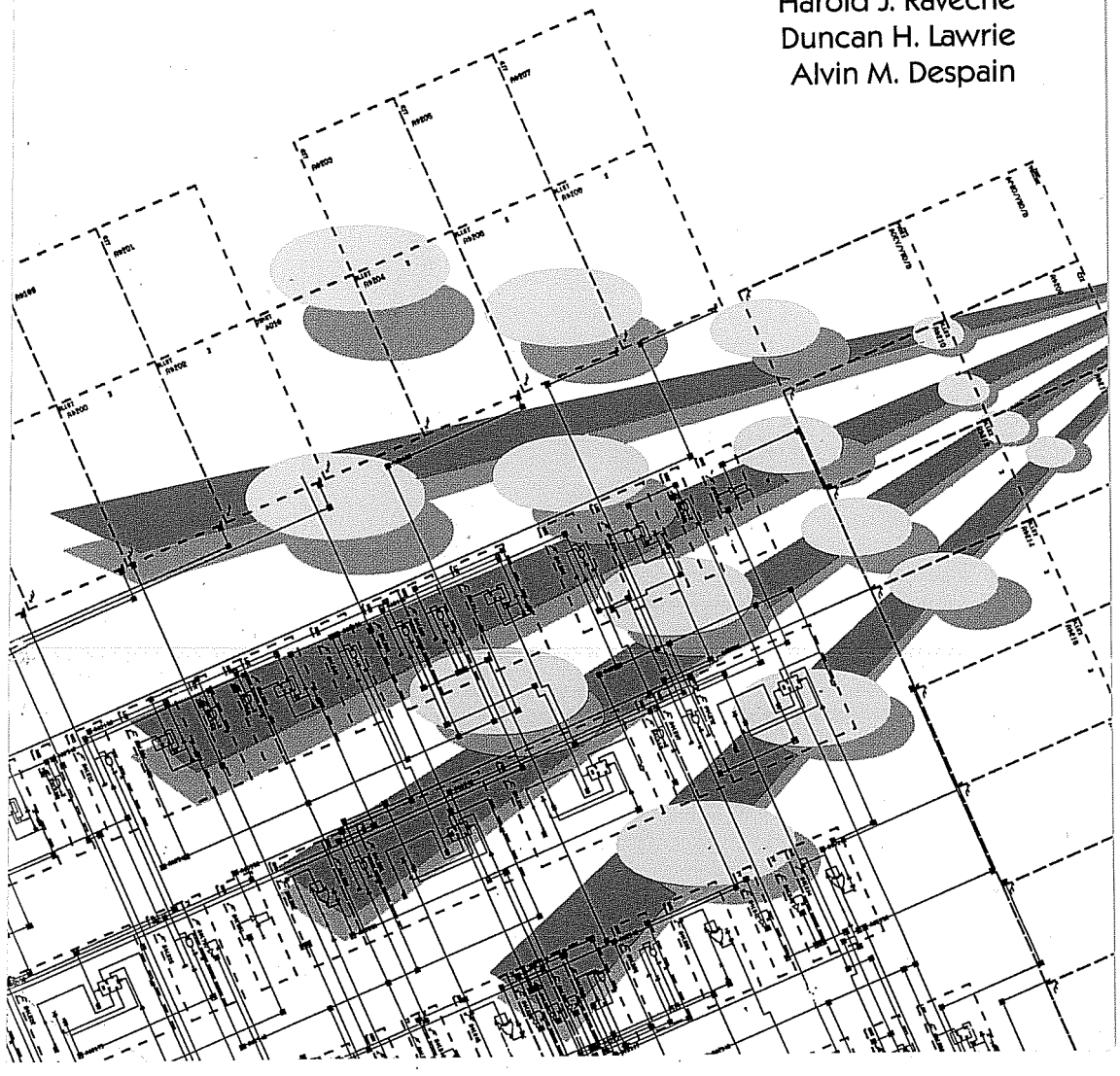
# A NATIONAL COMPUTING INITIATIVE

## The Agenda for Leadership

## Report of the Panel on Research Issues in Large-Scale Computational Science and Engineering

Harold J. Raveché
Duncan H. Lawrie
Alvin M. Despain



## Panel Members

*Sub-Panel on Applications*

Robert Brown, Massachusetts Institute of Technology
Melvyn Ciment, National Science Foundation
John M. Dawson, University of California/Los Angeles
David A. Dixon, E.I. du Pont de Nemours & Co., Inc.
John A. Dutton, Pennsylvania State University
Ron D. Ekers, National Radio Astronomy Observatory
John Geweke, Duke University
George H. Gilmer, AT&T Bell Laboratories
William Goddard, California Institute of Technology
James D. Gunton, Temple University
Karl Hess, University of Illinois/Urbana
David A. Hoffman, University of Massachusetts
Scott Kirkpatrick, IBM/T. J. Watson Research Center
Peter D. Lax, Courant Institute, New York University
(invited speaker)
William R. Martin, University of Michigan
Daniel F. X. O'Reilly, DRI/McGraw Hill
Steven A. Orszag, Princeton University
Harold J. Raveché, Rensselaer Polytechnic Institute
(Chair of Workshop and Sub-Panel)
Christopher A. Sims, University of Minnesota
Michael J. Werle, United Technologies Research Center

*Sub-Panel on Advanced Systems*

Dennis Gannon, Indiana University
Kenneth W. Kennedy, Rice University
Duncan H. Lawrie, University of Illinois/Urbana (Chair)
Joanne L. Martin, IBM/T. J. Watson Research Center
John Riganati, Supercomputing Research Center
Burton J. Smith, Supercomputing Research Center
Lawrence Snyder, University of Washington
Peter J. Weinberger, Bell Laboratories
Anthony Vacca, ETA Systems

*Sub-Panel on Parallel Computing*

Henry D. I. Abarbanel, Scripps Institute of Oceanography
Duane A. Adams, Carnegie-Mellon University
Donald M. Austin, Department of Energy
David R. Cheriton, Stanford University
Doug DeGroot, Texas Instruments, Central Research Labs, Dallas
Alvin M. Despain, University of California/Berkeley (Chair)
Jack J. Dongarra, Argonne National Laboratory
Robert M. Haralick, University of Michigan
Stephen Squires, DARPA

# A National Computing Initiative

## The Agenda for Leadership

*Report of the Panel on Research Issues in Large-Scale Computational Science and Engineering*


| | |
|---|---|
| Harold J. Raveché | Chair, SIAM Workshop<br>Chair, Sub-Panel on Applications |
| Duncan H. Lawrie | Chair, Sub-Panel on Advanced Systems |
| Alvin M. Despain | Chair, Sub-Panel on Parallel Computing |



SIAM Workshop held at Leesburg, Virginia, February 2 - 3, 1987. Sponsored by the National Science Foundation and the U.S. Department of Energy.

Distributed by the Society for Industrial and Applied Mathematics, Philadelphia, 1987.

Any opinions, findings, conclusions, or recommendations expressed in this report are those of the panels and do not necessarily reflect the view of the National Science Foundation, the U.S. Department of Energy, or the Society for Industrial and Applied Mathematics.

For additional copies write:

Society for Industrial and Applied Mathematics
1400 Architects Building
117 South 17th Street
Philadelphia, PA 19103-5052

# Contents

# Abstract of Recommendations

## Goal of the Initiative

To meet grand challenges in areas of strategic national importance, the U.S. research community must be able to exploit leading-edge computational technology. This requires, over the next few years, a thousand-fold increase in applied computational performance (i.e., in the range of one trillion floating-point operations, or one teraflop, per second, on a sustained basis) and action to ensure innovative applications of this extraordinarily powerful technology.

## Principal Recommendation and Impact

The striking commercial and scientific opportunities of modeling with supercomputers mandate that a National High-Performance Computing Initiative be established through the cooperative efforts of federal agencies, universities, and industries. New resources totalling at least $1.5 billion will be required over the next 5 years to achieve unprecedented computing power and exploitation of computing technologies in areas of strategic national importance. This initiative will accelerate the creation of innovative hardware and software, effective mathematical techniques, and new university curricula that will attract and educate students in computational engineering and science, and it will significantly leverage existing investments made by both the federal government and industry.

Realization of these goals will lead to major economic gains in a rich diversity of important applications, including high-speed aircraft; surface vehicles; medical imaging; biotechnology; and advanced structural, electronic, and optical materials. Key manufacturing industries will achieve higher productivity by (a) more effectively linking engineering analysis with product design and fabrication, (b) using computer modeling to supplement expensive product testing, and (c) having improved automation technology. Fundamental scientific discoveries will be accelerated in areas such as astronomy, virology, biochemistry, and the structure of nuclei, atoms, and molecules. Vital predictive methodologies such as weather forecasting and global economic modeling will also benefit from this technology.

# Executive Summary

## Background

In response to congressional inquiries and urged on by the extraordinary opportunities created by rapid developments in high-performance computing, the Federal Coordinating Council on Science, Engineering and Technology (FCCSET) recommended that several federal agencies convene expert panels to assess high-performance computing. The National Science Foundation, the Department of Energy, the Department of Defense, the Defense Advanced Research Projects Agency, and the National Aeronautics and Space Administration participated in a workshop convened by the Society for Industrial and Applied Mathematics (SIAM) in Leesburg, Virginia, on February 2-3, 1987. In attendance were 45 recognized leaders from industry, academe, and national laboratories. In three separate sub-panels -- on applications, advanced systems, and parallel computing -- they considered the steps necessary to grasp the opportunities and face the challenges of the next decade: in particular, to maintain U.S. leadership in computing technology and the strengthening of our competitive position vis-à-vis our trading partners. The three sub-panel reports follow this executive summary.

In the 1982 study, which led to the Lax Report [1], federal action was called for in four essential recommendations: (i) increased access for the scientific and engineering research community, (ii) increased research in computational mathematics, software, and algorithms, (iii) training of personnel in scientific and engineering computing, (iv) research and development basic to the design and implementation of new supercomputing systems. NSF, for example, responded with an internal study, the Bardon-Curtis Report [2], which spelled out the steps for NSF to follow up with the implementation in the 1985 budget year. One consequence was that NSF, DOE, and NASA initiated new programs and created what are now known as National Supercomputer Centers. Another example was the Rheinboldt Report [3], which was influential in encouraging initiatives for interdisciplinary research teams in computational science and engineering at NSF.

## High-Performance Computing

The SIAM workshop members unanimously urge that a powerful partnership of universities, federal agencies, and industries be organized to strengthen the position of the United States as the world's leader in the development and effective utilization of high-performance computing.

High-performance computing has emerged as a powerful and indispensable aid to scientific and engineering research, product and process development, and all aspects of manufacturing. This tool is critically important to the competitiveness of broad segments of America's technological industries and scientific enterprise. High-performance computing, sometimes called supercomputing, involves large-scale computations to perform specific tasks: designing new semiconductors; performing accurate three-dimensional medical imaging; examining turbulence around surface, airborne, and space vehicles; delineating the basic building blocks of nuclear matter; uncovering the structure-function relations of viruses; and modeling the global economy. It is now widely recognized that high-performance computing leads to economically significant benefits in such diverse industries as aerospace and pharmaceuticals, and that it is a cornerstone of the nation's defense system. Significant increases in performance can sometimes be achieved by reducing large calculations to a number of smaller segments that can be executed in parallel; massive parallelism is thus expected to be the computer architecture of the future. Simulations are performed not only on mainframe supercomputers, but also on an increasing number of custom-designed, parallel computing machines.

The SIAM workshop members, hereafter referred to as the Panel, considered the current state of the art in high-performance computing in light of the 5 years since the Lax Report, discussed the opportunities and challenges of computing in the next decade, and reported their conclusions which are found in the three sub-panel reports. Their findings are highlighted below.

## The National Computing Initiative

The Panel urges the creation of a comprehensive 5-year National Computing Initiative to ensure preeminence of the United States in the development and application of high-performance computing. This requires approximately $1.5 billion in new resources for computational science and engineering over the next 5 years.

New opportunities in parallel architectures, microelectronics, optoelectronics, and software, together with the broad enthusiasm of scientists and engineers to apply and integrate high-performance computing technologies, make the National Computing Initiative a technological imperative. In addition, significant advances have been made in the production of supercomputers by the Japanese; and supercomputers are heavily utilized in Japan, the United Kingdom, West Germany, and France for engineering and research by industry and government, as well as the academic community. The United States must regain the former leadership position in the design and manufacture of semiconductor components.

High-performance logic and storage devices are currently the most critical technologies. Inadequate systems software has resulted in less than optimal scientific productivity. Software is not readily portable between high-performance computers. A vigorous, cooperative effort of federal agencies, universities, and industry is required in order for the nation to gain significantly in economic competitiveness in the areas at both development and utilization of high-performance computing. The National Computing Initiative provides this thrust. Without decisive action, the United States will lose an exceptional opportunity to benefit from one of its best native technologies.

The foremost purpose of this Initiative is to provide significantly improved computational power to U.S. scientists and engineers over the next 5 years and to initiate actions that ensure innovative application of this powerful technology. Achievement of these goals requires the following: (1) accelerated development of ever faster VLSI logic and memory technologies, which entails major advances in electrical engineering and materials science; (2) improved architectures for single processors including vector processors and exploratory research on new architectures; (3) the coupling of multiple processors into highly parallel computers; (4) the organization of distributed high-performance computers into a single coherent processing system of unprecedented power; (5) the development of greatly enhanced optimizing compilers, improved languages, programming tools, and operating systems and more effective technologies for managing large databases; (6) support for computational science and engineering research in the interdisciplinary mode, combining applied mathematics, computational science, and scientific computing applied to fundamental problems in science and engineering; (7) innovative undergraduate and graduate curricula to educate engineers, scientists, economists, and so on in the effective use of advanced

computing technology and the continuous education of users of high-performance computing; and (8) the establishment of a reliable, national gigabit network to provide U.S. universities, industries, and government laboratories access to supercomputer facilities.

The special facilities required to promote the National Computing Initiative will be predicated on the existing National Supercomputer Centers undergoing rapid upgrades of computing equipment and common operating systems, including massively parallel computers as well as special purpose computers for specific applications, such as high-speed graphics and Monte Carlo simulations. In addition, the implementation of the National Computing Initiative will include the establishment of separate institutions or centers to achieve the specific goals outlined above. The Panel feels that three types of institutions are urgently needed to accomplish these goals: First, the "computational science and engineering centers" would be devoted to research in computational methods, numerical analysis, algorithm design, and development, and related areas such as pattern recognition, imaging, and nonlinear phenomena. Second, the "software institutes" will foster the development of powerful applications and systems in software to the prototype stage, through a mix of basic and applications-oriented research. Third, the "software libraries" will be responsible for software documentation, maintenance, and dissemination to the computational community. Computational scientists and engineers, computer scientists, programmers, and other researchers interested in developing methods, algorithms, and software to achieve the full power of high-performance computing will be members of these centers. All of these institutes and centers would be linked together with the National Supercomputer Centers on the national gigabit network.

The National Computing Initiative is a comprehensive and focused effort that addresses the key issues discussed below: basic research, access, and education.

## Basic Research

The Panel unanimously recommends significant increases in support, on a stable basis, for scientific and engineering research that underpins high-performance computing. While architecture and logic design have been and are the driving technology for supercomputers, the Panel urges attention to the increasing importance of developing improved computational methods, and

more efficient algorithms to exploit current and future architectures. Progress in these areas will be fostered by the creation of the three types of centers and institutions discussed above, as well as by the existing National Supercomputer Centers. Significant advances must occur in areas such as compilers, programming languages, computer design methodology, and operating systems. The Panel urges the development of powerful concepts and technologies for the management of large scientific and engineering databases. The "software institute" and "software library" would contribute to progress in these areas. Several aspects of applied mathematics, such as algorithm design, numerical analysis, and the theoretical understanding of nonlinear equations must also advance. The "computational science and engineering centers" will accelerate the solution of these problems. Massively parallel systems are the most promising routes to major improvements in speed and they will require not only more powerful software, but also the availability of reliable, cost-effective, and innovative semiconductor devices produced in the United States. Therefore, support for increased basic scientific and engineering studies of advanced semiconductor and other innovative materials, such as high temperature superconductors, is essential to restore the lost U.S. leadership in the design and manufacturing of semiconductor components.

## Access

The National Supercomputer Centers, developed by DOE, NASA, and NSF, have provided leadership and visibility and the Panel recognizes the dramatic improvements in access that these centers have been able to achieve in the last 5 years, but is unanimous in noting that access is a continuing problem. Many of the national centers are nearing saturation. Moreover, centers must be rapidly upgraded to have the most advanced computing systems and orders of magnitude improvement in nationwide network facilities. These centers must also reach out to developing fields, such as economics, whose future growth depends on the use of large-scale modeling. Another nationally significant shortcoming noted by the Panel is the lack of incentives to have broader segments of U.S. industry exploit the potential of large-scale computer modeling. It is estimated that there are approximately 30,000 computer centers nationwide in academe, government, and industry. Representatives of the centers that require high-performance computing should have access to the facilities on the gigabit network.

## Education

Effective use of advanced computers requires mastery of fundamental computational principles, much like the knowledge of applied mathematics required in science and engineering. The understanding of these principles is best accomplished through carefully designed courses involving such fields as applied mathematics and computer science. Regrettably, few universities have extensive programs that advance computational science and engineering. Training was one of the four recommendations of the Lax Report. Little progress, however, has been made.

The Panel is unanimous in urging that universities receive incentives to develop both undergraduate and graduate curricula for large-scale computing in engineering and science. The curricula must be interdisciplinary and foster a deeper understanding of the underlying problems in large-scale computer modeling. The Panel further recommends that state and federal agencies cooperate with industry to provide resources for curriculum development, undergraduate summer research, and major increases in graduate and postgraduate fellowships. It is essential that the nation attract, educate, and advance more researchers in high-performance computing.

## Bibliography

[1] Report of the Panel on Large Scale Computing in Science and Engineering, Peter Lax, Chairman, Sponsored by the U.S. Department of Defense and the National Science Foundation, in cooperation with the Department of Energy and the National Aeronautics and Space Administration, Washington, D.C., December 26, 1982.

[2] A National Computing Environment for Academic Research. Marcel Bardon and Kent Curtis, NSF Working Group on Computers for Research. National Science Foundation, July 1983.

[3] A Report of the Panel on Future Directions in Computational Mathematics Algorithms and Scientific Software. Werner C. Rheinboldt, Chairman. SIAM, 1985.

[4] David Report -- Renewing U.S. Mathematics, Critical Resources for the Future. Report of the Committee on Resources for the Mathematical Sciences, National Research Council, Edward E. David, Jr., Chairman, Washington, D.C., 1984.

[5] Scientific Supercomputer Committee Report. Institute of Electrical and Electronics Engineers, Sidney Fernbach, Chairman, Washington, D.C., Oct. 25, 1983.

[6] Federal Coordinating Council on Science, Engineering and Technology. Report of the Supercomputer Panel on Recommended Acting to Retain U.S. Leadership in Supercomputers. Washington, D.C., Dec. 1983.

[7] Federal Coordinating Council on Science, Engineering and Technology Committee on High Performance Computing FY 1985 Annual Report, Washington, D.C., March 19, 1986.

[8] Federal Coordinating Council on Science, Engineering and Technology Committee on High Performance Computing FY 1986 Annual Report, Washington, D.C., January 1987.

[9] Federal Supercomputer Programs and Policies Hearing before the subcommittee on Energy Development and Applications and the Subcommittee on Science, Research and Technology of the Committee on Science and Technology. Ninety-Ninth Congress, House of Representatives, June 15, 1985 (see additional references therein).

[10] Supercomputers: Government Plans and Policies. Congress of the United States, Office of Technology Assessment, March 1986.

[11] Software for High Performance Computers, A Report, prepared by the Subcommittee on Supercomputers of the Committee on Communication and Information Policy, Institute of Electrical and Electronics Engineers, December 1985.

# Report of the Sub-Panel on Applications of High-Performance Computing in Engineering and Science

## *Outline*

## Summary

The sub-panel on applications of supercomputing focused on astrophysics, economics, engineering, materials, mathematics, plasma physics, and weather forecasting. There was unanimous agreement that high-performance computing has become firmly established as the third mode of scientific research, thereby broadening the traditional methodologies of laboratory experimentation and theoretical analysis. Examples presented in this report indicate that this new approach can both lead and anticipate experiment and theory, as well as test the results of these longer-standing methodologies. In the area of molecular biology, for example, computer simulations of the hydration of the Na+ ion around B-DNA revealed ordered strings of solvent molecules before they were confirmed by laboratory experimentation. High-performance computing is known to be a powerful tool in unraveling other complex molecular level dynamics. For example, the mechanism for the twisting of the polymer tetrafluoroethylene (Teflon) has only been recently identified (with aid of computer studies), even though the phenomenon was observed in the laboratory over 30 years ago. The model for the structure of Teflon was found from large-scale calculations on perfluoro-n-alkanes, which are short, isolated chains of polymer. By building on these studies, chemists and material scientists are working to advance the design of new polymers with innovative properties. This progress accelerates gains in an economically vital area of materials.

We are unanimous in our endorsement of the National Computing Initiative and the need for increased graduate and postgraduate fellowships. A key feature of high-performance computing is the opportunity for different disciplines to focus their strengths on the same problems. Thus, aeronautical engineers working with applied mathematicians and computer scientists can share common interests in the modeling of three-dimensional supersonic flow around aircraft.

We urge the creation of undergraduate and graduate curricula that are designed to provide a strong foundation in large-scale engineering and scientific computation. Basic skills in high-performance computing are analogous to the pervasive use of applied mathematics in engineering and science. Researchers in the future will require not only knowledge of applied mathematical analyses, but also of areas such as numerical analysis, algorithm design, programming languages, and parallel processing. We also urge the creation of federally and industrially sponsored incentives to assist universities in the development of new curricula. These

curricula could extend to industry through the increasing use of off-campus education.

The National Supercomputer Centers have had a major impact on the significant advances made by the applications of high-performance computing. The Panel members, however, are in complete agreement about the need for these Centers to have the most advanced computing technologies linked through a national gigabit network. One important national goal is the development of computers that can be upgraded to massive levels of parallelism and that are operated by efficient, user-oriented software. There is broad consensus that the United States must strive to provide several orders of magnitude increases in computing power over 5-year periods. Without decisive action, the nation will not be strategically positioned to exploit one of its most valued native technologies. We see every major area of engineering and scientific research as benefiting from this thrust. United States leadership requires a focused effort to accelerate the methologies and technologies that advance high-performance computing, which include areas of computer science, applied mathematics, materials science, and electrical engineering. Three tables at the end of this report illustrate recent achievements in these fields and some of the grand challenges that lie ahead.

## Weather Forecasting

Advancing computational power has produced continually increasing improvement in weather prediction through simulation of future atmospheric conditions evolving from observed initial states. Because spatial resolution is one of the determinants of forecast accuracy, the major national and international weather prediction centers operate at the limit of computational technology and are among the first customers for advanced machines.

These national and international forecast centers all operate one or more supercomputers full time in data assimilation, numerical prediction, and model development. Currently, the most successful predictions are made with a Cray X-MP/48 by the European Center for Medium-Range Weather Forecasting (ECMWF), operated by a 17-nation consortium. ECMWF forecasts are as accurate at 6 days as are those made with the U.S. National Meteorological Center CDC 205 at 5 days, although factors other than machine power have impact on the relative performance. The record of both centers shows increased accuracy with each significant increase in computational capacity (see Fig. 1). The stated goal of the U.S. National Weather Service is to achieve the same accuracy at 8 days

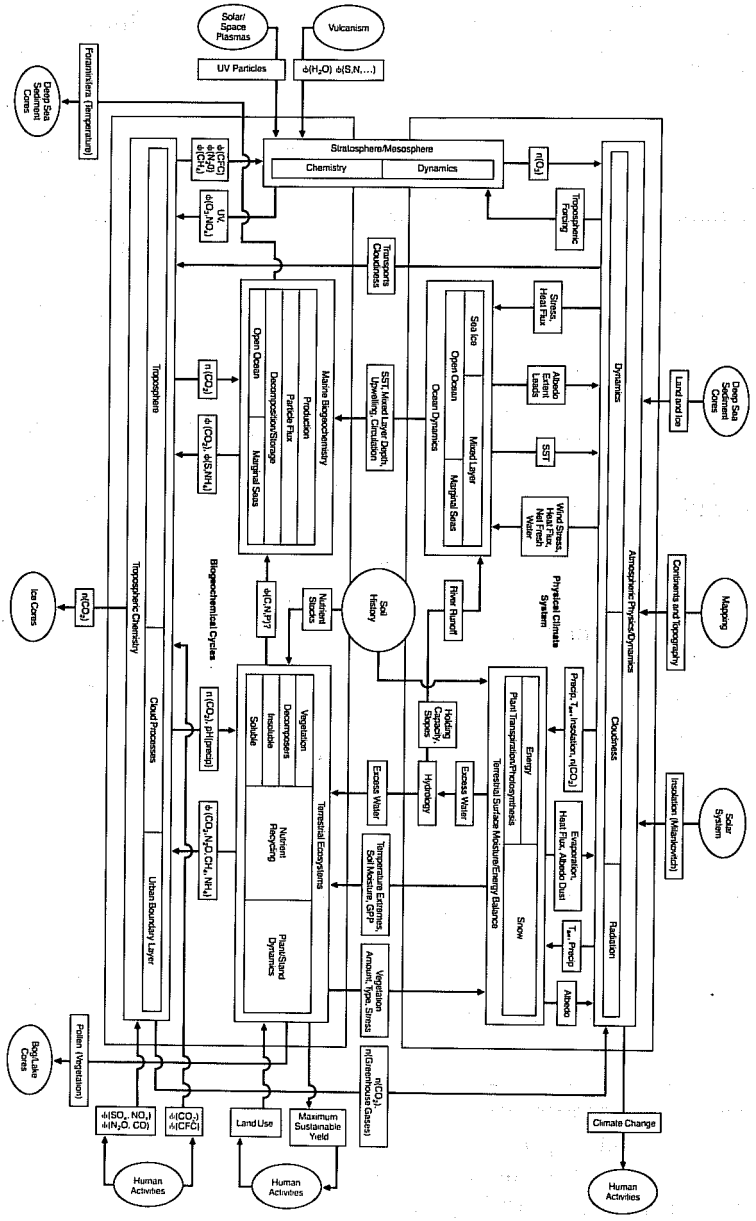

Solar/ Space Plasmas
Vulcanism
UV Particles
δ(H2O) δ(S,N,...)
Foraminifera (Temperature)
Deep Sea Sediment Cores
Stratosphere/Mesosphere
Chemistry
Dynamics
n(O3)
Tropospheric Forcing
UV
δ(O3,NO2)
δ(CO2) δ(N2O) δ(CH4)
Transports Ozone/Gas
Stress, Heat Flux
Albedo Ice/Land
SST
Wind Stress, Heat Flux, Net Fresh Water
Sea Ice
Open Ocean
Mixed Layer
Marginal Seas
Ocean Dynamics
Land and Ice
Deep Sea Sediment Cores
Dynamics
Atmospheric Physics/Dynamics
Clouds/ice
Radiation
Continents and Topography
Mapping
Insolation (Milankovitch)
Solar System
Marine Biogeochemistry
Production
Particle Flux
Decomposition/Storage
Open Ocean
Marginal Seas
n(CO2)
e(CO2) δ(S,NH4)
SST, Mixed Layer Depth, Upwelling, Circulation
Ice Cores
n(CO2)
Troposphere
Troposphere Chemistry
Cloud Processes
Urban Boundary Layer
Biogeochemical Cycles
δ(C,N,P)?
Nutrient Stocks
Soil History
River Runoff
Physical Climate System
Precip, Tair, Insolation, n(CO2)
Holding Capacity, Slopes
Hydrology
Excess Water
Energy
Plant Transpiration/Photosynthesis
Terrestrial Surface Moisture/Energy Balance
Snow
Heat Flux, Albedo Dust
Evaporation
Tair, Precip
Albedo
n(CO2), pH(precip)
δ (CO2,N2O,CH4, NH3)
Soluble
Insoluble
Decomposers
Vegetation
Nutrient Recycling
Plants/Stand Dynamics
Terrestrial Ecosystems
Temperature Extremes, Soil Moisture, GPP
Vegetation Amount, Type, Stress
Pollen (Vegetation)
Bog/Lake Cores
δ(SO2, NO3) δ(N2O, CO)
δ(CO2) δ(CFC)
Land Use
Maximum Sustainable Yield
n(Greenhouse Gases)
Climate Change
Human Activities
Human Activities
Human Activities

Fig. 1.

as is now obtained at 4 days, through increased resolution and improved observations, including data from polar-orbiting and geosynchronous satellites.

Supercomputers are also essential in modern atmospheric research. Among the notable successes and continuing challenges are detailed models of entire systems such as thunderstorms, squall lines and large-scale convective systems, and intense cyclones. Most of these models are run at government centers or at the National Center for Atmospheric Research (NCAR), where two Cray 1 computers are fully utilized for atmospheric and oceanic research; a Cray X-MP/48 is being acquired for the Center.

In simulation of the evolution of atmospheric systems, accuracy is largely dependent on adequate modeling of the nonlinear interactions that spread energy among the various spatial scales of the flow. Because of nonlinearity, small-scale processes combine to produce large-scale effects with the consequence that improved model resolution and improved portrayal of small-scale effects will produce improved large-scale accuracy. The coupling of high-resolution models with greatly improved observations, including detailed wind fields observed with a national network of Doppler radars now being deployed, is expected to provide a dramatic improvement in the prediction of local surface weather events that affect human activities. Current computational capabilities have led to the development of nested models, in which a high-resolution model is selectively imposed on a subregion of a large-scale model to produce detailed predictions of significant features, such as time and location of hurricane landfalls.

Simulation of the climate and prediction of climate anomalies require that the combined response produced by atmospheric and oceanic interactions be resolved and accurately modeled. Detailed three-dimensional modeling of oceanic circulations is just becoming feasible, as computational power becomes adequate to resolve the energy-containing features in the ocean -- structures an order of magnitude smaller than the energy-containing scales in the atmosphere.

Such interactive ocean-atmosphere models are a precursor to the grand challenge of geophysical modeling: developing a model that will simulate the evolution of the entire Earth System on the scale of decades to centuries. The increasing concentration of radioactively significant gases in the atmosphere, including carbon dioxide and methane, will produce changes in the earth's surface environment, including increased surface temperature and new precipitation patterns, whose magnitude and consequence are now only dimly perceived.

The scientific imperatives created by global change are recognized in national and international efforts -- the *Global Geosciences* initiative of NSF, the *Earth System Science* initiative of NASA, and the *International Geosphere-Biosphere Program: A Study of Global Change* recently endorsed by the International Council of Scientific Unions (ICSU). The diverse observational efforts of these programs will be aimed at understanding the interactions between the atmosphere, ocean, land surface, biosphere, and cryosphere that control the evolution of the Earth System and that will provide the empirical knowledge necessary to develop an Earth System model. Once verified by comparison with present conditions and with geological and paleoclimatological records of past climatic variations, such a model will be useful in scenario studies designed to estimate the consequences of human activities related to energy consumption, pollution control, and modification of the earth's surface and its vegetative cover. Eventually such a model of the physical, chemical, and biological components of the Earth System will be combined with models of the economic, demographic, agricultural, and resource usage patterns of human society to produce the simulations necessary to understand, and perhaps guide, the future of the planet.

The components of the Earth System, as with the atmosphere, evolve nonlinearly and interact in nonlinear ways with the other components. Thus, small-scale phenomena act in sum to modify large-scale evolution with the consequence that the accuracy of an Earth System model will depend in part on resolution. At the present state of knowledge about the system, highly aggregated models summarizing these interactions parametrically will provide an initial understanding of the challenge of modeling the Earth System. As knowledge about the earth and experience in modeling the system accumulate, the models will increase dramatically in resolution and complexity and then will be limited by the computational capacity expected to be available in the foreseeable future.

Before this limitation is reached, however, the efforts to understand and model global change are more likely to be hampered by an inability to use the expected flood of observed data effectively. The operational observations of atmospheric and oceanic conditions, the NASA earth science research missions now in development or planned, and the Earth Observation System (EOS) being developed for launch as part of the Space Station Initiative will produce data at unprecedented rates -- 1 terabyte per day for the EOS system alone. These data must be processed into scientifically meaningful forms, entered into active databases,

made available for research and education to earth scientists throughout the nation and world, and archived in order to permit us to detect global change.

Plans for extensive observational programs in the space and earth sciences and the data sets that will be created by computer simulations of complex processes in science and engineering mandate new approaches to the management of data. Current technology and current conceptual approaches would be overwhelmed by the data streams that will be produced before the end of the century. Even if these data could be processed and stored in active databases, the technology does not yet exist to provide the access necessary for effective scientific use. Networks must be created that will allow scientists to work interactively with these extensive data sets, to transport them for intensive local analysis, and to return meaningful results to the databases in order to increase the value of the holdings.

The most serious difficulty anticipated today will occur in the middle of the next decade with the data streams from satellite observations of the earth. Similar difficulties can be expected before the end of the century as computer simulations augment, or replace, laboratory experimentation in diverse fields, including materials science and engineering, biology, fluid dynamics, and aeronautical engineering.

Thus, we urge that the development of effective concepts and technologies for the management of databases containing $10^{16}$ bytes be assigned a high priority. Management issues and networking concepts are critical, because such databases within disciplines will undoubtedly be distributed among groups of scientists who must share data in interactive modes and use them for model verification and initialization.

## Engineering

High-performance computing is playing an increasingly important role in engineering research and development. As in other applications, computational approaches have rapidly achieved a status equal to that of the classical scientific methods of analytical theory and laboratory experiment. Computational approaches are currently embedded deeply in many aspects of engineering. Effective computing holds the potential to vastly improve both the productivity of engineers and the reliability of engineering designs. In addition, state-of-the-art supercomputers are an enabling technology, permitting advances that are not otherwise attainable.

These increases in productivity and innovation will help keep the United States competitive.

In the current globally sensitive economy, it is necessary to complete engineering analyses and product designs in a timely way to meet market demands. High-performance computing plays a crucial role in this accelerated development schedule. For example, the time required from design to market of a color picture tube has dropped from about 4 years to about 6 months. Computational modeling is an indispensable tool in the design process, by effectively isolating design options.

Despite the pervasiveness of computing in engineering today, computational engineering is difficult to define in a brief statement. However, key components can be identified. The field of computational engineering is interdepartmental, drawing on expertise from mathematics and computer science, but most effectively feeding on the driving applications. Indeed, the history of large-scale engineering and scientific computing has shown that the most effective and novel algorithms have been developed in response to the need for the solution to specific problems. Examples include the development of finite element methods by structural engineers, fast Fourier transforms by statisticians, and relaxation methods to solve problems in petroleum and nuclear engineering. Each of these developments was driven by large-scale scientific computing, because the development of the algorithms was necessary for solving these problems on the fastest machines of their time.

Computational engineering is the union of its areas of application so it cannot be isolated as a single discipline. What seems to be happening in many academic and industrial organizations is that computational engineering is being organized across traditional departmental boundaries, forming "virtual departments" or institutes of computational scientists with common interests. We believe that this organizational structure is good because the mix of engineers, computer scientists, and mathematicians is effective in problem solving.

In order to put the subject of computational engineering in perspective, we will look at some specific examples of the utility of supercomputing over the last decade.

### *Issues*

Major advances in engineering are made by seeking solutions to specific problems. Such solutions typically involve the development of new fundamental theory, mathematical tools,

algorithms, software, laboratory experimental techniques, data reduction methods, etc. Those projects that open major new lines of inquiry and lead to new agendas for science and engineering research usually require teams with strong focus. An intradisciplinary team of this sort may be composed of members from many classical departments (such as physics, aeronautical engineering, mathematics), who share a strong common interest in the specific application.

Two examples are given below.

*Example 1.* Analysis of the fluid and solid mechanics of microstructured materials is one of the most challenging areas of continuum mechanics. In this classification we include understanding of mechanical behavior of multiphase materials ranging in size from melt and solutions of polymer molecules with mean sizes of tens of Angstroms to the colloidal suspensions of micron-sized particles used in ceramic processing to suspensions of larger particles common in processing of slurries and in liquid-gas multiphase system. Detailed analysis of these interactions is important from several perspectives.

Flows can purposely or inadvertently alter the morphology of microstructured materials in many materials processes. Examples of such systems are the weld lines introduced in injected polymer parts of stagnation points in the flow pattern and the dependence of details of the dendritic morphology of a metal crystal on the conditions for solidification. Successful modeling of these interactions of the microstructure with the processing conditions is only possible when the mechanics and transport processes on the smallest length scales of the material are well understood. The production rates and product quality of other processes are often limited by more macroscopic instabilities caused by interactions of the rheology of the system with process conditions. Melt fracture at high pull rates of crystalline polymer fibers is an example.

Large-scale computing will serve two important roles in helping to solve these problems. First, accurate constitutive equations describing the mechanical behavior of these materials must be derived. The complexity of any continuum model that is derivable from even a simple micromechanical picture of the material makes closed-form analysis nearly impossible and amplifies the role of large-scale computing in probing the ramifications of the form of the constitutive equations on predictions for meaningful problems.

Large-scale computing must also play an increasingly important role in testing the framework for constructing continuum models from first-principle descriptions of fluids with

microstructure. For example, the validity of assumptions on the statistical averaging needed to describe the rheology of a nondilute suspension of macroscopic particles in a Newtonian solvent can be tested by calculation of the dynamics of large collections of these particles in an imposed flow. Statistical averaging of these dynamics gives the equivalent of molecular dynamics results that can serve as input into the derivation of averaged or continuum equations for description of macroscopic phenomena.

*Example 2.* Turbulence (meaning here fluid turbulence) is one of the most challenging problems of modern physics. Recently, major progress has been made by an integrated attack on the problem using advanced methods of theoretical physics, novel computer hardware and software, and new laboratory experimental techniques. On the one hand, the application of renormalization group techniques popularized in statistical physics and field theory have allowed the development of models of turbulence that appear to be both in good agreement with available data and to be sufficiently tractable that they allow efficient computer solution. This theoretical advance has suggested that experimental data of a different sort from that previously available are necessary to make further progress on the turbulence problem. In particular, it is now becoming apparent that information on the basic geometry and dynamics of turbulence can best be obtained through understanding of complete flow fields in space. Unfortunately, classical measurement techniques provide either accurate single (or few) point data or qualitative, pictorial projections of the full flow field in space.

In order to obtain quantitative data on full spatial flow fields, new experimental methods based on such advanced tools as laser sheet scanning, nuclear magnetic resonance, and electronic spin resonance are under development. These new experimental methods, developed by applied physicists, generate so much data that laboratory-sized computers are unable to properly process the generated data. In order to achieve the required data reduction, another team of fluid dynamicists-computer hardware/software designers have developed special-purpose laboratory-sized supercomputers under NASA's support. These desktop supercomputers, which are capable of processing data at speeds comparable to that of a Cray supercomputer, are the key to developing new insights into mechanisms of analysis and control of turbulence. At the same time these new computers provide tools to solve simultaneously the underlying fluid dynamical equations and turbulence models and to compare the results directly with experiment. This intradisciplinary team of mathematicians, physicists, applied physicists, and systems engineers gives a modern, effective way to achieve progress on these difficult problems.

*Recommendations*

Funding agencies must encourage and foster strong interdisciplinary proposals that combine supercomputing with experiments and theory to open new agendas in science and engineering.

The subject of high-performance computational science and engineering is now developing so rapidly to a mature state that it is necessary to begin formalizing the education, training, and retraining of the engineering community. Students now begin graduate study without an adequate introduction to scientific computing. At both the undergraduate and graduate level, there has been little effort to develop curricula in computational science and engineering. Since an increasing number of engineers will use the tools of high-performance computing as a major part of their work, such curriculum development is essential.

In addition, it is necessary to begin planning for extensive programs of continuing education for the engineering community in order to meet the competitive thrust of the global economy. It is noteworthy that many Japanese industries have been aggressively placing engineers in continuing-education programs in high technology and computing related fields at major U.S. universities. It is surprising that much of U.S. industry has been less aggressive in taking advantage of these opportunities. While the U.S. engineering industry is beginning to recognize the need for sophisticated computing, there is little opportunity in either academe or the national laboratories for industry to provide the required stimulus and training in the new ideas of high-performance computing. It is in our strong national interest to provide this stimulus immediately.

A concerted national effort, perhaps similar to that of the Physical Science Study Committee's curriculum development for physics and chemistry, should be initiated for curriculum development in computational science/engineering at the undergraduate/graduate level.

Efforts to integrate industrial researchers into supercomputing activities should be encouraged. This encouragement may take the form of continuing-education programs, summer workshops, and so on, perhaps coordinated by local universities and the National Supercomputer Centers.

## Mathematics

Computational methods are fundamental to all aspects of science, and mathematics is an important part of the foundation and intellectual basis of these methods. In order to utilize today's high-speed computing machines, new techniques have been devised. Not only has this provided a serious challenge to the applied mathematician, it has also placed new and difficult problems on the desk of the theorist; algorithms themselves have become an object of serious investigation. Their refinement and improvement has become at least as important to the speed and utility of high-speed computing, in some areas at least, as the improvement of hardware; for example, it is estimated that the speedup in solving elliptic boundary-value problems gained during the years from 1945 to 1975 due to improved numerical methods has been greater than the speed of the Cray 1 over the IBM 650.

Many mathematical ideas have been central to the development of new techniques for scientific computation, for example:

- The methods of alternating direction and fractional step now used in a wide variety of problems;

- Higher-order difference schemes have been used extensively in meteorological calculations;

- A collection of techniques and ideas that has come to be described as the implicit method is used to calculate fluid flows and has been used a great deal in the field of magneto-hydrodynamics;

- The problem of calculating flows with shocks -- a problem of great practical importance -- has generated a number of valuable mathematical ideas that have borne fruit (flux-corrected transport, artificial compression, solutions based upon multiple solutions to Riemann's initial value problem, complex coordinate methods);

- Ideas from advanced geometry have recently made an impact on the speed of calculation of solution to linear optimization problems;

- Advanced statistical methods and ideas from modern probability have led to new algorithms, based on the idea of Gibbs distributions, that are inherently parallelizable and have already been successfully applied to routing problems, network layout on chips, and digital image restoration.

At the same time, computation has become an integral part of many fields of mathematics. Not only those areas whose direct concerns include matters of computation (complexity theory, numerical analysis, for example) but others, even ones long considered to be abstract, have increasingly relied upon numerical examples. These examples have come in the form of suggestive numerical simulations:

- Several long-standing conjectures have recently been proved -- in number theory and in statistical matrix theory, for example -- guidance and, equally important, the courage to continue extremely intricate and difficult theoretical work;

- Recent advances in knot theory have suggested algebraic computations that are now being carried out on computers to gain further understanding (This work has direct application to the topology of molecules and DNA.);

- Analysis of numerical computation involving iterated maps led to conjecture of universal laws about their behavior that are now being established by rigorous mathematical means. The theory of the chaotic behavior of dynamical systems depends fundamentally on numerical simulations; the concept of a strange attractor was formulated to understand the results of a series of numerical computations;

- Several important estimates for the eigenvalues of the Laplace operator for more general geometries were first found by numerical estimation on a computer. The computations by machine do not play a part in the final proof, but they were essential in establishing what was to be proved;

- In the field of geometry and the calculus of variations, computers have been used recently to discover new examples of minimal surfaces, examples whose analytic form was too difficult to analyze directly without numerical simulation. The simulations were understood by the use of computer

graphics and led to the explicit construction of infinite families of new examples.

Computation and abstraction are symbiotic processes. Computing power has matured to the point where mathematical examples too complicated to be understood analytically can now be computed and observed. Insights gained in this manner have led to theoretical advances in the areas mentioned above as well as others, including those as "abstract" as the topology and geometry of three-manifolds.

The modern computer is the first laboratory instrument that mathematicians have ever had. Not only is it being used increasingly for research in pure mathematics, but, equally important, the prevalence of scientific computing in other fields has provided the medium for communication between the mathematician and the physical scientist. The fruitful reunion should be fostered by continuing support for high-quality interdisciplinary work centered around well-defined, important scientific and mathematical problems.

Some of the research mentioned above and all of the work mentioned at the end of this section could make good use of 10 to 1000 times the currently available computing power. Special computing centers should continue to provide the most powerful equipment available and should be concerned with making this computational power accessible to researchers in these areas. There should be, at the same time, support of many more centers with the "next level down" of computational capability. These secondary centers should have very high speed communication with the larger centers and make it a priority to have, as much as possible, equipment compatible with national centers; the use of machines on both sites for a single computational task should be as transparent as possible. In addition, there is relatively little research support for professional programming assistance for mathematical research. This will be an increasing need that will be made all the more important by the implementation of the recommendations of this report.

It should be emphasized that mathematics is one of the last scientific disciplines to be computerized. More than in other fields, there is a lack of instrumentation and training. This prevents the mathematician from using high-performance computing in attacking research problems and at the same time isolates him/her from productive communication with scientific colleagues. This is a serious problem of education, training, and research funding priorities.

In conclusion, we mention two of many areas of mathematics that will require increased computing power and are likely to produce significant breakthroughs by the use of this power:

- Computational study of the partial differential equations associated with three-dimensional problems involving turbulence, unstable regimes, or unusual geometry, particularly those associated with fluid dynamics and gas flow (the Navier-Stokes equation and the Boltzmann equation). Some computations now being done in the study of roll-up and twisting of vortex sheets have taken approximately 100 hours on a Cray 1 and are relatively simple compared to the full range of problems;

- Partially free boundary-value problems in the calculus of variations arising from the study of interfaces arising in microemulsions, the behavior of thin films, block copolymers, and crystals. Here, new equilibrium structures might be identified by computational methods.

## Materials Science

The application of high-performance computing to materials science is still in an early stage. There are, however, numerous areas in which it has had a significant impact. The future will see even greater progress, as more scientists and engineers have access to powerful computing facilities.

Indeed, we are entering an era in which the use of supercomputers allows scientists to "design" materials with innovative properties, based on the predictions of computer modeling. Predictive simulations will provide a cost-effective alternative to trial and error design. We will discuss a few examples in which computer models have provided insight and, to some degree, prediction. We will also briefly summarize critical areas of technological development that may influence future supercomputer performance and then summarize our recommendations.

### *Recent Progress*

- The use of high-performance computing and its increased availability have greatly influenced the simulations of electronic transport in semiconductors and semiconductor devices. Two- and three-dimensional device and process

simulations using supercomputers have been successfully applied to silicon technology and are cost-effective in device optimization problems. Transient effects related to impact ionization and hot electron emission into silicon dioxide have been treated and require Monte Carlo simulations that make the use of large-scale computational resources imperative since the inclusion of complicated band-structure effects is necessary.

Simulation of Gallium arsenide (GaAs) devices necessitates even more the use of vector and parallel computing because of the inherent complexity of the equations of motion in this material and the pronounced transient effects (overshoot, ballistic transport).

First attempts have been made to include more precise Monte Carlo computations into standard device simulators, Monte Carlo "windows" in the PISCES code and high-precision results have been achieved. Experimental research is directly influenced by the optimization of device structures and in turn provides parameters necessary for the computation.

- A massive Monte Carlo simulation has led to a major breakthrough in our understanding of spin glasses, which provide an important example of the role of randomness in magnetic phase transitions. This study resolved a major issue in the theory of second order phase transitions, which had previously been unresolved by either theoretical or computer studies. In the latter case this was due to the unusually long equilibration times with the three-dimensional model of interest.

- Atomistic simulation of semiconductor crystal growth has provided information on solidification (growth from the melt) and on epitaxial growth from olecular beams. This has been achieved by the calculation of atomic trajectories using classical mechanics (molecular dynamics) and three-body interatomic potentials for silicon and germanium. The crystal-melt interface exhibits a large-amplitude hill-and-valley structure on the Si (100) interface which contrasts sharply with the atomically flat (111). Simulations of crystal growth on these two orientations have been attempted for crystal-melt and crystal-vapor systems. In both cases, material deposited on the (111) becomes defective or amorphous at a higher temperature than for the

(100) face. These results are consistent with experimental observation of rapidly solidified silicon using pulsed laser techniques and have stimulated attempts to monitor the structure of these two interfaces during crystal growth. The stability of strained epitaxial layers of GeSi alloys on Si has also been examined by computer simulation. These studies predict that a defect-free film may remain in a metastable coherent state at a strain that is considerably larger than the maximum value obtained in experiments. This result suggests that misfits larger than the 4% Ge/Si could be accommodated in carefully prepared systems. This could have an impact on the design of strained layer devices.

- Computer simulations often provide insights that influence theoretical development. The existence of a surface roughening transition was indicated in early Monte Carlo simulations of close-packed Ising model surfaces. Qualitative changes were observed in the structure and dynamics above a critical temperature specific to the surface orientation. These results were later confirmed when analytical work showed that the transition is of the Kosterlitz-Thouless type.

- Density functional calculations have provided important insight into surface science, including the origin of surface reconstruction in several solids. Since surfaces play an important role in many areas of materials science, including catalysis, these calculations have provided a breakthrough in a field previously only understood by semi-phenomenological theories. Interaction energies between atoms of low coordination number can exceed those in the bulk by factors of two. Good agreement is obtained between theory and experiments on the dissociation energy of small clusters. Some progress has been made on techniques to combine density functional theory and molecular dynamics to produce "first principles" results for nuclear trajectories. Systems containing up to 64 atoms have been followed for short times (fractions of a picosecond). System size is limited by the fact that he number of computations increase as the third power of the number of nuclei.

*Future Challenges*

Theory and experiment in materials engineering and science will profit by high-performance computing. Simulation will

provide important guides to both theory and experiment and be used to predict novel properties of new materials "designed" by computers, such as metal/semiconductor and composite structures. A few of the major applications of large-scale computing to be expected in materials science are given below.

- Future challenges in the area of electronic materials include the development of a complete quantum transport theory that is capable of handling tunneling phenomena, electron-phonon interactions in the presence of a large electric field, as well as transient phenomena in complex geometries (real-time quantum dynamics).

  Computation of these effects using, for example, Feynman path integrals, calls for even larger computational resources and will be important for the understanding of the ultimate limitations of new forms of heterolayer devices.

- One of the major challenges in materials science is the development of interatomic potentials that describe the interaction between atoms over a range of conditions. These should be cast in a form suitable for efficient computation, and would ideally incorporate quantum mechanical results and fit experimental measurements of materials properties.

Examples of materials for which realistic interatomic potentials are necessary include semiconductors such as GaAs, semi-conductor alloys such as GeSi/Si, and metal silicides. Such potentials will allow one to "design" the properties of semiconductor heterostructures, metal-semiconductor heterostructures, and possibly structural composites by computer simulation. This will in some cases involve computer simulation of crystal growth by, for example, molecular beam epitaxy (MBE).

Preliminary molecular dynamics studies of growth under MBE conditions for both a Lennard-Jones model and Si(100) (involving two- and three-body forces) suggest that realistic studies, given sufficient computational facilities, are quite possible.

- Systems far from equilibrium constitute a fundamental research area that will benefit substantially from increased computer power. These include dendrite growth and directional solidification, spinodal decomposition and coarsening, and rapid solidification. Large-scale computational studies of these topics in the past few years have indicated that solutions of such problems are possible,

although they may require more powerful computing facilities than currently available to completely determine system size effects and initial transients.

### *Issues*

Cross-fertilization of supercomputing and research on VLSI, as well as newly emerging technologies such as GaAs MBE, is imperative. It has already been established that the design of small devices, the components of future supercomputers, requires current supercomputers for efficient simulation and optimization. Recent investigations have shown the validity of the concept of the optically interconnected computer, and new areas are opening up which include semiconductor heterolayers such as GaAs silicon.

The cross-fertilization of high-performance computing with theory and experiment is coming to fruition in the application of picosecond and femtosecond transient phenomena. This has been clearly demonstrated in the area of very small electronic devices. Optoelectronics is a fertile ground for future research in large-scale computation as witnessed by recent developments in femtosecond spectroscopy.

### *Recommendations*

The significant progress in large-scale computational materials research and the future challenges make it clear that this area deserves increased resources and attention. Some of the barriers in the way of success and solution can be removed by further increases of computer power, including special purpose machines. Other problems call for increased interaction of the investigators and the formation of national thrusts. We, therefore, recommend the development of:

- National thrust areas in materials science that are related to and fostered by the National Supercomputer Centers. Some of these thrusts should be focused on the outstanding problems that have been discussed above including the development of realistic interatomic potentials in materials science, electronic semiconductor materials and new forms of heterolayer structures as well as metal-semiconductor multilayers. These thrusts should be geared to deal with important fundamental questions such as standardization of parameters needed for the computations, exchange of

preprints and avoidance of unnecessary redundancies, and ultimately the exchange of software.

- More emphasis on visitor programs and financial support of visitors to support the national thrusts and experience exchange and to foster new and promising areas.

- Continued support in large-scale computational science projects for single investigators, interdisciplinary research groups (as recommended in the Rheinboldt Report), and the development of undergraduate and graduate curricula for large-scale computing in engineering and science consistent with the recommendations on education in the Lax Report.

- Progress in many areas of materials science would be greatly expedited by access to computational facilities of a teraflop capability. The cost-effectiveness of this facility should, however, be compared with that of a distributed system with the equivalent computing power (plus concomitant memory and I/O capacity). Particular areas in materials research that could profit from such enhanced computing power include determination of realistic interatomic potentials, particularly from first principles quantum mechanical calculations; solution of nonequilibrium problems involving dendritic growth, directional solidification, and spinodal decomposition and coarsening in a variety of different materials (growth mechanisms in the kinetics of phase transitions can be material-dependent, e.g., different for binary alloys and for long-chain polymer molecules); computer modeling of crystal growth under MBE conditions, with the goal of designing novel materials in semiconductor physics and composite structures that will have enhanced properties.

*Special Purpose Computers at National Supercomputer Centers*

The national centers are appropriate locations for the development of special purpose computers. The success of these projects requires the support of a staff with expertise in circuit design, microprocessor technology and software, and with research interests that require intensive computation. The advantage of these computers is the ability to tailor the hardware to the

application and thus make efficient use of the microprocessors and/or floating point accelerators.

The Monte Carlo Ising model computer designed at AT&T Bell Laboratories is an excellent example. The Monte Carlo code runs faster by a factor of 5-10, depending on the system, than Fortran programs on a Cray 1. The design, implementation, and programming required about one staff year. A special purpose computer for molecular dynamics computations with two-body potentials was constructed at the Technical University in Delft and has been applied to studies of two-dimensional melting in large systems (16,000 particles) and to liquid-vapor interface studies. Machines with more flexibility in applications, but designed primarily for molecular dynamics, are being constructed at AT&T Bell Laboratories and IBM/San Jose. These computers will incorporate parallel architectures, with the ability to increase their power by adding floating point processor boards. Typically, the hardware cost for a system in the 100 M flop range is $50,000.

## Plasma Physics

### *Plasma*

Plasmas, high-temperature ionized gases, are one of the most complex fluids one has to deal with, because they are subject to electric and magnetic forces as well as pressure forces. Plasma motions are often nonlinear and turbulent and exhibit many subtle phenomena due to the highly energetic particles present. The only way to attack many of the problems encountered (like other complex fluid problems) is through high-performance computing. Thus, the impact of this approach on plasma physics has been fundamental and profound.

The confinement and control of very hot plasmas ($T = 10^8$ K) is central to achieving controlled fusion energy. Great strides have been made on this research in recent years; strides that almost certainly would not have been made without the help of large-scale computing. Almost all theory related to plasma confinement and its behavior in fusion devices involves large-scale computing as measured by the fraction of fusion theory papers that involve it (over 50% in a recent issue of *Physics of Fluids*, Vol. 29, No. 8 [Aug. 6]). Codes to calculate plasma equilibrium in a magnetic field, plasma stability properties and plasma and energy transport are integral parts of experiments in this area. They are used to interpret experimental results; they are a central part of designing new experiments by predicting performance. Large-scale computing is

involved in the engineering design of the complex machines used in this research.

In the area of space plasma physics, the use of large-scale computer models has also come to be an essential tool. The interaction of the solar wind with the earth's magnetic field determines conditions in the earth's magnetosphere, a region where many satellites operate. The solar wind-magnetosphere interaction is a complex dynamical one whose properties can be probed at only a limited number of points by satellites; computations are required to fill in the gaps. These involve many space scales (the thickness of the atmosphere, the size of the earth, the size of the magnetosphere, the scale of the magnetotail in the wake of the earth, the size of the shock wave surrounding the earth). Important advances have been made in recent years of our understanding of the interaction between the solar wind and earth magnetosphere interaction (also that for other planets).

More specifically, the aurora borealis that occurs in the polar regions of the earth has awed man since he first saw it. There was little understanding of the phenomena until recent exploration of the magnetosphere by satellites. Even now our understanding is more qualitative than quantitative. Aside from this intrinsic interest, the precipitation of energetic particles into the atmosphere (associated with an aurora) affects the atmosphere. Magnetic storms are associated with such activity and can disrupt communications and cause large electric currents in such things as the Alaska pipeline and in power and telephone grids. With the operation of satellites in the earth's magnetosphere and the coming of the proposed space station, understanding of the magnetosphere is becoming increasingly more important. Important examples of computational successes in plasma physics are provided by magnetospheric modeling. These would not have been possible without the present generation of supercomputers. As one example, such calculations have shown that when the solar wind interacts with the earth's magnetosphere large vortices are produced which act like large magnetohydrodynamic generators driving the auroral currents. The modeling results, although still crude, show very encouraging correspondence with observations (the so-called O aurora was found almost simultaneously by satellite observation and computer modeling).

New plasma physics has been discovered using computer modeling. An outstanding example of this is computations that show the possibility of using plasmas in high energy physics as an accelerating structure and as focusing elements. The acceleration possibilities arise because plasmas can support waves moving at

close to the speed of light that produce very large electric fields (109 V/cm). These waves can be generated by the nonlinear beating of two intense laser beams or by bunches of energetic electrons passing through the plasma. The lens properties arise from the electromagnetic forces which act on a bunch of charged particles passing through a plasma. All these processes are highly nonlinear and would involve complex, expensive, experimental apparatus to explore. Computer modeling has proved to be a powerful, relatively inexpensive, but also a rather detailed method for investigating the physics involved. It has proved itself to be an excellent guide to experiments. Modeling would be an invaluable design tool if applications of this plasma technology for accelerating particles are found in industry, medicine, and defense.

Other examples from plasma physics are modeling of the operation of microwave devices for radar, plasma heating, and particle acceleration where computing is becoming an important part of the procedure for designing new devices. Very large-scale computer modeling is showing ways plasma can be used to improve particle accelerators, which are used in high-energy physics, materials research, industry, and medicine. Plasma modeling is important for the design of high-powered electrical switches and for an understanding of the effects of lightening. Finally, the importance of computer plasma modeling in fusion research has been absolutely vital for the success and interpretation of fusion experiments. All applications mentioned above are still limited by grid resolution and by our inability to deal with either a wide range of space and time scales or to include many aspects of the physics which we know are important. As a measure of how present computers are limited, a typical running time for the problems mentioned above was in the range of a few hours on a Cray X-MP; some of them took 100 hours. Computing power was a severe limiting factor on the size and degree of realism that could be included in the calculation. Several orders of magnitude increase in power would find immediate and productive use.

### *Astronomy*

*Theoretical Astrophysics.* Many problems involve using the computer to simulate situations known to exist on a large scale in the universe, but for which the parameter space is not accessible in the laboratory. Simulations now being conducted include celestial mechanics, star clusters, star formation, supernovae, galactic structure, galaxy collisions, hydrodynamics of stellar atmospheres,

relativistic plasma jets from active galactic nuclei and quasars, and the formation of structure in the early universe.

The dynamical evolution of gravitationally bound systems of particles, such as in star clusters, galaxies, or cosmology requires the self-consistent calculation of forces in large aggregates of particles. For example, to understand the stability of the disk of a galaxy, and hence the problem of the persistence of spiral structure, requires about $3 \times 10^5$ particles. Every particle "feels" the effect of 300 neighbors resulting in $10^8$ interactions to be calculated for each time step. To follow the evolution of such a galaxy over the 10 billion year age of the universe takes about 30,000 steps. At 30 flops/calculation the evolution of a single galaxy would require one day on a 1 G flop computer. At present the largest simulations on a Cray X-MP have made it possible to begin to understand the stability of disks of galaxies -- the persistence of spiral structure and the bar forming instabilities. However, the resolution is still too coarse to model the effects of higher-density regions such as in the nucleus of a galaxy.

The fully relativistic radiative transfer calculations needed to model the explosion of a star, a supernovae, have only just become practical through use of the current generation of supercomputers.

Astrophysical calculations of the behavior of plasma are similar to those discussed under plasma physics but extend these calculations to lower densities and to vastly larger spatial scales. One problem in plasma astrophysics is now being attacked using the Cray X-MP at the University of Illinois National Center for supercomputing. By studying the spatial growth of Kelvin-Helmholtz kink instabilities in a two-dimensional magnetized plasma jet, the filamentary appearance seen in the radio emitting lobes of a radio-galaxy (Fig. 2) has been reproduced for the 320 zones. Fully realistic simulations will require computations in three dimensions and these only become practical with a capacity 10 times that of a Cray 2.

It is the "grand challenge" in theoretical astrophysics to make this transition from two- to three-dimensional simulations. The capabilities of the current generation of supercomputers are now close to this borderline.

*Observational Astronomy.* In observational astronomy the demand for more computational capacity has increased as new instrumentation produces larger amounts of digital data, and as the demands for processing this data into usable scientific forms increase.

An extreme requirement occurs in imaging at radio wavelengths. Because of the relatively long wavelength of radio waves, a conventional monolithic radio telescope would require an aperture many kilometers in diameter to achieve the angular resolution of an optical telescope. Although the construction of such a telescope would be prohibitively expensive, the development of the digital computer made it possible to achieve this resolution with "aperture synthesis" radio telescopes. These telescopes use many relatively small antennas distributed over a large area to measure the coherence of the wavefront. An image is then formed in a digital computer.

The most powerful aperture synthesis array is the Very Large Array (VLA) in New Mexico. In a typical observation, the VLA measures the correlation between all pairs of the 27 antennas, producing 200,000 bits of information every 10 seconds. These data are then Fourier transformed into images with up to 4,000 pixels on a side. The resulting images are degraded because the wavefront is only sampled at discrete points, and because the earth's atmosphere distorts the wavefront before it is measured. Recently developed imaging algorithms have made it possible to correct for both these defects by using a priori constraints (e.g., possitivity and finite support) in the image plane to estimate missing information and atmospheric-induced errors in the aperture plane. The computations required can take many days unless supercomputer capacity is available. Figure 2 illustrates the results of such processing on the powerful radio galaxy Cygnus A. The final processed image reveals the tenuous thread of matter that connects the galaxy in the center to the lobes. This image unequivocally demonstrates the role that the central galaxy plays in fueling the radio lobes and, when combined with the extraordinary energy flux flowing from the center, leads to the inference that highly condensed forms of matter, such as a black hole, exist in the center of this galaxy.

Problems in observational astronomy need the computer capacity of a medium-size supercomputer, but they need this to be matched by very good I/O bandwidth, very large amounts of data storage, a convenient means to get large external observational databases into the supercomputer and to get the images produced out. In order to effectively steer the numerically intensive computations it is essential to have high bandwidth image display and interactive control of the supercomputer.

Similar imaging problems occur in many other areas such as medical imaging, crystallography, microscopy, and synthetic aperture radar.

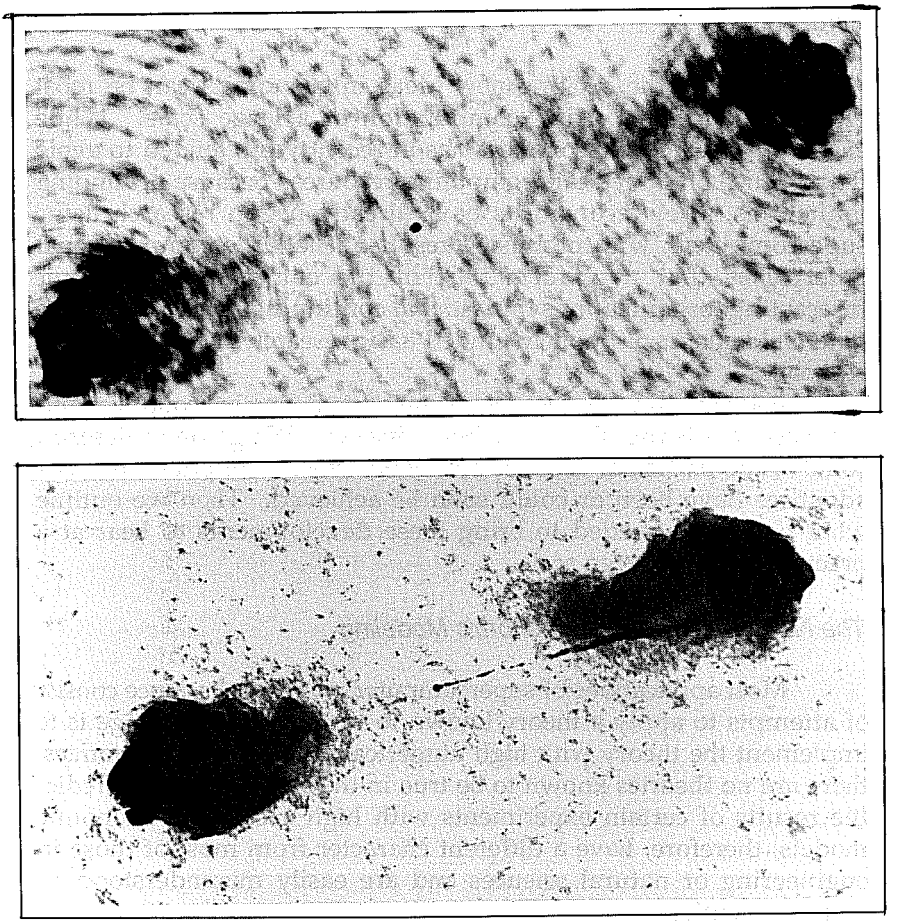


Fig. 2. The powerful radio galaxy Cygnus A: (top) before any computer image enhancement, (bottom) after (1) use of the deconvolution, or "cleaning" algorithm to correct for incomplete measurements and (2) removal of atmospheric "seeing" errors ("self-calibration"). Supercomputers greatly speed up the process. Courtesy of the National Radio Astronomy Observatory, operated by Associated Universities, Inc., under contract with the National Science Foundation. *Observers:* R. A. Perley, J. W. Dreher, and J. J. Cowan.

## Economics

### *Introduction*

Many of the exciting new ideas in economics can be most effectively pursued using computation at a large scale. While academic research in economics has in recent years tended to avoid dependence on large-scale computing resources because of funding limitations, a few economists have already begun research using supercomputers, with promising results. The large computational enterprises that now form part of the commercial economic forecasting organizations provide a continuing example of the potential benefits of publicly funded basic research in computational economics.

Below we first discuss the nature of modeling in economics and the reasons it is useful, despite being inexact. We go on to describe some of the particular areas of economics where computation-based advances have been or could soon be achieved. Then we outline what would be needed to bring these developments to bear at a practically useful large scale.

### *The Nature and Uses of Economic Modeling*

Most applications of supercomputers to natural science consist of attempts to apply a theory known to be true; the challenge is to implement the theory with high numerical accuracy. In economics, there are no theories known to be true in the sense that they predict the results of certain experiments with high accuracy. Economic models, therefore, have a different character from most of those in engineering or natural sciences and are easily misunderstood by natural scientists.

There are some areas of natural science which, like economics, offer little opportunity for experimentation and yet must deliver expert opinion relevant to important decisions. Nuclear reactor safety studies, epidemiological analysis of the effects of weakly toxic pollutants, ecological studies of the effects of acid rain, and analysis of the nuclear winter hypothesis are examples of areas in which modeling takes on some of the character of economic modeling. Experiment cannot resolve conflicts among theoretical ideas before decisions must be reached, so models are either based on controversial theory or include alternative, conflicting theories. In these areas, as in economics, complex computer models are valuable aids to making coherent use of large amounts of data and various theoretical ideas; but inevitably the results from the models are not

in themselves decisive. In these areas models help us understand the implications of existing data and theoretical hypotheses, but they cannot by themselves tell us which of conflicting hypotheses are correct.

Computer models are used pervasively to guide economic decision making. Commercial organizations sell forecasts and customized modeling services to private companies and to governments. Many models are maintained by government agencies, in the United States, in other countries, and in international organizations. For example, the Congressional Budget Office maintains its own model and uses outside models as well in making the projections that guide the Congressional budget process; the Federal Reserve Board in Washington maintains several models, as do most of the regional Federal Reserve Banks, to assist in the formation of monetary policy; models were used to project the economic impact of various legislative proposals during the recent tax reform debate; and computer models are used widely for financial planning and arbitrage in the private sector. All these models are imperfect, and there is no reasonable prospect that they will become exact. Because they are used in making important decisions, however, there may be great value in improving them.

### *New Developments*

*Stochastic Dynamic Optimization.* Much of economic theory has always consisted of tracing out the implications for the economy as a whole of the interactions among a population of optimizing individuals. Only within the last 10 years or so, however, has it begun to seem possible to produce practically useful models in which individual behavior is modeled taking account of both uncertainty and the passage of time. Such models have been labeled rational expectations models. They have already changed economists' thinking about the consequences of persistent inflation. They are promising in application to a wide range of other problems, for example, in understanding how families' decisions about the timing and spacing of childbirths interact with wages and job opportunities for women. But stochastic dynamic programming problems are well known to be computationally demanding. In an economic model where several different types of people are each taken to be solving such a problem, and the solutions are mutually interdependent, the computational burden can quickly become enormous. Economists are nevertheless pursuing such models and indeed have in the last two

or three years come up with ideas, based on the special character of the economics application, for cutting the computational burden.

*Computable General Equilibrium.* Models in which many different types of optimizing individuals are supposed to interact in markets are difficult to solve, yet have some special mathematical structure. Techniques have improved in recent years for handling such structure so that models with richly diverse types of individuals can be solved. A special impetus behind such work is the need to evaluate the effects of tax law changes and the many distinct types that emerge when people are categorized by age, race, sex, employment status, and a few brackets of each of several types of income.

*Nonparametric and Semiparametric Econometrics.* When economists try to interpret data, it has been common practice to reduce general mathematical theories containing functions of unknown form to special, convenient forms containing only a few unknown parameters. The statistical problems of taking account of the arbitrariness of the special forms used for estimation have seemed insurmountable. Reduced costs of computing have made it appear possible to use more realistic procedures, however, and there are now applications in economics of, for example, semiparametric regression (which allows an essentially unrestricted form for the distribution of disturbance terms in a regression) and nonparametric regression (which allows an essentially unrestricted form for the regression function itself). These procedures allow use of more realistic and easily interpreted empirical models.

*Bayesian Model Integration.* When there are several competing models with different implications among which the data do not cleanly discriminate, a decision maker using the model faces a difficult problem in judiciously taking account of all their results. Bayesian methods provide a formal analytic solution to the problem, but implementing the solution has seemed impractical. Some recent econometric work has applied Monte Carlo integration to achieve a computation-intensive implementation of the Bayesian solution.

*Bayesian Multivariate Time Series Modeling.* Multivariate macroeconomic models based on the econometric methods of the 1960s, like those most widely used in forecasting and policy analysis today, do not produce reliable forecasts on their own. Results from them are adjusted by ad hoc procedures to eliminate aspects of the forecasts that seem implausible. This is essential, because though the models are probabilistic, their internal structure does not take account of the greatest source of uncertainty about their forecasts -- uncertainty about whether the form of the model is itself correct.

The forms of the models are in fact changed every year or two, in most cases. Models which contain enough free parameters to realistically claim that most "unknowns" have been included as parameters turn out to exhaust the degrees of freedom in existing data.

A Bayesian approach can allow for something closer to the true degree of uncertainty about model structure, yet still make use of data with existing sample sizes. The size of models which can be handled with this approach has been increasing in the last few years, and they have a forecasting record comparable to that of economists who use standards models with ad hoc adjustment.

### *The Scale of Computation in Economics*

Though the list of topics in the preceding section is not exhaustive, it shows the existence of some gaps and frontiers along which progress depends on the availability of large-scale computing resources. Computable general equilibrium models ought to take account of the sophisticated dynamic behavioral modeling going on in other fields. Business cycle modelers who have concentrated on careful treatment of dynamic stochastic optimization should allow a more realistic level of diversity among types of agents in their models. Computable general equilibrium models use results from the kind of microeconometric studies to which nonparametric and semiparametric methods are now being applied, but they account for the uncertainty in those results crudely or not at all. Bayesian multivariate time series modeling has mainly stuck to conditionally linear and Gaussian structures, while the work on nonparametric and semiparametric methods has largely ignored serial dependence. The reason these fields stay mutually isolated is that each deals with computational challenges which are in themselves limiting, assuming existing bounds on computational resources in economics.

There is some large-scale computation going on now in economics. For example, Bayesian multivariate time series modeling has used NSF supercomputer resources, as have some studies of models based on stochastically dynamically optimizing agents. However, for economists ready to scale up from personal workstations even to heavy use of a standard mainframe computer, funding is scarce. So long as there remains a discrete jump from difficult-to-finance local computation to cheap but inconvenient computation on remote supercomputers, migration of academic economists to large-scale computation is likely to remain slow.

The largest scale computation going on in economics is probably that in the commercial forecasting organizations. At least one of them maintains a database of tens of thousand of economic time series which are regularly forecast with the aid of a 1500-equation nonlinear model. The convenient linkage of the model to the database is critical. Though the scale of the computation is large, the emphasis on database management and interactive connection to a large number of users is not. With existing supercomputer operating systems, it is not likely that this kind of computation will make much use of supercomputers soon.

The commercial models were for the most part originally based on publicly funded computational research in economics, but they have not maintained much contact with academic economic research. Nonetheless, the eventual connecting of such models, which have been adapted to meet the needs of actual decision makers to new research ideas, is an important goal. One can imagine a system in which data on financial variables that become available daily or hourly are continually incorporated into a worldwide model that decision makers could access interactively for forecasts of the likely effects of their policy choices. Such a system could, for example, use rational expectations theory and use Bayesian methods to take account of model specification uncertainty. It would be a major intellectual and computational project, but with appropriate resources it is not an unrealistic goal.

*Recommendations*

- In economics, as in other disciplines where large-scale computating is in its early stages, productive use of supercomputers requires that they be accessible and that the software migration path to them from smaller computers be transparent. Research support for software development on supercomputers is essential. So long as academic supercomputing remains concentrated at a few locations, network access must be improved far beyond its present state.

- Research funding for economics should recognize that some productive projects will involve large amounts of computation. These projects will require budget levels for equipment, computer time, research assistance, and data acquisition and management beyond what has become the standard for social science research projects in recent years.

**TABLE 1. Selected Recent Achievements of Supercomputing in Engineering**

| *Problem* | *Methods and Computational Resource* | *Results and Impact* |
|---|---|---|
| Aerodynamics airfoil calculations | Finite difference solution of transonic flow (requires hrs. of Cray 1 time) | Two orders-of-magnitude reduction in number of airfoils tested in wind tunnels |
| Design of artificial heart valve | Three-dimensional finite-difference simulation of flow/structure interaction | Design of long lasting heart valves with minimal mechanical degradation |
| Nuclear reactor simulation | Vectorized Monte Carlo of the engineer | "Turnaround time" decreased from 2 weeks to 1/2 day; resulting from order-of- magnitude speed up in computation |
| Laminar turbulent flows | Finite difference/ element methods | Effective heat and mass transfer analysis for the design and manufacturing of advanced mechanical devices |

**TABLE 2. Selected Ongoing Supercomputing Projects Ripe for Impact**

| *Problem* | *Methods* | *Computer Capacity* G flops/M words* | *Impact* |
|---|---|---|---|
| Aerodynamic simulation of full aircraft | Triangular solution of transonic Euler equation | 10/100 | Full aircraft system design |
| General purpose laminar/ turbulent flow simulator | Spectral element methods | 10/100 | "Black box" 3D tool for engineering design analysis |
| Simulator of complex materials processing systems | Finite element methods | 1/10 | Design and optimization for advanced materials systems such as chemical vapor deposition, composite materials, solidification processing, polymer processing |
| Direct simulation multiphase | Molecular and Stokesian dynamics | 1/10 | Direct of simulation of problems in Flow reactor safety microstructure and particula flows |
| Turbulence | Direct simulation/dynamic | 10/100 | Fundamental understanding of geometry of turbulence and transport properties |
| Semiconductor device simulation | PDE description particle transport | 10/100 | VLSI design |

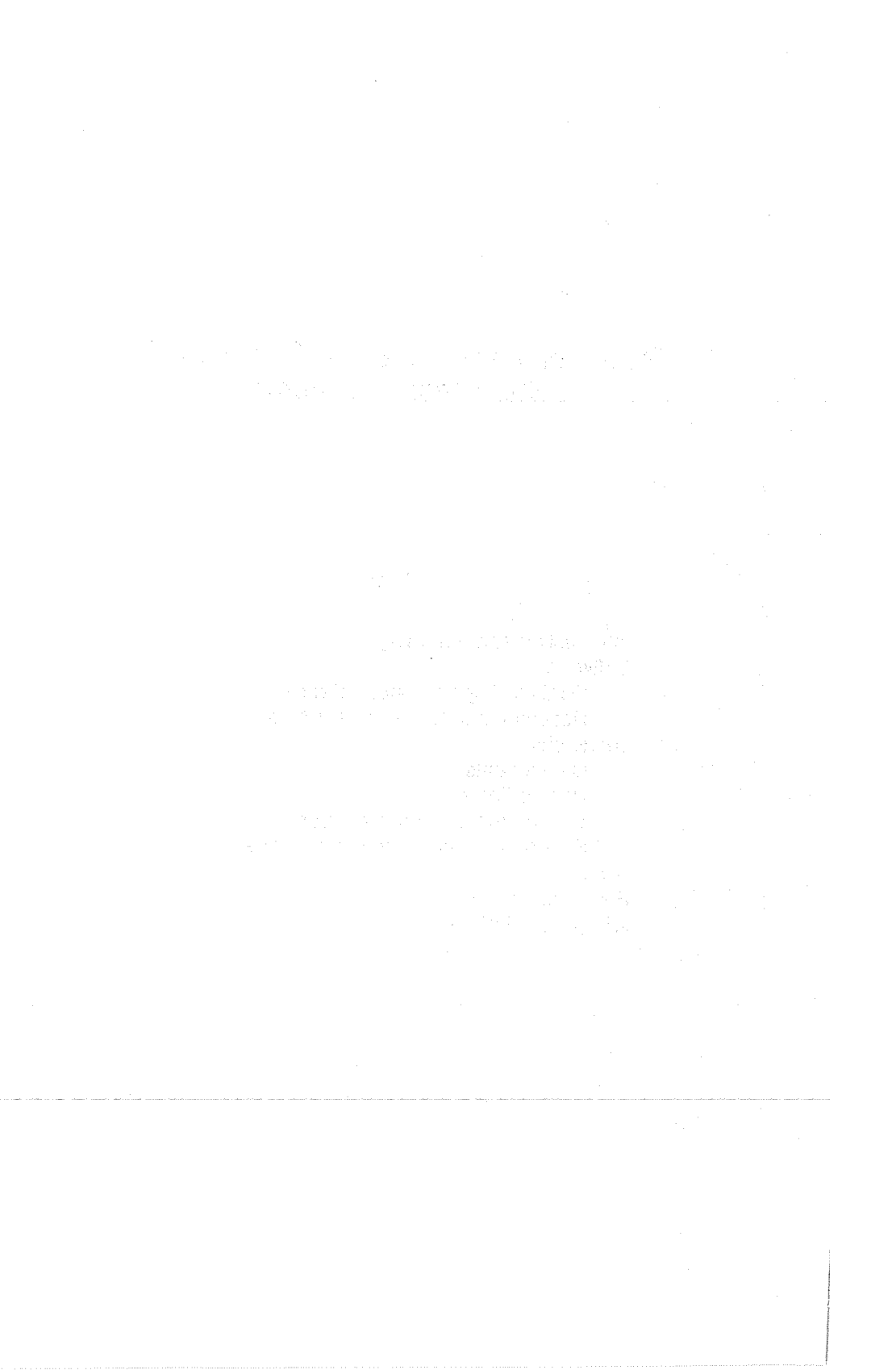
*For solution in about one hour of computer time.

**TABLE 3. Some Grand Challenges in Computational Engineering**

| *Problems* | *Impact* |
|---|---|
| Physics of systems with many length scales | Understanding and exploitation of turbulence, micro composites, and multiphase phenomena |
| Molecular engineering | Process design to build molecules that yield specific functional and bulk material properties |
| Numerical wind tunnel | Complete design of aircraft, including hypersonic vehicles |
| Vision and speech recognition | Design of interactive expert systems |

# Report of the Sub-Panel on Advanced Computational Systems

## *Outline*



## Introduction and Summary

The members of the Sub-Panel on Advanced Computational Systems identified four areas of continuing concern:

- Software
- Leadership
- Training
- Access.

We unanimously agreed that software is the most significant problem, both for the development of supercomputing systems and applications, and for maintaining leadership in the field of computing.

Most major speed improvements in the past have been the result of faster components. While faster components are still important to supercomputer development, it is becoming more difficult to achieve performance improvement solely by the use of faster components. Instead, designers are relying more on multiple processor systems to obtain higher performance. Yet software for multiple processor systems is significantly more difficult to produce and its development lags far behind that of hardware. Software is not only *the* key to the productivity of our engineers and scientists, but the enabling factor that makes advanced computing systems feasible.

We believe a national initiative is essential and appropriate to rapid progress in the supercomputer area. We also believe that in addition to software, the following problems are likely to further weaken our leadership position in computing, especially supercomputing:

(1) Lack of access to a continuing supply of high-quality, domestic semiconductor parts;

(2) Lack of support for a healthy supercomputer industry able to compete in the supercomputer marketplace;

(3) Lack of support for new computer architecture projects;

(4) Lack of better design tools.

We note that great progress has been made in the area of access to supercomputers through the NSF, DOE, and NASA programs. Nevertheless, there are weaknesses in this area as well.

Networking software, and other network facilities are weak and hard to use. Furthermore, mini-supercomputers should play a larger role in the national access program. And two segments of users -- industrial users and computer scientists and engineers -- need more consideration in the access program.

Finally, training and education are a continuing problem, especially the lack of appropriate courses and curricula in the areas of advanced computational methods and advanced computer design and engineering.

The following four sections summarize our consensus on the four issues we identified as major problems and opportunities: software, leadership, access, and training. A final section discusses the issue of national supercomputer initiatives.

## Software

The productivity of our scientists and engineers is a precious national resource that we cannot afford to waste. Access to National Supercomputer Centers has made it possible to solve problems that could not be attacked before, yet solutions on today's supercomputers continue to require an enormous effort in low-level programming. The main impediment to more productive use of supercomputers is the lack of adequate system software for these machines.

United States supercomputer vendors traditionally have not provided system software that is as sophisticated as the system hardware. This is true because (1) hardware can be sold without such advanced system software; and (2) vendors have assumed that the user shares the responsibility for providing this system software. These assumptions have not been inappropriate over the last 20 years in the United States because the user base was largely composed of sophisticated scientists and engineers such as those found in the major national laboratories and research institutions. However, the next round of advances in science and technology will be made by users from a much broader cross section of the technical communited -- experts in their own areas, but less sophisticated in the principles and techniques of advanced computation. For supercomputing systems to be employed in this manner, the system software must be as advanced as the hardware. Our foreign competitors have realized this and have already capitalized on the publications of leading American universities and research organizations. The United States is not well postured to make similar advances nor to make use of its own advanced research.

Supercomputer software must no longer be treated as an afterthought, or as a necessary evil in hardware development. The facts are

- Supercomputer software requires more investment in time, money, and quality engineering than the hardware.

- The state-of-the-art of parallel software lags far behind the other critical supercomputer technologies of architecture and components.

As a consequence, we recommend that supercomputer software research and development must become our highest priority.

### *Goals of Supercomputer Software Research*

The goals of supercomputer software research, though similar to those of conventional computing situations, assume a much greater importance because of the novel nature of the architecture and the massive needs of the user community. The key goals follow.

*Feasibility.* The need to take advantage of the complex architectures characteristic of present and proposed supercomputers leads to more complex programs. These programs are extremely difficult, even impossible, to write without a substantial amount of software support, especially sophisticated compilers, debugging systems, and algorithm libraries. Often the difference between whether a program will be undertaken or not will depend on the availability of this support software.

*Efficiency.* Compilers play a critical role in supercomputer performance when they translate the programmer's instructions to machine code, so they have a major effect on the proportion of the rated performance that is actually observed. We are just beginning to see the emergence of compilers that can do this efficiently for specific computers, but much more needs to be done to expend this work to broader ranges of architectures including the newer, highly parallel machines.

*Productivity.* Better programming environments are capable of substantially improving the amount of correct code a programmer can produce per day. They are essential in order to help manage the increased complexity of code required for supercomputers.

*Portability.* For supercomputers to attract users, they must have programs that work on other computers as well, including the next generation. New, high-level languages have the potential to

facilitate this portability, but better compilers also are required to efficiently translate these languages to a variety of architectures. In addition, while some excellent algorithm libraries exist, they are often not applicable to new architectures; we need to learn how to design algorithms that will be useful on a wider variety of architectures.

*Research and Development Areas*

To retain leadership in computing, we need a major initiative to build a new generation of software technology. This will require research and development in the following key areas.

- Fundamental principles for designing architecture-independent programming languages must be developed. Such languages will allow the programmer to specify a solution at a high level of abstraction while facilitating efficient translation of the resulting program to a variety of supercomputer architectures.

- Powerful new compilation techniques will be needed to extract more parallelism from programs and to map these programs to complex architectures. Current techniques need to be extended to advanced languages and a more varied class of architectures. Methods for optimizing programs as a whole, while maintaining the convenience of separate compilation, must be developed. Because future supercomputers will have more complex memory hierarchies, compiler management of these hierarchies will be critical to achieve acceptable efficiency, to allow easier portability between systems, and to extend the lifetime of applications software.

- To achieve software portability, methods for generating compilation and debugging systems from descriptions of machine architectures should be studied. Without such systems, a prohibitive effort is required to implement languages efficiently on emerging parallel architectures. This research may in turn influence the design of machines.

- We must continue the high level of effort on design of efficient parallel algorithms for challenging problems. A major goal of the software effort should be to make retargetable implementations of such algorithms easy to produce.

- New research in programming environments and distributed systems will be needed to build the software-rich, multiple-machine programming support systems that are prerequisites for high programmer productivity and efficient compilation and checkout. Special attention must be given to developing parallel debugging systems that can be used and understood by the average scientific programmer. This will entail inventing new approaches to deal with irreproducible errors and similar problems characteristic of parallel architectures. In addition, tools are needed to identify performance bottlenecks in applications programs.

- We need a fresh approach to the design of operating systems. In the past, operating systems research has concentrated on multiprogrammed interactive systems. New parallel computers will need operating systems that allow the algorithms embedded in the user's program to schedule and synchronize parallel processors. Such systems could be based on industry standard user interfaces and file systems networked with standard workstations.

The programming tools that will result from this research are as essential to helping scientists do computational research as good laboratory equipment is in doing experimental research. In particular, these tools enable the researchers to focus their energy on that part of algorithm design that cannot be automated, namely, the conceptualization needed to understand where physical processes can be partitioned into naturally parallel components.

## Leadership

We believe that this country's leadership in supercomputing continues to weaken. Software will begin to play an increasingly important role in correcting this weakening position, yet other issues are still of great concern. It should be no surprise that we identified the availability of components as one of these issues. We also identified the lack of good design tools, the need to foster new architectural ideas, and the need to support a supercomputer industry.

### *Components*

The United States is losing its competitive edge in the design and manufacture of semiconductor components that will be required

for future generations of supercomputers. Reliable and cost-effective semiconductor devices demand high volume manufacturing philosophies. High-performance supercomputer components occupy a relatively small niche in the overall semiconductor market. This niche market, coupled with short-term goals for profitability, has caused domestic suppliers to place less emphasis on these required devices. If this trend continues, supercomputer manufacturers will be forced to become increasingly dependent upon Japanese sources. Some Japanese system manufacturers, by contrast, seem to take a longer-term, vertically integrated approach. This is consistent with their agenda of seeking worldwide supercomputer dominance.

Government-supported semiconductor initiatives, such as VHSIC, while meeting certain Pentagon-specific goals, have been less effective in providing technology for commercial supercomputer applications. Usually the contractors involved have been large defense system suppliers who have neither the necessary manufacturing experience nor the commitment to provide the necessary components to supercomputer vendors. This is due to the fact that fulfillment of the contracts often requires only low volume production of a few key semiconductor components. Another example, the DARPA gallium arsenide initiative, has (appropriately for DARPA's mission) emphasized low-power, radiation-hardened devices at the expense of very fast circuits that are needed for commercial exploitation in supercomputer systems.

Key members of the United States semiconductor industry must guarantee timely access to the necessary volume of state-of-the-art semiconductor *logic* devices exhibiting the highest possible performance at any given time. The definition of the necessary technology parameters is best provided by supercomputer companies. These same U.S. semiconductor industry members must also provide the most cost-effective *storage* devices in sufficient quantities when required in the development of the next generation of supercomputers.

These two technologies, logic and storage, must continue to be an integral segment of key U.S. semiconductor supplier product portfolios in order to allow U.S. supercomputer manufacturers to regain their former leadership position. Any government initiative whose goal is to bolster the semiconductor industry must give every consideration to support both high-performance storage *and logic*, since both are crucial to the successful development and manufacturing of supercomputers.

### *Design Tools*

There is need for research leading to better tools for the logic design of supercomputers. Although this research is important for logic design generally, supercomputer logic design is more difficult than any other sort for two reasons. First, the size and complexity of supercomputers, coupled with the limit on the number of designers that can coordinate successfully on a single design of this complexity, lead to extended logic design times, long product introduction cycles, and high risk. Second, the need for high speed from the logic circuitry leads to greater interaction between logic design and correct timing. The result is design by trial and error. The more customary "design rule"-based approach to this problem is too wasteful of space and time for critical supercomputer applications. This tendency is exacerbated by the newness and incomplete electrical characterization of the devices needed to obtain very high speeds.

### *Support for New Architectures*

Architectural research initiatives are crucial to leadership in supercomputing and should be funded at much higher levels than are presently available. Further experimentation with advanced computer architectures is essential because many important problems remain unsolved in the design of truly useful supercomputer architectures; this experimentation tends to be very expensive. At least some of this funding should encourage joint efforts between public and private sectors. The potential of such creative funding (perhaps similar to the ESPRIT program in Europe) is to foster better understanding of architectural concepts and the development of mature systems, beyond the basic research project, providing a mechanism for the direct transfer of technology into the commercial domain.

Researchers in universities and industry should be encouraged to develop innovative designs, to simulate their designs for feasibility and utility, and to follow through on construction and demonstration of the concepts that show a reasonable chance of success. Because the financial investment in the construction of a supercomputer design is extremely high, the greatest leverage of government funding will come from the support of ideas that are plausible, but too risky for industry alone. That is, industry developers must be concerned with shorter-term returns on investments, and so are unlikely to be able to commit significant resources to concepts that diverge from proven designs. Although

the potential to fail is an important ingredient of the research process, and much can be learned from a carefully executed failure, industrial investors are not likely to view architectural failure as a reasonable return on investment.

For the architectures that evolve from paper designs to full implementations, careful evaluations of performance must be conducted. These evaluations should be based on a range of nontrivial problems, and should wherever possible be conducted by researchers who are independent of the design process. Performance and programmability questions should be addressed in the evaluation process, with the ultimate goal of a new system being to increase the productivity of the end users.

Mission-oriented funding does not usually produce designs that will result in commercial supercomputers usable by the general scientific community. Because the requirements for producing a commercially viable, general-purpose supercomputer are so demanding, it is unlikely that the more narrowly focused effort usually appropriate for mission-oriented programs will produce a design that will have the vision to emerge as a full-scale, general-purpose supercomputer. Such vision would require not only a mature architectural design, but also an advanced technology base for components, the ability to manage hierarchically or otherwise the large memory required of a supercomputer system, and the capability of attaching and utilizing advanced peripheral equipment -- especially high-performance I/O. Issues of scalability and extensibility are difficult to manage late in the life cycle of an architectural implementation and should be part of the original plan. These considerations have not been at the forefront of most mission-oriented funding, hence the designs produced under those programs are unlikely to result in commercial supercomputers usable by the general scientific community. A policy for fostering the design and development of more general-purpose, broadly based supercomputers, through the combined efforts of scientists and engineers in the public and private sectors, should be adopted.

### *Support for New Supercomputer Ventures*

The desirability of the long-term existence of a U.S. supercomputing industry that is in a position of world leadership is unquestioned. Cray and ETA, which are relatively small compared to their foreign competition, are constrained somewhat in their innovation by compatibility with their own installed base. From the U.S. national perspective, a healthy, stable industry requires greater economic depth, greater technical freedom, and longer-term

stability. Current mini-supercomputer ventures provide learning vehicles for parallelism, but none are likely to provide a base for future supercomputer companies because of their relatively small size and their concentration on shorter-term, market-related goals. One of the reasons is simply the amount of funding required for a supercomputer ($100 million or more), but perhaps more significantly, they have discovered that they have already found a far simpler way to make money. Thus, maintaining long-term leadership will require major changes in U.S. federal policy and procedures to encourage greater activity in the supercomputer industry.

During and following World War II, the U.S. government played a major role in bringing about a dramatic U.S. lead in computing, principally by (1) direct funding of industrial efforts, and (2) procurement procedures that encouraged U.S. industrial innovation. Examples of direct funding are LARC, Stretch, and Illiac IV. Less direct support was demonstrated by the government when Los Alamos accepted a no-cost loan of the first Cray 1 without software. The government was motivated by keeping an impartial stance for all; however, such behavior is now illegal and the symbiotic relationships of the past are no longer possible. In fact, the environment has become positively antagonistic.

When ETA was formed in 1983, there was no possibility of the favorable treatment accorded Cray a decade earlier. Supercomputer development requires both a very large capital investment and a long-term horizon, facts not in consonance with the venture-capital-based approach of new U.S. startups. This, coupled with the environment, has led to no successful new supercomputer startups other than those directly traceable to U.S. government policies of World War II vintage. Under the current "fly before buy" philosophy, the U.S. government is no longer shouldering any of the burden associated with a healthy supercomputer industry.

Two existing programs provide a model for a federal solution: (1) BIRD (the Bi-national Research and Development program between the United States and Israel); and (2) the DOD IR&D program. BIRD provides loans at reasonable rates for joint U.S.-Israeli business ventures. The loans are repaid if the venture is successful and are essentially forgiven if the venture is unsuccessful. The National Bureau of Standards serves as an impartial technical arbiter in the United States. United States entities that are part of the DOD IR&D program can expect up to 80% research and development costs to be reimbursed by DOD. The efforts must be chosen from an approved list (but not preapproved), and audited for quality every 3 years. A U.S. civilian R&D program for the supercomputer industry modeled on these efforts would be appropriate.

## Access

Of the four original recommendations of the Lax committee, access to supercomputers is where the most significant progress has been made. The establishment of national centers has been a major step forward in maintaining this country's leadership in supercomputing. Yet we should not become complacent about this issue -- there is still much room for both enhancement and expansion of this program. We note, for example, that supercomputer centers in foreign countries still outnumber similar centers in the United States.

- Network technology needs significant improvement both in performance and software support. The thrust should be toward a fully integrated, distributed facility rather than just ad hoc remote logons.

- Software, as discussed elsewhere, is another issue of importance to quality of access, including the issue of portability.

- Use of mini-supercomputers should be encouraged as a path for the development of prototype software for supercomputers. In addition, these systems can be used as pre- and post-processing facilities by interdisciplinary teams at remote sites.

- Supercomputers must be accessible to computer scientists involved in language, algorithm, operating system, and performance evaluation research interests. The requirements of this type of user are somewhat different from those of researchers in applications areas; in some cases they involve sole use of the entire system for the duration of an experiment, and it may take many hours of downtime to install an experimental operating system for a very short test. This is in conflict with the management of a multi-user production system, but it is critical to the advancement of research in supercomputer systems issues.

- Industrial access to supercomputers is available almost exclusively through affiliates programs that require million-dollar investments from the affiliates. Researchers in the private sector should be able to purchase time at the centers, for work approved by the allocation committees, in blocks significantly smaller than 1,000 hours. In addition,

the entire scientific and engineering community -- including the portion in industrial research organizations -- should have access to the national computer networks.

- Directors of our national supercomputer centers, from all agencies, should meet annually to foster planning and coordination of services.

## Education and Training

In 1982, the Lax Report (see Executive Summary) identified training as one of the major areas of concern:

Another important component of this national program is the development of an imaginative and skilled user community in supercomputing. There is a considerable shortage of appropriately trained personnel and of training opportunities in this area. Forms of institutional encouragement, such as NASA's special fellowships in the area of numerical fluid mechanics, special summer schools, and special allocation of access time to supercomputers for those projects that involve graduate students, should be considered. Some of the more mathematical aspects of these activities can be accomplished independently of the machines on which the actual calculations are done; however, the true integration of methods and their implementation cannot be done without access to supercomputers. The nature of the machine architecture has a very profound effect on the numerical methods, on the algorithms, and of course on the software. Thus, while being trained, students must have access to state-of-the-art computers. Today such training is virtually nonexistent; yet the skills gained from such training are essential to science and engineering.

We reiterate this concern. In many areas of science and engineering, advanced computational methods are becoming as important as experimental and analytical methods, yet few curricula include even basic courses in computational mathematics. While universities are beginning to gain access to supercomputers through federally sponsored programs, they still need to be encouraged to develop courses and curricula in this area.

In addition, we need the development of curricula in areas important to the design of new computer systems, including supercomputers. These curricula need to include aspects of hardware design, compilers, operating systems, and algorithms, as well as advanced computational methods.

## A National Initiative

One of the difficulties of promoting the state-of-the-art in advanced computation is the wide diversity of disciplines needed. Too often research becomes uncoordinated because of a lack of focus on the critical issues and because of a lack of understanding of the complex interrelationships between machine architectures, operating systems, languages, programming paradigms, algorithms, and applications. We believe a major initiative to produce a new generation of supercomputers and software would provide the necessary focus and would stimulate the necessary research and development in the areas of concern to supercomputing (see also [11] in Executive Summary). We stress that software should be the major thrust of this initiative -- hardware should not proceed until software problems are better understood. Additionally, development of components, architectures, and design tools is important to the overall goals of this initiative and must be included.

A national initiative would focus U.S. efforts on the development of the technology for building a new generation of supercomputer orders of magnitude faster than existing machines. The resulting machines could be applied to problems of national interest, with considerable fallout to the private sector. Software should be the major thrust of this initiative, but the development of components, architectures, and algorithms should also be major components.

A national supercomputer initiative must have the following fundamental components:

- It must expand university and laboratory research into the fundamental problems described in this report. In embarking on this research, it will be necessary to find projects of sufficient size and duration to build experimental software systems of practical usefulness.

- We believe that the hardware for a teraflop computer could be built today. But the result would be unusable except by a few very sophisticated users, and it would not form a good foundation for the future of the industry in this country. Instead, what is needed is an initiative to develop the software and technology for teraflop computer systems. Building such a system is an undertaking of complexity far greater than that of simply building prototype hardware and should be funded accordingly. Exact estimates of the funding

requirements require further study, but we believe that hundreds of millions of dollars would be needed.

- Another important aspect of this initiative is diversity. The results of this initiative are too important to this nation to risk on any single architectural concept, research institution, or agency. Multiple, major projects will need to be funded to provide the necessary diversity. In addition, continued funding of single-investigator grants will be required to ensure a continual flow of fresh ideas into the field.

- Mechanisms for effecting a smooth transfer of technology from university and research laboratories into commercial supercomputer companies need to be developed. Two approaches appear promising for this effort. First, increased funding for research, as proposed above, should stimulate the production of computer science and mathematics Ph.D.s in supercomputer software. These Ph.D.s may then be hired by industry, bringing research ideas into commercial practice. Second, a network of institutes to foster technology transfer might be developed. The role of these institutes would be to develop promising software systems to near product quality for transmission to commercial firms.

# Report of the Sub-Panel on Parallel Computing

## *Outline*

Introduction
- Parallel Computing
- Stability
- Balance
- New Initiatives

A National Network
Experimental Machines
Neural Networks and Cellular Automata
Advanced Architecture Research
Theoretical Underpinnings
Conclusions

## Introduction

High-performance computing is vital to modern science, engineering, and the management of the U.S. economy and defense. It is generally agreed that in the future, further improvements in high-performance computing will be achieved mainly through the exploitation of parallelism. Many problems must be solved to accomplish this goal. The question addressed here is what sort of infrastructure, program management, and research will be needed to realize high-performance, large-scale computing through parallelism.

### *Parallel Computing*

The next generation of high-performance computing will be achieved by utilizing parallelism at all possible levels. At the highest level, National Supercomputer Centers will not only perform calculations, but will remotely access massive data banks, send special jobs to experimental machines in outside laboratories, and transmit results to innovation support centers (e.g., MOSIS). The National Supercomputer Centers themselves will be composed of several, perhaps specialized, supercomputers capable of cooperating on a single calculation. Within a given computer (e.g., the Cray X-MP), multiple parallel processors, each containing multiple functional units, will operate in parallel. At the bottom level, logic circuits also will exploit extensive parallelism.

To understand how to design, manage, program, and effectively use such systems is a tremendous challenge. It is essential that the entire science and engineering community cooperate with the computer science community in order to properly direct the development of these next generation systems.

*Recommendation*: Establish a series of broad national computing initiatives to realize the potential of parallel execution. These initiatives, detailed below, address issues of management, access, infrastructure, development, research, and education. The primary goal is to create more powerful supercomputer centers by employing parallelism, and to interconnect these and other major computing centers into a single, coherent computing system.

### *Stability*

A stable environment is especially essential to the conduct of research in high-performance systems. This includes both stable funding sources and a stable infrastructure. Funding for computer

science research has been continuously increasing over the last few years. Large national programs such as DARPA's Strategic Computing Program, the Strategic Defense Initiative (SDI), and NSF's Engineering Research Centers have all contributed to this increase. Interdisciplinary research teams, particularly at universities, have been assembled to respond to these initiatives. These teams frequently take years to assemble and to develop their expertise; by the nature of high-performance computing they themselves must be quite large. Funding gaps, policy changes, or program terminations can have a devastating effect on these research teams, not only because they result in the loss of key personnel, but also because they create a climate in which it becomes difficult to establish future teams.

During the past 2 years, several research teams have experienced funding disruptions. The problem could become much more serious in the future because larger, more focused programs (such as SDI) could cause even greater disruption in the research community. While some initiatives are primarily focused on computing technology, such as the DARPA Strategic Computing Program, other initiatives may have several enabling technologies, and computing may not be the critical one at the moment. It is all too easy to first institute and then reduce support for these teams. It is important to preserve research in the basic computing technologies that are funded by a number of agencies and through several different programs.

The United States does not have a long-range policy that guides research in computing. Yet every year the U.S. economy is becoming more and more information and computation driven. In the future, it may be wise to establish a high-level organization that would be directly responsible for computing and for other information sciences. Would it be presumptuous to suggest a "Department of Information" at the cabinet level? Or at least an "Advanced Computer Research Agency"?

*Recommendation*: FCCSET should provide increased coordination across programs and agencies, and particular attention should be paid to the impact that funding perturbations can have on our research base. FCCSET should also consider the merits of establishing a new, more formal, organization for computing and information technology.

### *Balance*

There is concern in the community that misdirection of new computing initiatives could significantly set back progress in the

exploitation of parallel computing. For example, a large, very expensive, highly visible, but ill-conceived project to demonstrate massive parallel computation could easily back-fire and thus set back the field many years. A carefully balanced program to develop the full range of capabilities is necessary to assure that the pursuit of massive parallelism will be successful.

*Recommendation*: A balanced program for introducing the next generation of high-performance parallel computing systems should be established. The emphasis should be on increasing understanding and capability, rather than on "demonstrating" some particular performance (such as a Teraflop) that can be achieved for some restricted problem. This will require that FCCSET develop a carefully balanced and coordinated research, development, and implementation plan to cover theoretical and experimental research, development of techniques, algorithms, and tools, as well as the carefully considered acquisition of equipment.

### *New Initiatives*

Making the transition from today's generation of serial machines to parallel architectures requires several levels of action. We recommend:

(1) the establishment of a new national computer network;

(2) increased access to experimental parallel machines;

(3) the exploration of neural network and cellular automata ideas on the currently available machines;

(4) an aggressive development of new parallel computer architectures;

(5) increased support for research of the theory of parallel computation.

These proposed initiatives are discussed in detail in the sections that follow.

## A National Network

A new national network offers the United States a payoff with high leverage in maintaining its technology lead. There are several reasons why an initiative is essential now.

First, using an analogy proposed by Senator Gore, a nationwide network will be the superhighway system of the information age. The critical computing problems of the future will rely increasingly on computations distributed across the nation. We see the information systems problems as being "communication-bound" rather than computation-bound. For instance, distribution of satellite data to scientists is hindered by low-speed, lossy lines. Projects such as the national knowledge bank (advocated by Kahn of NRI) are limited by inadequate networking. Even now, access to supercomputers is limited by poor networking. The supernetwork proposed here would not only make access possible, it would fundamentally change the way we think about the availability of distant information and processing resources. For example, a doctor treating a critically ill patient from out of state would be able to call up medical image data in a few seconds from a remote site rather than foregoing that information because of time constraints. Similarly, scientists feasibly could explore the use of multiple supercomputers to carry out a single computation in parallel, effectively forming a super multicomputer from geographically dispersed machines. A simulation of a proposed monetary policy, for example, might employ resources of a national data bank of economic time series, a large supercomputer complex, and an advanced graphics center.

A supernetwork could catalyze work in the commercial and military areas. It could also open up a market for nationwide (value-added) information and computation services that are currently infeasible. Moreover, it could facilitate the transition of interorganization interaction to electronic modes, including mail, conferencing, purchase orders, and so on, with the attendant benefits in efficiency.

We see this type of major effort essential to counter the current international movement to freeze communication standards into relatively restrictive, antiquated, and technically inferior protocols, the International Standards Organization's Open Systems Interconnection (OSI) effort being the prime example. The parameters of the supernetwork would allow the United States to leapfrog the political barriers raised by this standards effort, since OSI protocols appear unworkable in this performance range.

A second timely aspect of this effort is the maturation of research in the area of distributed operating systems and the perspective gained from our previous work on networks from the 1960s and 1970s, particularly the ARPA network and DOD Internet. We now have the basic techniques needed to exploit a supernetwork and to stretch the practice of distributed computing in the local

network out to a national scale. In particular, work on distributed databases and file systems could be applied nationwide. Similarly, work on cluster-based parallel processing, successfully applied on local networks, could then be applied to nationwide clusters. An extreme example here again is a cluster of supercomputers. Such a technology also would allow high-performance computation to ride advances in supercomputers, limited only by the speed of the individual supercomputer multiplied by the number of such computers available.

Another timely aspect of this initiative is the recent maturation of fiber optics technology. In fact, the project might be piggybacked to some degree on the current installation of fiber by phone companies nationwide. However, the provision of the fiber channels marks but a starting point for the research work required to realize this supernetwork.

The supernetwork initiative also would be a tremendous technology "pusher" in the following key areas of communications and computer science:

(1) Switching technology: how to build very fast, low delay, nonlossy packet switches. Optical switching and optical computing devices may be applicable.

(2) Network interface architectures: how to interface computers to this range of bandwidth without significant delay or overloading the host processor. Architecture, VLSI, and protocol designers will have to work closely together. Note: this work should also contribute to the general problem of supercomputer peripheral design.

(3) Protocol design: current transport, presentation, and application level protocols are not suitable for supernetwork performance parameters. A new generation protocol architecture is required, addressing performance, reliability, and security concerns, and this requires a major advance in research as the foundation.

(4) Network management: managing a system on this scale, and of such complexity, with these performance, reliability, and security requirements is a difficult task. Again, research building on our experience with the Internet is required. High-performance gateways will also figure prominently.

This list is a mere sampling of the key computer communications research that is required. In addition, computer research is being stimulated in several other areas, including: (1) high-performance distributed computing, (2) cluster-based parallel computation, and (3) multi-person distributed cooperation.

Finally, a closer cooperation between researchers would be possible with a supernetwork, thus changing the nature of collaboration and interaction in research. A supernetwork clearly would make us a "national village," to use Marshall McLuhan's vision.

We envision a program roughly as follows. A preliminary study would identify key critical areas of research required before development began. It would also identify budgeting requirements and current opportunities to acquire basic facilities. For instance, telephone companies might be willing to install extra fiber now in anticipation of supernetwork requirements, provided that these requirements are identified and incentives are supplied early on.

The preliminary study would lead to the funding of research in the critical and peripheral areas, followed by further steps in the direction of installation, as appropriate. Planning would commence to identify possible contractors for deployment. In addition, a substantial effort to educate the research, commercial, and military communities to the opportunities of the supernetwork would be useful. This would be a tremendous stimulus to these communities to revise their research and development agendas in line with the new opportunities. We would expect a plethora of auxiliary research projects growing out of the revolutionized environment provided by the supernetwork.

Valuable experience can be gained from the development, deployment, and maintenance of the ARPANET in this undertaking, both in successes and failures. The ARPANET also provides a precedent for this endeavor. The cost of this proposed research would be significant. However, this investment would be paid back many times over through the stimulation of the United States' industrial enterprise in all aspects of information technology as it fights to maintain the lead in both the commercial and military areas.

*Recommendation*: A major national initiative for research and development leading to the installation of a nationwide network with several orders of magnitude improvement over current networking technology should be mounted. We refer to this as a "supernetwork" since its relation to current networks would be analogous to the relation between supercomputers and other computers. We propose the following design goals: (1) 1 gigabit/sec.

user-level transfer speed; (2) 30 millisecond propagation delay nationwide; (3) $10^{-10}$ error rate, including packet loss.

These network parameters would exceed those of the most ambitious local network in existence, limited only by the inevitable propagation delay (which is determined by the speed of light).

## Experimental Machines

The future of scientific computing is clearly aimed in the direction of parallel processing. While we are exploiting state-of-the-art supercomputing today, we must prepare for the arrival of the next generation of high-performance processors in the near future. To be in a position to use such systems we must develop techniques, algorithms, and tools on the existing experimental (prototype) parallel architectures. While these machines lack the ability to solve our most demanding problems, they provide an important vehicle for investigation. The goal is to explore how to perform efficient parallel computations rather than demonstrate high performance through the use of fast semiconductor technology.

In addition to developing a general framework for highly parallel, general purpose computing, there is a need to investigate special purpose hardware for certain applications. By exploiting such hardware, a cost-effective solution for those applications can be obtained. The most important issue is to understand how to do this well over as wide a variety of different problems as possible.

Parallel machines that have heterogeneous structure will be well matched to solve certain applications. Heterogeneous machines could also be the best way to match special purpose, high-performance processors to generic computing problems. Care should be taken not to assume that massive parallelism will be most effective when each processor is identical.

Each vendor of parallel computers has provided its own set of parallel extensions. This has led to widely different approaches ranging from explicit synchronization to automatic detection of loops. A standard set of tools and extensions to our current programming languages are needed that will assist in the portability of programs across a wide range of parallel architectures, thus avoiding locking the software into a particular architecture. Simple extensions to existing languages will remedy the problem in the short term. Portability is often at odds with efficiency; it is our hope that both will be accommodated in these new environments. A long-term solution involves the development of new innovative languages which will permit a more direct mapping of applications onto parallel architectures.

As they become available, new parallel machines must be placed in an experimental scientific and engineering environment in order for the active user community to learn how to develop programming techniques and skills in the new environment. By experimental, we mean that the emphasis initially should not be on how fast one can compute on standard problems, but on how one can compute on these machines for problems seen as unconventional and inaccessible on existing serial machines.

These machines should be made widely available to the scientific and engineering community over the proposed national network (see above). Then a significant set of experiments on the development of software for various applications as well as experiments on strict computer science issues, such as new operating systems and innovative languages, can be performed.

It is critical that there be broad access to the new experimental machines as they become available. Many different problems from many different domains must be tried on the experimental machines.

As parallel processing comes of age we will need to reexamine existing traditional algorithms. We will have to rework existing algorithms and develop new ones to fit into the parallel setting. It is also critical that new theoretical results be incorporated into this work as they develop. Theoretical underpinnings are discussed more extensively in a section below.

*Recommendation*: It is essential that experimental parallel processing machines be provided to users on a timely basis. Access through a network should be satisfactory for most users. In conjunction with this delivery of hardware to experimental users, we urge the support of joint efforts among discrete mathematicians and theoretical computer scientists for the investigation of key problems in the application of theory to parallel computing. This program should encourage, and perhaps insist on, strong interaction among application specialists, theoretical computer scientists, software tool builders, and experimental computer scientists.

## Neural Networks and Cellular Automata

Massively parallel processing machines offer a significant opportunity to examine anew some longstanding and controversial computational matters. We have picked two as deserving attention in the near future on machines already available; other examples may be as fruitful. The two currently stirring up much interest and controversy follow.

(1) Simulating the "learning" capabilities of simple processing elements coupled to a large number of other simple elements. These neural networks can store patterns taught by an outside source and recall them from noisy inputs. Exploration of various algorithms for coupling the simple processors from the points of view of efficiency of computation, accuracy of pattern recognition, and ability to filter out input noise can be done in a rapid and systematic fashion on several newly available multiprocessors. These pattern recognition algorithms are attractive for gaining experience on massively parallel machines (e.g., The Connection Machine) and for testing their performance.

(2) Simulating the partial differential equations of science and engineering, especially those of fluid flow and heat transfer, by simplified computational models of the microphysics, which is naturally suited to parallel processing environments. Configuration and velocity space rules for the microscopic physics can be implemented as few degree of freedom (small bit) representations, called cellular automata of the local physics. Averages over this gross microphysics can yield the fluid dynamics equations, for example, for low Reynolds' number and low Mach number flows. The method, which is 20 years old, needs massively parallel machines with simple processors and a small amount of memory at each node to accurately and efficiently represent the micro-dynamics in two or three spatial dimensions. Exploration of flows in complex geometries and in regimes of transition to turbulence may be very efficient and revealing when done on existing and near-future parallel computers.

We urge careful development of parallel processing codes for these problems both as a test bed for learning how to use present and near-future parallel architectures and for the promise in illuminating aspects of these important problems. Performing the calculations described could be done in conventional ways on existing serial machines, but both problems lend themselves to parallel computation in a natural fashion. The goal will be to explore the advantages of performing parallel computations. Preparing for future parallel computers with enhanced nodal capabilities and increased communications bandwidth is the main goal here.

*Recommendation*: Since there are several massively parallel machines now available, experiments should be conducted to simulate neural nets and cellular automata. Considerable study of

these problems is needed before the design of specialized machines is considered.

## Advanced Architecture Research

Parallel processing has the potential of breaking through the performance barriers faced by conventional supercomputers. The emerging parallel architectures will provide a basis for a new generation of supercomputers that are not only fast but highly parallel. It will take time for the research community to change to parallel computing. Now is the time to start this process because the available generation of parallel systems has reached a stage of maturity sufficient to support experimental applications (see above).

While this process of transition is occurring, the next wave of parallel architectures must be developed in such a way so as to both exploit the experience gained by the experimental users and to employ the power of the new machines to help accelerate the design process itself. The machines so designed will then become the first wave of highly parallel supercomputers. This can be achieved by aggressively pushing all aspects of the "technological system" toward this new frontier.

The following activities should be pursued:

(1) Accelerate the development of the underlying technologies required to support the aggressive implementation of the next wave of parallel architectures in areas such as submicron and wafer-scale integrated circuits, and optical coupling of system components.

(2) Accelerate the insertion of advanced technology in the initial generation of parallel architectures through larger and faster memories, faster coupling, and faster processors of advanced design. Develop the technology to enable this process to be accelerated through support of the design, prototyping, and production process. This will enable the development of "tera-operation" systems through appropriate combinations and specializations.

(3) Accelerate the development of software technology required to support the effective exploitation of parallel systems that will enable users to focus on specifying the problem to be solved and suggesting approaches to a solution rather than

the current approach of explicit programming for parallel execution. This will require the development of the following: innovative models of computation; nonprocedural languages; new optimizing compilers; integration of symbolic, numeric, and adaptive computing techniques.

(4) Aggressively develop new computer architectures to directly support nonprocedural languages. Such languages do not require explicit parallel programming constructs but do offer efficient means to achieve parallel execution. Current parallel machines do not execute these languages well, so new computer architectures must be developed to support these languages directly.

(5) Develop scalable I/O and mass storage technologies.

(6) Develop an open common software and system design environment that may be used to support new design approaches and new architectures as they emerge.

While it is clear that new and innovative parallel computer architectures are the key to high-performance computing, new parallel computer architectural ideas must go through a critical evaluation phase before the hardware implementation phase. In the evaluation phase the new architecture must be shown to be superior in some way to the parallel architecture classes that are currently in existence. The evidence for superiority must come from an extensive set of simulations on parallel algorithms for a given problem domain. The simulations should be written in a language suited to the architecture and have a built-in comprehensive set of performance measurements that are automatically made when the simulation is run. Evidence of the superiority can then be based on a comparison of the new architecture against the performance measurements of the existing parallel computer architectures.

*Recommendation*: A new and aggressive development of highly parallel supercomputers that incorporate new architecture developments in the support of nonprocedural languages and innovative models of computation should be initiated. It is critical that government, university, and industry researchers be marshaled to develop this next generation of innovative parallel computers so that the United States can maintain its lead in this technology.

## Theoretical Underpinnings

Parallel computation represents a significant departure from the experience and practice of nearly 40 years of serial machines. Much is to be learned from putting multiprocessor machines into the scientific and engineering community and encouraging experimentation on both conventional and unconventional problems. Enormous benefits will result from learning more about parallel computation rather than concentrating on how fast a single process can be executed. However, this investigation of parallel computation requires a fuller fundamental understanding of the manner in which parallel processors can work than we presently possess. Thus we must develop the theory of parallel computing. Computer systems represent one of the most complex and dynamic systems ever constructed. Parallelism raises the issue of the complexity of these systems by a qualitative degree. This complexity makes these systems prone to error, if not complete failure. In addition, the complexity causes algorithm development and performance analysis to be extremely difficult. These problems represent a significant and important challenge for the computing community as the use of parallel computers increases. Numerous engineering failures have occurred due to lack of adequate theoretical foundation in areas related to parallelism.

In this view, it is widely recognized that testing can only demonstrate the presence of engineering errors, not the absence of errors. Only mathematical techniques can prove the correctness of designs under all cases. Such techniques become essential as parallel-computer systems are introduced into life-critical systems.

The theory of parallel systems is in a preliminary stage. Early results are very promising but point to the need for fundamental new results and understanding in areas of discrete mathematical logic. Theoretical computer science and discrete mathematics have been underfunded relative to the actual needs of the nation. In addition, there is a need to encourage a focus by the theoretical community on the problems of particular importance to experimentalists and practitioners. A careful increase in the funding and research management with this focus would strengthen the discrete mathematics and theoretical computer science community over the long term, as well as bridge the gap with practice. We see great benefit in coordinating the funding agencies in providing a stable, directed funding base for this work, particularly by encouraging more students in these areas. This approach to funding would represent a true commitment to long-term support, and it need not require a particularly large budget for that purpose. Autonomy

of funding by different agencies would not be affected. We suggest a basic program as follows.

Each funding agency should be encouraged to establish a program for research by discrete mathematicians and theoretical computer scientists in which the investigators identify and establish a close connection with some experimental work in parallel systems. Funding levels should be established at higher than conventional levels for theory and mathematics. Specific emphasis should be placed on encouraging students and investigators to gain familiarity with practical issues and to interact with experimentalists and developers. Evaluation of investigators should be based not only on research excellence in their areas, but also on their contributions to solving the key practical problems of parallelism.

*Recommendation*: A program of research funding that specifically encourages discrete mathematicians and theoretical computer scientists to investigate key problems in the theory of parallel computation should be instigated. The funding program should facilitate and encourage interaction with experimental and development work.

## Conclusions

For a given technology, the speed of light limits the speed of a simple processor. Parallelism at all machine organization levels is needed to gain a substantial improvement in performance. A number of initiatives are needed to achieve this. We must build a national network to enable parallel computation at the largest scales feasible. We need to investigate and develop new parallel machine architectures. We need to experiment with the new machines. We need to support research in the theoretical underpinnings. Only by building the infrastructure, technology, and research base in parallel computing will the United States be able to maintain its lead in supercomputing and all the scientific and engineering disciplines that are rapidly becoming dependent upon that base.