descriptions of the conformational free energies and binding free energies of biological macromolecules. When combined with sophisticated conformational search techniques, simplified lattice models, and sophisticated statistical techniques that identify sequence and structural homologies, there is every reason to expect major progress on the protein folding problem.

There will be parallel improvements in structure based design of biologically active compounds such as pharmaceuticals. Moreover, the development of new compounds based on biomimetic chemistry and new materials based on polymer design principles deduced from biomolecules should become a reality. All of this progress will require increased access to high performance computing for the reasons given above. The various simulation and conformational search techniques will continue to benefit dramatically from increased computational power. This will be true at the level of individual workstations, which a bench chemist for example might use to design a new drug. Work of this type often requires sophisticated three dimensional graphics and will benefit from progress in this area. Massively parallel machines which will certainly be required for the most ambitious projects. Indeed, it is likely that for some applications the need for raw computing power will exceed what is available in the foreseeable future.

New developments in the areas covered in this section will have major economic impact. The biotechnology, pharmaceutical and general health industries are obvious beneficiaries but there will be considerable spin-off in materials science as well.

Recommendations

- Support the development of software in computational biology and chemistry. This should take the form of improved software and algorithms for workstations as well as the porting of existing programs and the development of new ones on massively parallel machines.
- Make funds available for training that will exploit new technologies and for familiarizing biologists with existing technologies.
- Funding should be divided between large centers, smaller centers involving a group of investigators at a few sites developing new technologies, and individual investigators.

Molecular Modeling and Quantum Chemistry

by William Lester

The need for high performance computing has been met historically by large vector supercomputers. It is generally agreed in the computer and computational science communities that significance improvements in computational efficiency will arise from parallelism. The advent of conventional parallel computer systems has generally required major computer code restructure to move applications from vector serial computers to parallel architectures. The move to parallelism has occurred in two forms: distributed MIMD machines and clusters of workstations with the former receiving the focus of attention in large multi-user center facilities and the latter in local research installations.

The tremendous interest in the simulation of biological processes at the molecular level using molecular mechanics and molecular dynamics methods has led to continuing increase in demand of computational power. Applications have potentially high practical value and include, for example, the design of inhibitors for enzymes that are suspected to play a role in disease states and the effect of various carcinogens on the structure of DNA.

In the first case, one expects that a molecule designed to conform within the three-dimensional arrangement of the enzyme structure should be bound tightly to the enzyme in solution. This requirement, and others, make it desirable to know the tertiary structure of the enzyme. The use of computation for this purpose is contributing significantly to the understanding of those structures which are then used to guide organic synthesis.

In the second case, serial vector supercomputers typically can carry out molecular dynamics simulations of DNA for time frames of only picoseconds to nanoseconds. A recent calculation of 200-ps involving 3542 water molecules and 16 sodium ions took 140 hours of Cray Y-MP time. Extending such calculations to the millisecond or even the second range where important motions can occur remains a major computational challenge that will require the use of massively parallel computer systems.

Although in molecular mechanics or force field methods, computational effort is dominated by the evaluation of the force field that gives the potential energy as a function of internal coordinates of the molecules and non-bonded interactions between atoms, a popular approach for small organic molecules is the ab initio Hartree-Fock (HF) method which has come into routine use by organic and medicinal chemists to study compounds and drugs. The HF method is ab initio because the calculation depends only on basic information about the molecule, including the number and types of atoms and the total change. The computational effort of HF computations scales as N4, where N is the number of basis functions used to describe the atoms of the molecule. Because the HF method describes only the "average" behavior of electrons, it typically provides a better description of relative geometries than of energetics. The accurate treatment of the latter requires proper account of the instantaneous correlated motions of electrons which inherently is not described by the HF method.

For systems larger than those accessible with the HF methods, one has, in addition to molecular mechanics methods, semiemperical approaches. Their name arises out of the use of experimental data to parameterize integrals and other simplifications of the HF method leading to a reduction of computational effort to order N3. Results of these methods can be informative for systems where parametrization has been performed.

Recently, the density functional (DF) method has become popular, overcoming deficiencies in accuracy for chemical applications that limited earlier use. Improvements have come in the form of better basis sets, advances in computational algorithms for solving the DF equations, and the development of analytical geometry optimization

methods. The DF method is an ab initio approach that takes into account electron correlation. In view of the latter capability, it can be used to study a wide variety of systems, including metals and inorganic species.

The move to parallel systems has turned out to be a major undertaking in software development for the approaches described. Serious impediments have been encountered in algorithm modification for methods that go beyond the ab initio HF method, and in steps to maximize efficiency with increased numbers of processors. These circumstances have increased interest in quantum Monte Carlo (QMC) methods for electronic structure. In addition, QMC methods have been used with considerable success for the calculation of vibrational eigenvalues, and in statistical mechanics studies.

QMC, as used here in the context of electronic structure, is an ab initio method for solving the Schroedinger equation stochastically based on the formal similarity between the Schroedinger equation and the classical diffusion equation. The power of the method is that it is inherently an N-body method that can capture all of the instantaneous correlation of the electrons. The QMC method is readily ported to parallel computer systems with orders of magnitude savings in computational effort over serial vector supercomputers.

In the statistical mechanical studies of complex systems, one is often interested in the spontaneous formation and energetics of structure over large length scales. The mesoscopic structures, such as vesicles and lamellars, formed from self assembly in oil/water-surfactant mixtures are important examples. The systematic analysis of these phenomena have only recently begun, and computer simulation is one of the important tools in this analysis. Due to the large length scales involved, simulation is necessarily confined to very simple classes of models. Even so, the work presses the capabilities of current computational equipment to their limits. While the stability of various nontrivial structures have been documented, we are still far from understanding the rich phase diagram in such systems. The work of Smit and coworkers using transputors demonstrates the utility of parallelization in these simulations. Future equipment should carry us much further towards understanding.

Competing interactions and concomitant "frustration" characterizes complex fluids and the resulting mesoscopic structures. Such competition is also a central feature of "random polymers" - a model for proteins and also for manufactured polymers. Computer simulation studies of random polymers can be extremely useful. Though here too, the computations of even the simplest models press the limits of current technology. To treat this class of systems, Binder and coworkers and Frenkel and coworkers have developed new algorithms, some of which are manifestly parallelizable. Thus, this area is one where the new computer technology should be very helpful.

Along with large length scale fluctuations, as in self assembly and polymers, simulations press current computational equipment where relaxation occurs over many orders of magnitude. This is the phenomena of glasses. Here, the work of Fredrickson on the spin-facilitated Ising system demonstrates the feasibility of parallelization in studying long time relaxation and the glass transition by simulation.

Polymers and glasses are examples pertinent to the understanding and design of advanced materials. In addition, one needs to understand the electronic and magnetic behavior of these and other condensed matter systems. In recent years, a few methods have appeared, especially the Car-Parrinello approach, which now makes feasible the calculation of electronic properties of complex materials. The calculations are intensive. For example, studying the dynamics and electronic structure of a system with only 64 atoms, periodically replicated, for only a picosecond is at the limits of current capabilities. With parallelization, and simplified models, one can imagine, however, significant progress in our understanding of metalinsulator transitions and localization in correlated disordered systems.

Mathematics and High Performance Computing

by James Sethian

A. Introduction

Mathematics underlies much, if not all, of high performance computing. At first glance, it might seem that mathematics, with its emphasis on theorems and proofs, might have little to contribute to solving large problems in the physical sciences and engineering. On the contrary, in the same way that mathematics contributes the underlying language for problems in the sciences, engineering, discrete systems, etc., mathematical theory underlies such factors as the design and understanding of algorithms, error analysis, approximation accuracy, and optimal execution. Mathematics plays a key role in the drive to produce faster and more accurate algorithms which, in tandem with hardware advances, produce state-of-the-art simulations across the wide spectrum of the sciences.

At the same time, high performance computing provides a valuable laboratory tool for many areas of theoretical mathematics such as number theory and differential geometry. At the heart of most simulations lies a mathematical model and an algorithmic technique for approximating the solution to this model. Aspects of such areas as approximation theory, functional analysis, numerical analysis, probability theory, and the theory of differential equations provide valuable tools for designing effective algorithms, assessing their accuracy and stability, and suggesting new techniques.

What is so fascinating about the intertwining of computing and mathematics is that each invigorates the other. For example, understanding of entropy properties of differential equations have led to new methods for high resolution shock dynamics, approximation theory using multipoles has led to fast methods for N-body problems, methods from hyperbolic conservation laws and differential geometry have produced exciting schemes for image processing, parallel computing has spawned new schemes for numerical linear algebra and multi-grid techniques, and methods designed for tracking physical interfaces have launched new theoretical investigations in differential geometry, to name just a few. Along the way, this interrelation between mathematics and computing has brought breakthroughs in such areas as material science (such as new schemes for solidification and fracture problems), computational fluid dynamics (e.g., high order projection methods and sophisticated particle schemes), computational physics (such as new schemes for Ising models and percolation problems), environmental modeling (such as new schemes for groundwater transport and pollutant modeling) and combustion (e.g. new approximation models and algorithmic techniques for flame chemistry/fluid mechanical interactions).

B. Current State

Mathematical research which contributes to high performance computing exists across a wide range. On one end are individual investigators or small, joint collaborations. In these settings, the work takes a myriad of forms; brand-new algorithms are invented which can save an order of magnitude speedup in computer resources, existing techniques are analyzed for convergence properties and accuracy, and model problems are posed which can isolate particular phenomena. For example, such work includes analysts working on fundamental aspects of the Navier-Stokes equations and turbulence theory (c.f. the following section on Computational Fluid Dynamics), applied mathematicians designing new algorithms for model equations, discrete mathematicians focussed on combinatorics problems, and numerical linear algebraists working in optimization theory. At the other end are mathematicians working in focussed teams on particular problems, for example, in combustion, oil recovery, aerodynamics, material science, computational fluid dynamics, operations research, cryptography, and computational biology.

Institutionally, mathematical work in high performance computing is undertaken at universities, the National Laboratories, the NSF High Performance Computing Centers, NSF Mathematics Centers (such as the Institute for Mathematical Analysis and the Mathematical Sciences Research Institute, and the Institute for Advanced Study), and across a spectrum of industries. In recent years, high performance computing has become a valuable tool for understanding subtle aspects of theoretical mathematics. For example, computing has revolutionized the ability to visualize and evolve complex geometric surfaces, provided techniques to untie knots, and helped compute algebraic structures.

C. Recommendations

Mathematical research is critical to ensure state-of-the-art computational and algorithmic techniques which foster the efficient use of national computing resources. In order to promote this work, it is important that :

1. Mathematicians be supported in their need to have access to the most advanced computing systems available, both through networks to supercomputer facilities, on-site experimental machines (such as parallel processors), and individual high-speed workstations.
2. State-of-the-art research in modeling, new algorithms, applied mathematics, numerical analysis, and associated theoretical analysis be amply supported; it is this work that continually and continuously rejuvenates computational techniques. Without it, yesterday's algorithms will be running on tomorrow's machines.
3. Such research be supported on all levels; the individual investigator, small joint collaborations, interdisciplinary teams, and large projects.
4. Funding be significantly increased in the above areas, both to foster frontier research in computational techniques, and to use computation as a bridge to bring mathematics and the sciences closer together.

Computational Fluid Dyanmics

by James Sethian

Introduction

The central goal of computational fluid dynamics (CFD) is to follow the evolution of a fluid by solving the appropriate equations of motion on a numerical computer. The fundamental equations require that the mass, momentum, and energy of a liquid/gas are conserved as the fluid moves. In all but the simplest cases, these equations are too difficult to solve mathematically, and instead one resorts to computer algorithms which approximate the equations of motion. The yardstick of success is how well the results of numerical simulation agree with experiment in cases where careful laboratory experiments can be established, and how well the simulations can predict highly complex phenomena that cannot be isolated in the laboratory.The effectiveness and versatility of a computational fluid dynamics simulation rests on several factors. First, the underlying model must adequately describe the essential physics. Second, the algorithm must accurately approximate the equations of motion. Third, the computer program must be constructed to execute efficiently. And fourth, the computer equipment must be fast enough and large enough to calculate the answers sufficiently rapidly to be of use. Weaving these factors together, so that answers are accurate, reliable, and obtained with acceptable cost in an acceptable amount of time, is both an art and a science. Current uses of computational fluid dynamics range from analysis of basic research

into fundamental physics to commercial applications. While the boundaries are not sharp, CFD work may be roughly categorized in three ways: Fundamental Research, Applied Science, and Industrial Design and Manufacturing

CFD and Fundamental Research

At many of the nation's research universities and national laboratories, much of the focus of CFD work is on fundamental research into fluid flow phenomena. The goal is to understand the role that fluid motion plays in such areas as the evolution of turbulence in the atmosphere and in the oceans, the birth and evolution of galaxies, atmospherics on other planets, the formation of polymers, the physiological fluid flow in the body, and the interplay of fluid mechanics and material science such as in the physics of superconductors. In these simulations, often employing the most advanced and sophisticated algorithms, the emphasis is on accurate solutions and basic insight. These calculations are often among the most expensive of all CFD simulations, requiring many hundreds of hours of computer time on the most advanced machines available for a single simulation. The modeling and algorithmic techniques for such problems are constantly under revision and refinement. For the most part, the major advances in new algorithmic tools, from schemes to handle the associated numerical linear algebra to high order methods to approximate difference equations, have their roots in basic research into CFD applied to fundamental physics.

CFD and Applied Science

Here, the main emphasis is on the application of the tools of computational fluid dynamics to problems motivated by specific problems such as might occur in natural phenomena or physical processes. Such work might include detailed studies of the propagation of flames in engines or fire research in closed rooms, the fluid mechanics involved in the dispersal of pollutants or toxic groundwater transport, the hydraulic response of a proposed heart valve, the development of severe storms in the atmosphere, and the aerodynamic properties of a proposed space shuttle design. For the most part, this work is also carried out throughout the national laboratories and universities with government support. While cost is not a major issue in these investigations, the focus is more on obtaining answers to directed questions. Less concerned with the algorithm for its own sake, this work links basic research with commercial CFD applications, and provides a stepping stone for advances in algorithms to propagate into industrial sectors.

CFD and Industrial Design and Manufacturing

The focus in this stage of the process is on applying the tools of computational fluid dynamics to solve problems that directly relate to technology. A vast array of examples exist, such as the development of a high-speed inkjet plotter, the action of slurry beds for processing minerals, analysis of the aerodynamic characteristics of an automobile or airplane, efficiency analysis of an internal combustion engine, performance of high-speed computer disk drives, and optimal pouring and packaging techniques in manufacturing. For the most part, such work is carried out in private industry, often with only informal ties to academic and government scientists. Communication of new ideas rests loosely on the influx of new employees trained in the latest techniques, journal articles, and professional conferences. A distinguishing characteristic of this work is its emphasis on turnaround time and cost. Here, cost is not only the cost of the equipment to perform the calculation, but the people-years involved in developing the computer code, and the time involved performing may hundreds of simulations as part of a detailed parameter study. This motivation is quite different from that in the other two areas. The need to perform a large number of simulations under extremely general circumstances may mean that a simple and fast technique that attacks only a highly simplified version of the problem may be preferable to a highly sophisticated and accurate technique that requires many orders of magnitude more computational effort. This orientation is lies at the heart of the applicability and suitability of computational techniques to a competitive industry.

Future of CFD

- Modeling and Algorithmic Issues.

  In many ways, the research, applied science, and industry agenda in CFD has changed, mostly in response to increased computational power coupled to significant algorithmic and theoretical/numerical advances. On the research side, up through the early 1980's, the emphasis was on basic discretization methods. In that setting, it was possible to develop methods based on looking at simple problems in two, or even one space dimension, in simple geometries, and in a fair degree of isolation from the fluid dynamics applications. Over the last five years, however, there has been a transition to the next generation of problems. These problems are more difficult in part because they are attempting more refined and detailed simulations, necessitating finer grids and more computational elements. However, a more fundamental issue is that these problems are qualitatively different from those previously considered. To begin, they often involve complex and less well-understood physical models - chemically reacting fluid flows, flows involving multiphase or multicomponent mixtures of fluids or other complex constitutive behavior. They are often set in three dimensions, in which both the solution geometry and the boundary geometry are more complicated than in two dimensions. Finally, they often involve resolving multiple length and time scales, such as boundary and interior layers coming from small diffusive terms, and intermittent large variations that arise in fluid turbulence. Algorithmically, these require work in several areas. Complex physical behavior makes it necessary to develop a deeper understanding of the physics and modeling than had previously been required. Additional physics requires the theory and design of complex boundary conditions to couple the fluid mechanics to the rest of the problem. Multiple length scales and complex geometries lead to dynamically adaptive methods, since memory and compute power are still insufficient to brute force through most problems. For example, the wide variation in length and time scales in turbulent combustion require a host of iterative techniques, stiff ode solvers, and adaptive techniques. Further algorithmic advances are required in areas such as domain decomposition, grid partitioning error estimation, and preconditioners and iterative solvers for non-symmetric, non-diagonally dominant matrices. Mesh generation, while critical, has not progressed far.

  All in all, to accomplish in three-dimensional complex flow what is now routine in two-dimensional basic flow will require theory, numerics, and considerable cleverness. The net effect of these developments is to make the buy-in to perform CFD research much higher. Complex physical behavior makes it necessary to become more involved with the physics modeling than had previously been the case. The problems are sufficiently difficult that one cannot blindly throw them at the computer and overwhelm them. A considerable degree of mathematical, numerical, and physical understanding must be obtained about the problems in order to obtain efficient and accurate solution techniques. On the industrial side, truly complex problems are still out-of-reach. For example, in the aircraft industry, we are still a long way from a full Navier-Stokes high Reynolds number unsteady flow around a commercial aircraft. Off in the distance are problems of takeoff and landing, multiple wings in close relation to each other, and flight recovery from sudden changes in conditions. In the automotive industry, a solid numerical simulation of the complete combustion cycle (as opposed to a time-averaged transport model) is still many years away. Other automotive CFD problems include analysis of coolant flows, thermal heat transfer, plastic mold problems, and sheet metal formation.

- High Performance Computing Issues.

The computer needs to continue the next five years of CFD work are substantial. As an example, an unforced Navier-Stokes simulation might require 1000 cells in each of three space dimensions. Presuming 100 flops per cell. and 25,000 time steps, this yields $10^9 \times 10^2 \times 2.4 \times 10^4 = 2.5 \times 10^{15}$ flops; a typical compute time of 2 hours would then require a machine of 300 gigaflops. Adding forcing, combustion, or other physics severely extends this calculation. As a related issue; memory requirements pose an additional problem. Ultimately, more compute power and memory is needed. The promise of a single, much faster, larger vector machine is not being made convincingly, and CFD is attempting to adapt accordingly. Here is an area where the dream of parallelism is both tantalizing and frustrating. To begin, parallel computing has naturally caused emphasis on issues related to processor allocation and load balancing. To this end, communications cost accounting (as opposed to simply accounting for floating point costs) has become important in program design. For example, parallel machines can often have poor cache memory management and a limited number of paths to/from main memory; this can imply a long memory fetch/store time, which can result in actual computational speeds for real CFD problems far below the optimal peek speed performance. To compensate, parallel computers tend to be less memory efficient than vector machines, as space is exchanged for communication time (duplicating data where possible rather than sending it between processors). The move to parallel machines is complicated by the fact that millions of lines of CFD codes have been written in the serial/vector format. The instability of the hardware platforms, the lack of a standard global high performance Fortran and C, the lack of complete libraries, and insecurity associated with a volatile industry all contribute to the caution and reluctance of all but the most advanced research practitioners of CFD.

Recommendations

In order to tackle the next generation of CFD problems, the field will require:

- Significant accessibility to the fastest current coarse-grained parallel machines.
- Massively parallel machines with large memory, programmable under stable programming environments, including high performance Fortran and C, mathematical libraries, functioning I/O systems, and advanced visualization systems.
- Algorithmic advances in adaptive meshing, grid generation, load balancing, and high order difference, element, and particle schemes.
- Modeling and theoretical advances coupling fluid mechanics to other related physics.

High Performance Computing In Physics

by James Sethian and Neal Lane

INTRODUCTION AND BACKGROUND

High Energy Physics

Two areas in which high performance computing plays a crucial role are lattice gauge theory and the analysis of experimental data. Lattice gauge theory addresses some of fundamental theoretical problems in high energy physics, and is relevant to experimental programs in high energy and nuclear physics. In the standard model of high energy physics, the strong interactions are described by quantum chromodynamics (QCD). In this theory the forces are so strong that the fundamental entities, the quarks and gluons, are not

observed directly under ordinary laboratory conditions. Instead one observes their bound states, protons and neutrons, the basic constituents of the atomic nucleus, and a host of short lived particles produced in high energy accelerator collisions. One of the major objectives of lattice gauge theory is to calculate the masses and other basic properties of these strongly interacting particles from first principles, and provide a test of QCD, as well as suggest that the same tools could be used to calculate additional physical quantities which may not be so well determined experimentally. In addition, lattice gauge theory provides an avenue for making first principle calculations of the effects of strong interactions on weak interaction processes, and thus holds the promise of providing crucial tests on the standard model at its most vulnerable points. And, although quarks and gluons are not directly observed in the laboratory, it is expected that at extremely high temperatures one would find a new state of matter consisting of a plasma of these particles. The questions being addressed by lattice gauge theorists are the nature of the transition between the lower temperature state of ordinary matter and the high temperature quark-gluon plasma, the temperature at which this transition occurs, and the properties of the plasma. In the general area of experimental high energy physics, there are three primary areas of computing:

1) The processing of the raw data that is usually accumulated at a central accelerator laboratory, such as the Wilson Laboratory at Cornell.

2) The simulation of physical processes of interest, and the simulation of the behavior of the final states in detector.

3) The analysis of the compressed data that results from processing of the raw data and the simulations.

Atomic and Molecular Physics

In contrast to some areas of theoretical physics, the AM theorist has the advantage that he understands the basic equations governing the evolution of the system of particles under consideration. However, the wealth of phenomena that derive from the many-body interactions of the constituents and their interactions with external probes such as electric and magnetic fields are truly astounding. Computation now provides a practical and useful alternative method to study these problems. Most importantly, it is now possible to perform calculations sophisticated enough to have a real impact on AM science. These include high precision computations of the ground and excited states of small molecules, scattering of electrons from atoms, atomic ions and small polyatomic molecules, simple chemical reactions involving atom-diatom collisions, photoionization and photodissociation and various time dependent processes such as multiphoton ionization and the interaction of atoms with ultra strong (or short) electromagnetic fields.

Gravitational Physics

The computational goal of classical and astrophysical relativity is the solution of the associated nonlinear partial differential equations. For example, simulations have been performed of the critical behavior in black hole formation using high accuracy adaptive grid methods and which follow the collapse of spherical scalar wave pulses over at least fourteen orders of magnitude, of the structure we currently see in the universe (galaxies, clusters of galaxies arranged in sheets, voids etc.) and how it may have arisen through fluctuations generated during an inflationary epoch, of head-on collisions of black holes, and of horizon behavior in a number of black hole configurations.

CURRENT STATE

Definitive calculations in all of the areas mentioned above would require significantly greater computing resources than have been available up to now. Nevertheless steady progress has been made during the last decade due to important improvements in algorithms and calculational techniques, and very rapid increases in available computing power. For example, among the major achievements in lattice gauge theory have been: a demonstration that quarks and gluons are confined at low temperatures; steady improvements in spectrum calculations, which have accelerated

markedly in the last year; an estimate of the transition temperature between the ordinary state of matter and the quark-gluon plasma; a bound on the mass of the Higgs boson; calculations of weak decay parameters including a determination of the mixing parameter; and a determination of the strong coupling constant at the energy scale of 5 GeV from a study of the charmonium spectrum. Much of this work has been carried out at the NSF Supercomputer Centers. In the area of atomic and molecular physics, until quite recently most AM theorists were required to compute using simplified models or, if the research was computationally intensive, to use vector supercomputers. This has changed with the widespread availability of cheap, fast, Unix based, RISC workstations. These "boxes" are now capable of performing at the 40-50 megaflop level and can have as much as 256 megabytes of memory. In addition,it is possible to cluster these workstations and distribute the computational task among the cpu's. There are a few researchers in the US who have as many as twenty or thirty of these workstations for their own group. This has enabled computational experiments using a loosely coupled, parallel model on selected problems in AM physics. However, this is not the norm in AM theory. More typically our most computationally intensive calculations are still performed on the large mainframe, vector supercomputers available only to a limited number of users. The majority of the AM researchers in the country are still computing on single workstations or (outdated) mainframes of one kind or another.

FUTURE COMPUTING NEEDS

In the field of high energy physics, the processing of raw data and the simulation of generic mixing processes are activities that are well-suited to a centralized computer center. The processing of the raw data should be done in a consistent, organized, and reliable manner. Often, in the middle of the data processing, special features in some data are discovered, and these need to be treated quickly and in a manner consistent with the entire sample. It is usual and logical that the processing take place at the central accelerator center where the apparatus sits, because the complete records of the data taking usually reside there. In contrast, the high energy physics work in simulation of specific processes and analysis of compressed data are well matched to individual University groups.In terms of computing needs, all of the above are well served either by a large, powerful, central computer system, or by a cluster or farm of workstations. For example, CERN does the majority of its computing on central computers, while the Wilson Lab has a farm of DECstation 5000/240's to process its raw data. High energy computing usually proceeds one event at a time; each event, whether from simulation or raw data, can be dealt to an individual workstation for processing. There is no need to put the entire resources of a supercomputer on one event. However, massively parallel computers have the potential to handle large numbers of events simultaneously in an efficient manner. These machines will also be of importance in the work on lattice gauge theory. Conversely, university groups are well served by the powerful workstations now available, which has freed them from dependence on the central laboratory to study the physics signals of interest to them. These groups need both fast CPUs, for simulation of data, as well as relatively large and fast disk farms, for repeated processing of the compressed data. In the field of atomic and molecular modeling, the future lies in the use of massively parallel, scaleable multicomputers. AM theorists, with rare exceptions, have not been as active as other disciplines in moving to these platforms. This is to be contrasted with the quantum chemists, the lattice QCD theorists and many materials scientists who are becoming active users of these computers. The lack of portability of typical AM codes and the need to expend lots of time and effort in rewriting or rethinking algorithms has prevented a mass migration to these platforms.

Accomplishments in Computer Science and Engineering since the Lax Report

by Mary K. Vernon

While vector multiprocessors have been the workhorses for many fields of computational science and engineering over the past ten years, research in computer science and engineering has been focused both on improving the capabilities of these systems, and on developing the next generation of high-performance computing systems — namely the scalable, highly parallel computers which have recently been commercially realized as systems such as the Intel Paragon, the Kendall Square Research KSR-1, and the Thinking Machines Corporation CM-5.

A variety of factors make scalable, highly parallel computers the only viable way to achieve the teraflop capability required by Grand Challenge applications. These systems represent far more than an evolutionary step from their modestly parallel vector predecessors. Realizing the teraflop potential of massively parallel systems requires advances in a broad range of computer science and engineering subareas, including VLSI, computer architecture, operating systems, runtime systems, compilers, programming languages, and algorithms. The development of the new capabilities in turn requires computationally intensive experimentation and/or simulations that have been carried out on experimental prototypes (e.g., the NYU Ultracomputer), early commercial parallel machines (such as the BBN Butterfly or the Intel iPSC/2), and more recently on high-performance workstations as well as the emerging massively parallel systems such as the Thinking Machines CM5 and the Intel Paragon.

Computer science and engineering researchers have made tremendous progress in the past ten years in the development of high performance computing technology, including the development of ALL of the major technologies in massively parallel systems. Among the specific accomplishments are:

- development of RISC processor technology and the compiler technology for RISC processors, which is used in all high performance workstations as well as in the massively parallel machines.
- development of computer-aided tools to facilitate the design, testing, and fabrication of complex digital systems, and their constituent components.
- invention of the multicomputer and development of the message-passing programming paradigm which is used in many of today's massively parallel systems.
- refinement of shared memory architectures and the shared memory programming model which is used in the KSR-1, the Cray T3D, and other emerging massively parallel machines.
- invention of the hypercube interconnection network and refinement of this network to lower-dimensional, 2-d and 3-d, mesh networks that are currently used in the Intel Paragon and the Cray T3D
- invention of the fat-tree interconnection network which is currently used in the Thinking Machines CM-5.
- refinement of the SIMD architecture which is used for example in the Thinking Machines CM2 and the MasPar/1.
- invention and refinement of the SPMD and data parallel programming models which are supported in several massively parallel systems.
- development of the technology underlying the mature compilers for vector machines (i.e., compilers that give delivered performance that is a substantial fraction of the theoretical peak performance of these machines.)

- development of the technology underlying all of the existing compilers for parallel machines,
- development of the Mach operating system, which provided the basis for the OSF/1 standard used, for example, in the Intel Paragon.
- development of light-weight and wait-free synchronization primitives
- development of performance debugging tools
- development of high-performance database technologies, including both algorithms and architectures that have influenced emerging systems, for example from NCR, Teradata, and IBM.
- development of parallel algorithms for high-performance optimization
- development of parallel algorithms for numerical linear algebra
- development of machine learning technology for computational biology

In other words, key hardware, system software, and algorithm technologies are directly the result of computer science and engineering research across a broad range of subdisciplines. Much of this work has been highly experimental, and has made extensive use of current-generation, early commercial, and prototype high performance systems. For example, simulations of next-generation architectures, multi-user database systems, and the like, as well as the development and testing of new algorithms for large-scale optimization, numerical linear algebra, computational biology, and the like, often require days of simulation time on the most advanced platforms available.

Research efforts today are focused on improving the capabilities, performance, and ease of use of parallel machine technology, including the capabilities of workstation networks. Experiments to evaluate the technology for next-generation systems, like many other applications that would be classified as "computational engineering", require the highest performance systems available. In addition, simulation and/or testing of innovations in computer architecture or operating systems sometimes involve modifications to the host hardware and/or operating software. These modifications can be developed and debugged on medium-scale versions of the high-end systems. Support for such initial development, as well as porting working modifications to larger-scale systems for further test, is critical to the rapid development of new HPC technologies.

NATIONAL SCIENCE FOUNDATION
WASHINGTON, D.C. 20550

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE $300

RETURN THIS COVER SHEET TO ROOM 233A IF YOU DO NOT WISH TO RECEIVE THIS MATERIAL ☐ , OR IF CHANGE OF ADDRESS IS NEEDED ☐ , INDICATE CHANGE, INCLUDING ZIP CODE ON THE LABEL (DO NOT REMOVE LABEL).