# 27

# *Supercomputing Opportunities for the Protein Sciences*

*Larry Smarr*

Science has historically relied on two basic methodologies: theory and observation/experiment. In addition to these two modes, which are many centuries old, a third mode has arisen in the last half century: computational science. Computational science involves the use of digital computers to manipulate either equations or data. Although theory provides the equations, the analytic tools of theory are only adequate for obtaining solutions for special geometries or for simplified versions (linearized or perturbations) of the equations. Solutions of these equations that exhibit the realistic complexity seen in nature require numerical computations. Massive data sets, arising from modern laboratory instruments, are meaningless by themselves until analyzed by numerical means. Not only can computational science serve as a strong link between theory and experiment, it can also lead to new insights and knowledge in its own right.

The power of digital computers has increased exponentially for decades and should continue to do so for decades to come. This means that the computational scientist must gain access to new equipment on a time scale of every few years or fall behind competitively. The United States has been struggling with how to provide such technological access to university researchers for the last three decades.

In the 1960s, the federal government provided the funds needed to set up first-rate university computing centers. However, for the 15 years between 1970 and 1985, the federal government removed itself from maintaining these facilities at the state of the art. During that period of "feudalism," scientists were on their own in obtaining computing resources. Very few had access to the newest computational technologies. There was a radical reversal of this policy of "benign neglect" in 1985 when the National Science Foundation (NSF) formed the national supercomputer centers and began the national NSFNET network. That this change will also transform American industry, and thereby the national economy, has been explicitly recognized. Both the Executive Branch (Office of Science and Technology Policy, 1989) and the Congress (U.S. Congress, 1989; Senate Bill 1067, 1989) have issued strong calls for a national initiative in high-performance computing and networking. These programs in the 1990s will involve new forms of collaboration between government, universities, and industries.

In the last four years, more than 20,000 university scientists, engineers, social scientists, and humanists at over 250 universities and colleges have gained access to virtually all forms of frontier computing technologies housed in the national centers. Such access requires peer review of proposals in order to maintain a high standard of research excellence. There is a factor of 100 to 1000 times the

computing speed, memory, and storage capacity in the national centers as is available in the desktop computer of the typical individual scientist. The national network, which allows the researcher to "reach out" and grab that extra power when needed, has 1000 times the bandwidth of a user's access path four years ago. There are roughly 100 staff at each of the five NSF supercomputer centers to support the remote user and to provide a high level of training and consulting.

By the mid 1990s, the computing power on both the desktop computers and on the central supercomputers will be at least 100 times what it is today. The bandwidth of the national network will rise by another factor of 1,000. We are also seeing a fundamental change in software. User interfaces are moving from command line, character-oriented, single screens to menu-driven, bit-mapped, multiple-windowed environments. Easy-to-use tool sets in the public domain are being developed to handle routine tasks in file management, graphics, and telecommunications. Many important concepts in computer science (e.g., object-oriented programming, distributed computing, data structures) are becoming practical tools for computational scientists. Finally, a wide range of visualization technologies are changing the primary unit of information from number to image. Because of this enormous rate of technological change and the great democratization of access that the NSF has provided, the period in which we are now is one of historical change in computational science.

This chapter will give an overview of the emerging national information infrastructure that will enable many exciting discoveries in the computational sciences, including the protein sciences. The near future will see most researchers having access to powerful and flexible desktop computers linked over a national network. "Supernodes" on this network will contain high-value resources such as supercomputers and national data banks. Scientific visualization capabilities will be ubiquitous. The ability to bring together on the desktop the results of both complex simulations and detailed observations, and to be able to interact with each data set visually as well as quantitatively, could profoundly influence the progress of the protein sciences.

## Computational Protein Sciences

There are many current applications of advanced computing to the protein sciences, some of which are described in detail in other chapters in this volume. Since my research area is relativistic astrophysics, I will not presume to review the biology or biochemistry involved, but will instead give a brief overview of some of the research areas in the protein sciences in which I see computational progress being made: imaging of biological polymers, genome sequencing, determination of protein structure and function, and the protein folding process.

The direct imaging of the polymers of proteins and nucleic acids is bringing new technologies such as the scanning tunneling microscope into the biological laboratory. These images of individual molecules are digital and require image processing to reveal fine details of structure. Because imaging instruments create digital data sets, "multisensor fusion" is becoming practical. This approach combines data from various modalities such as optical, x-ray, and scanning tunneling microscopy into a single, digital, synthetic "object" that can be explored interactively on a workstation. It would be very desirable to make these data sets available electronically to the national community.

The rapid rise of interest in genome sequencing is producing ever larger numbers of the amino acid sequences that code for proteins. Steps are being taken in the community to assure that these sequences are available to all researchers in national data banks. However, the sequences require annotations in order to be scientifically useful. In many cases researchers would also like to see the original gel images. Thus, the need arises for a single file structure that can accommodate symbolic, numerical, and image data. Furthermore, a new generation of application software is being written to allow for com-

parisons and searches among the accumulating data banks.

The three-dimensional (3D) atomic structure is known for some 300 proteins. Since this is only 10 percent of the number of genes on a single bacterial DNA (which itself is only 1/1000 of the length of the human genome), it is clear that researchers will soon have far more linear sequences than they will have the folded 3D structure for proteins. Hence the interest in computational protein folding. Nonetheless, advances in x-ray crystallography and 2D nuclear magnetic resonance (NMR) hold out the hope for many more structures to be revealed experimentally. Some supermolecular structures such as viruses and protein/nucleic acid complexes have been solved as well. For the largest structures, supercomputers are required to reconstruct the 3D structure from the experimental data. The researcher again requires remote electronic access to national data banks (such as at Brookhaven National Laboratory). In addition, workstations capable of running 3D graphics software interactively are necessary to comprehend the hierarchical, organized complexity of protein structures.

An increasingly common use of supercomputer time is the determination of protein function from protein structure. Proteins carry out their function in an aqueous solution. In such an environment, they continuously flex and interact with other molecules. Molecular dynamics software can calculate the motions of the atoms in the protein, and computer graphics software can translate these computed trajectories into visual animation. However, even thousands of hours of time on the fastest supercomputers can compute only a tiny fraction of the possible motions of a large protein. Quantum chemistry codes can approximately determine the electronic structure of proteins. From this structure, researchers are beginning to study chemical interactions between molecules and electron or proton tunneling within molecules.

The one key area that is missing from the computational protein sciences today is the capability to predict the three-dimensional folded protein structure from the linear amino acid sequence. This seminar has been very useful for bringing together many of the leading researchers in the field to exchange views on how to proceed. It seems clear that there is not one "royal road" to the solution of the protein folding problem. Instead, I expect to see new algorithms brought into this field from other disciplines such as statistical mechanics (Wolynes, in press), computer science, and mathematics that will qualitatively change the computational approach. In any case, for the foreseeable future, this central problem of biology will require immense computing resources. Since other chapters in this volume deal with some specific techniques for computational protein folding, I will not discuss details of the relevant algorithms further here.

Rather, I will describe the wide variety of computational and communications technologies necessary to support the broad range of protein sciences described above. From these thumbnail sketches of the major subfields of the protein sciences, we can see a number of common technologies are required. Remote digital access to data banks of images, sequences, and structures as well as applications software, electronic mail, and the scientific literature itself needs to be easily available. Both local and remote storage capacity for exponentially increasing data sets must be provided.

A wide range of visualization capabilities—from x-y plots to color three-dimensional, real-time animations—are required to probe the complexity of proteins. All of the computers used by the researcher should be transparently integrated by distributed system software. A computational environment on a workstation configured with a Graphical User Interface (GUI) should tie together all the software tools and remote computing resources needed to perform protein research. One might term this device a "protein workstation."

In addition to these technological requirements, a higher level of human teamwork is necessary. Modern computational science demands not only detailed knowledge of the science itself but also of the computer science and technology that is being used to solve the

problem computationally. Therefore, consulting, training, and education services covering all aspects of the computational technologies and their applications to scientific research should be available to the end user. Particularly in the protein sciences, research progress will also require a high level of interdisciplinary collaboration between individuals and institutions in universities, industry, and the government.

These computational infrastructure needs of the researcher are not specific to the protein sciences. Indeed, it is their universality that makes it possible to design a national information system by spreading the total cost over all areas of science and engineering in both pure and applied research.

## The Emerging National Information Infrastructure

The 1990s will witness the establishment of a fiber optic network that will link together most major research facilities, whether in universities, industry, or government, to provide a comprehensive national computational environment with the capabilities described above. The beginnings of this infrastructure have already been laid. As mentioned above, NSF has created a number of national supercomputing centers to serve the university community and begun a national network, NSFNET, which today hooks most research universities together through over 1000 local, regional, and state networks tied to the national backbone (Fig. 1). In addition, nearly 20 universities have purchased their own supercomputers, and many more have created campus networks. Other government agencies with national laboratories as well as many large corporations have supercomputers with limited access to the university community but with strong in-house research groups. Personal computers and workstations are becoming very widespread on researchers' desktops. Finally, there is a rapidly improving family of application software in both public domain and commercial forms that can attack most of the problems that arise in protein science.

In this section, I will review some of the key computational technology trends that protein researchers should be aware of as this infrastructure emerges. I will divide the discussion into desktop computers, networks, and supernodes.

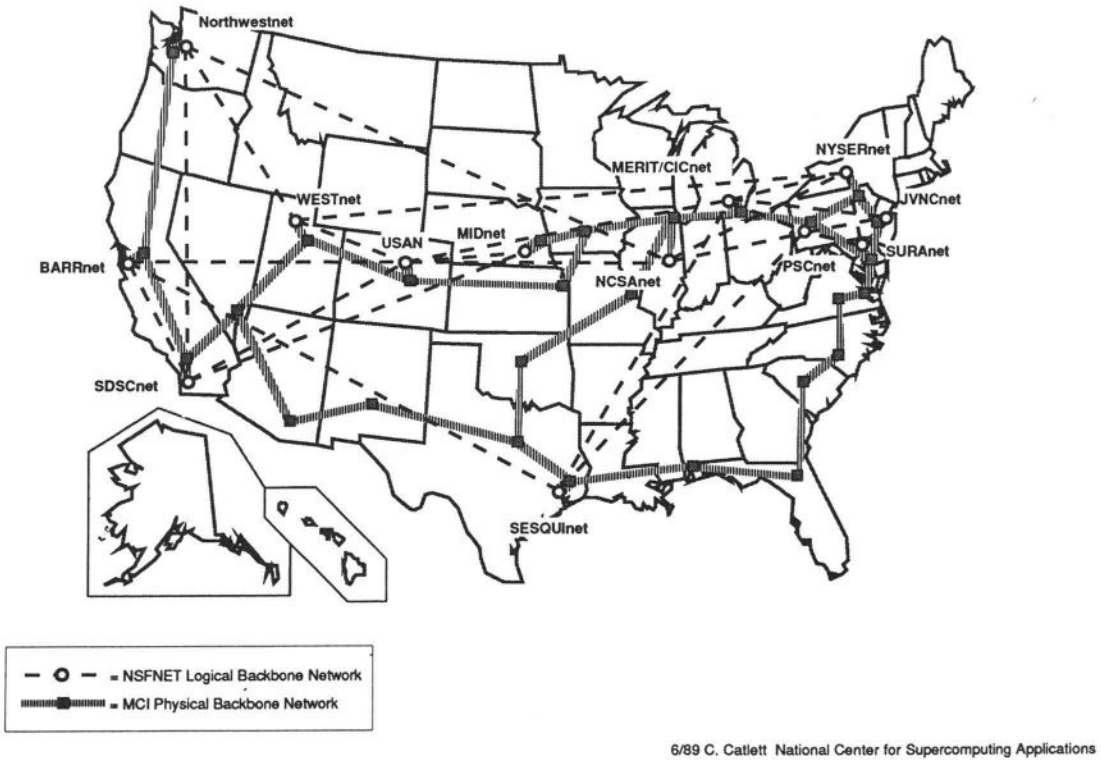



**Fig. 1.** The NSFNET national network. Shown is the logical and physical backbone with labels for the regional networks.

### *Desktop computers*

In 1989, there are well over 20 million personal computers and workstations on desktops in America. Although the great majority of these are engaged in business activities, an increasing number are being used by researchers. Originally, these machines were character oriented and command line driven single screens, but, in the last few years, they have moved to bit-mapped graphics, "point and click" interface, with multiple windows. These desktop computers, which have previously been used primarily as stand-alone machines, are increasingly being used as "smart ports" into a networked distributed computing environment. As the speed of all the computers in the network increases exponentially, users are turning to visual representations of the massive data sets they are producing.

A key trend for all of the major vendors of desktop computers is to provide a standard GUI with a "cut and paste" capability across all application software. At this writing, this is complete only for the Apple Macintosh. Therefore, I will focus below on what can be achieved in this environment. During the next 18 months, the other major vendors will complete their transition over to this type of advanced interface. Complete GUIs are now emerging for each major operating system, for instance: Microsoft's Windows for DOS personal computers, Presentation Manager for IBM's OS/2 operating system, and OPEN-LOOK (AT&T, Sun Microsystems, Inc., and others), Motif (Open Software Foundation—IBM, DEC, HP, and others), and NextStep (NeXT and IBM) for UNIX workstations. Since the industry has not standarized on a complete GUI, many third-party application software developers have not yet modified their codes to tie tightly into the GUI. In the interim, most vendors have agreed to a partial GUI, "X-Windows," as a standard windowing system, but one which does not provide as easily for moving across applications.

These desktop machines run commercial productivity software such as word processing, data bases, spreadsheets, and modem communication software, which plays a crucial role in the day-to-day life of a researcher. In addition, one can purchase a variety of 2D and 3D computer graphics software packages. Except on the Macintosh, these are stand-alone applications. While extremely valuable and powerful, they will become even more so when one can uniformly move data, images, and annotations easily across applications. This will result not only in greater productivity for the end user but also in a much reduced training barrier than exists today.

I will illustrate some of the potential capabilities of such future software tools, integrated with a standard GUI, by describing the current National Center for Supercomputing Applications (NCSA) software suite. These are general tools needed by anyone computing in today's distributed environment. These tools link personal computers, remote workstations, and supercomputers into a cooperative computing environment to enable the researcher to do computations, share files and data across machines, generate images from data sets, analyze and animate those images, produce color maps, and create presentation-quality graphics. Because these software tools were not available commercially, NCSA has created a software development group to produce public domain[1] software tools for the desktop. This group consists of computer scientists and computer professionals who receive functional design input from the computational scientists using the facilities of NCSA.

In today's distributed computing environment, the most fundamental need is for connectivity. As local ethernets, campus networks, and regional networks merge with the national backbone, a user needs to be able to run multiple computers simultaneously from

---

1 The NCSA software suite is released into the public domain and may be downloaded at no charge from an anonymous file transfer protocol (FTP) server at NCSA (via Internet addresses ftp ftp.ncsa.uiuc.edu or ftp 128.174.20.50), or ordered through the NCSA Technical Resources Catalog. To obtain an NCSA Technical Resources Catalog, contact: NCSA Documentation Orders, 152 Computing Applications Building, 605 E. Springfield Avenue, Champaign, IL 61820 (217) 244-0072.

his or her desktop and transfer files between them. Fortunately, NSF has insisted that all NSFNET access follow the standard "TCP/IP" protocols. NCSA Telnet provides the user with TCP/IP connections from DOS, Macintosh, and UNIX workstations to remote hosts with multiple window sessions, FTP file transfer, and Tektronix and raster image capabilities. Figure 2 shows a photo of a Macintosh II screen running NCSA Telnet. Each window is running a session on a different computer on the national network (here, a remote Cray Research, Inc., supercomputer, a DEC VAX superminicomputer, and a Sun Microsystems network server) and each session is independently scrollable in each window. The information in the window can be cut and pasted into any other Macintosh application package. We estimate that NCSA Telnet has tens of thousands of users today.

Once one has data sets, either from computation, experiment, or observation with modern sensors, the user needs a simple set of tools to explore the data visually and quantitatively. At this writing, the available NCSA software can handle two-dimensional data sets, of which I will give examples below. We have similar tools for three-dimensional data sets in development. NCSA Datascope can be used to create a color raster image representation of 2D data by simply pulling down a menu.

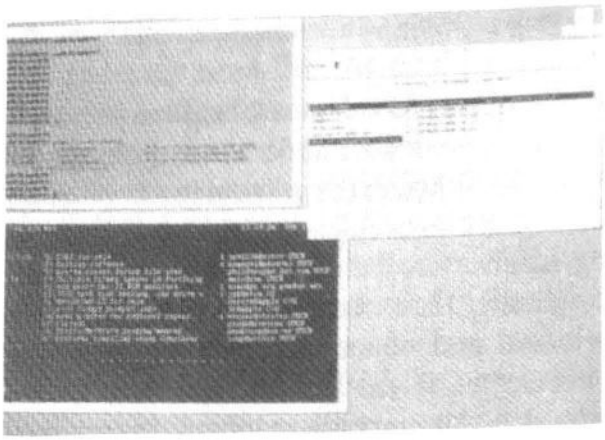
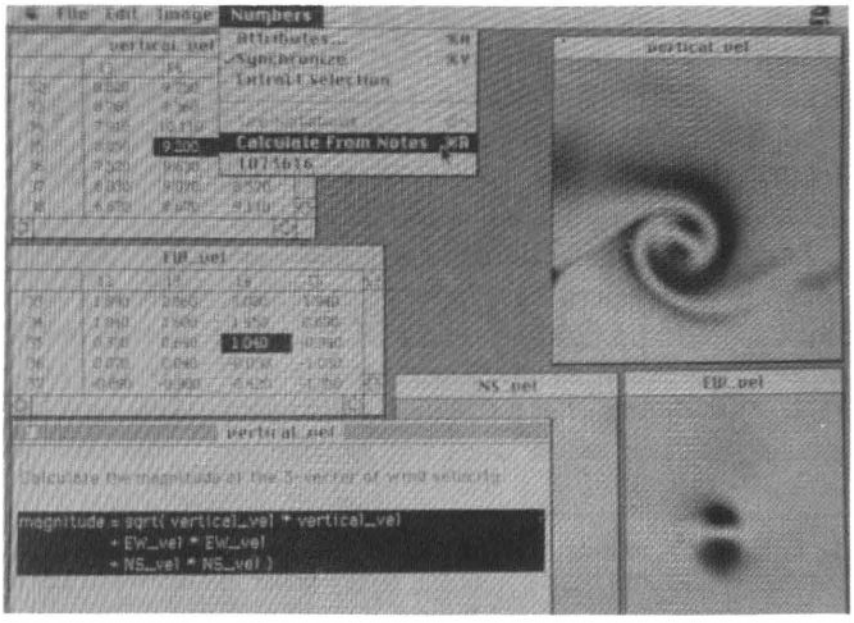


**Fig. 2.** NCSA Telnet running three sessions on different remote computers.

Figure 3 shows a fluid flow computation being explored on a Macintosh II using Datascope. (The original screen was in color but is printed as black and white in this volume.) Note that the data are held in electronic spreadsheet form, so that the numbers remain "live." An electronic notebook allows one to compute functions of the data by simply writing down the desired formula. The mouse can be used to pick out any subset of the data and this area is simultaneously shown on the image.

This image can be easily moved into NCSA Image, which allows one to obtain line graphs of cuts through the data, contour and 3D surface plots of the data, or animations of a series of images. Rudimentary image

**Fig. 3.** NCSA Datascope analyzing a vortex fluid flow.



processing can be achieved by manipulations of the color transfer function, which is the method by which data values are turned into colors. These are all mouse driven and require no programming. NCSA PalEdit is another tool with which one can construct custom color transfer functions or "palettes." Finally, NCSA Layout allows one to design presentation graphics containing raster images. Again, one can cut and paste all these images, palettes, graphs, words, and equations into any Macintosh application. This allows one to create an "electronic research notebook."

In order to make the capabilities of the NCSA suite available on other vendor platforms which do not yet have a complete GUI, NCSA has developed a set of tools (e.g., NCSA X Image, X DataSlice) that run on any workstation that supports the X-windows standard. Figure 4 shows a screen where a user is studying another fluid flow simulation with NCSA X Image. This greatly extends the number of end user desktop computers that can work in the modern bit-mapped, multiple-windowed environment. However, until the other vendors GUIs are complete, one will not be able to move as freely among all application software on these machines as one can today on the Macintosh.

Today these tools handle only 2D data sets, which are being used by protein biologists and chemists to study such widely varying topics as sequencing gels, electronic orbitals, and microscopic images. The study of protein structure and function, of course, requires a 3D visualization capability. There are many powerful, stand-alone commercial packages for molecular graphics, many of which are demonstrated in figures elsewhere in this volume. Both these software developers and NCSA will be working to integrate these capabilities into future standard GUIs.

The cost of desktop computers capable of not only displaying 3D graphics but interacting with them as well has recently decreased to the point where most researchers can afford one. The introduction of the Personal Iris, from Silicon Graphics, Inc., in fall 1988, is indicative of this important trend. As very large systems integration (VLSI) drives specialized chip prices ever lower, this situation will continue to improve rapidly. It is still too computationally intensive to have full digital real-time 3D dynamics on an inexpensive desktop machine. Some high-end graphics workstations are entering the market and some centers like NCSA have experimental "numerical laboratories" for real-time interactivity with dynamical systems. This capability will be widely available by the mid 1990s.

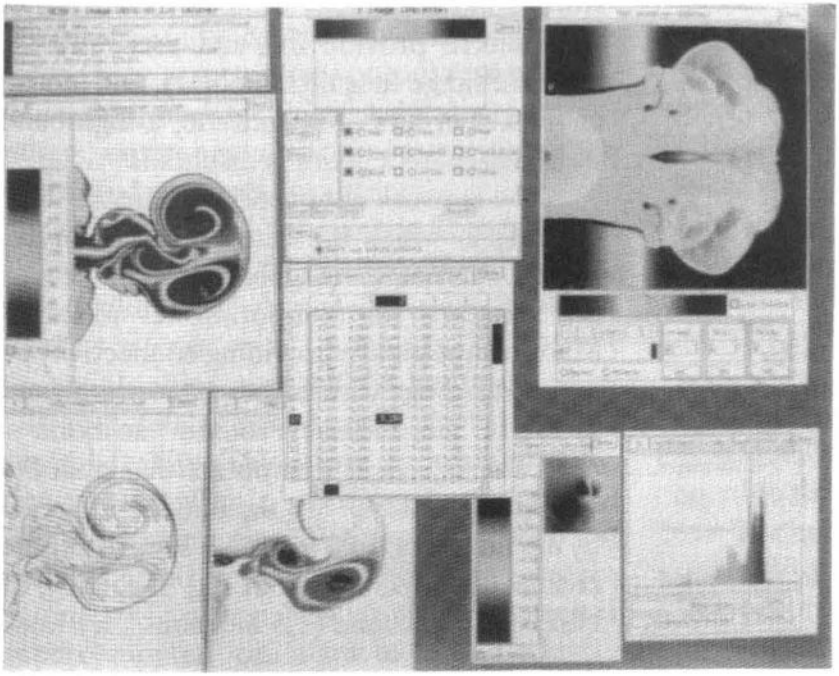


**Fig. 4.** NCSA X Image software used on a fluid flow simulation.

Finally, one needs a method for storing and transferring multi-object files across different machines on the network. For example, multidimensional floating point data arrays, palettes, images, and annotations are preferably kept together under one file name, as in the genome example mentioned above. Furthermore, one does not want to have to bundle and unbundle these objects by hand. NCSA developed the Hierarchical Data Format (HDF) to accomplish this. Figure 5 shows how the HDF allows different vendors' computers on the network to automatically access these combined files from selected software. The user's application code can read and write HDF files. NCSA's HDF is not yet a national standard, but we are working with others to create such a standard file format to ease the data exchange problem between researchers and computers.

A hardware breakthrough is going to radically alter the power of the desktop computer very soon in certain aspects of protein science. Parallel arrays of microprocessors called "transputers" can now be placed on add-in boards for workstations. Mature applications codes such as molecular modeling or genome sequencing are beginning to be mapped onto these transputers. The computational speed of these boards can approach that of a supercomputer, for a price of tens of thousands of dollars[2]. Of course, they do not have the flexibility of a general-purpose supercomputer, but, for certain molecular design or sequencing studies that do not require the large memory of supercomputers, they can allow a researcher to interact rapidly with the molecule under study. When these transputers are combined with 3D interactive graphics software by the early 1990s, we will be close to the dream protein workstation described in the first section of this chapter.

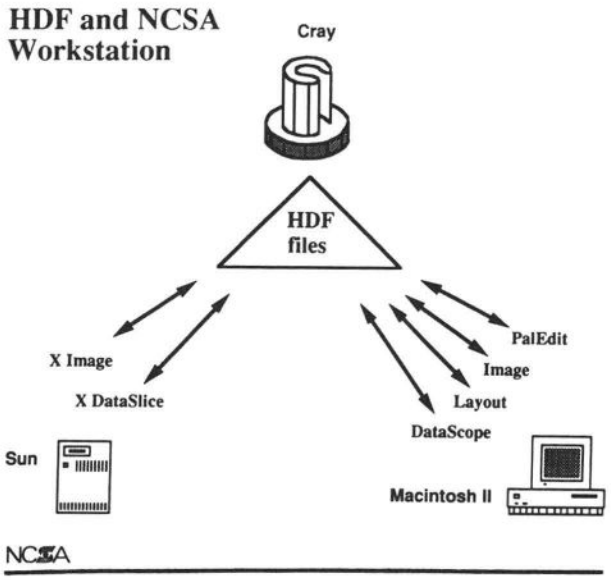



**Fig. 5.** NCSA Hierarchical Data Format and Distributed Computing.

## *Networks*

Even if one can increase the speed of specialized applications on desktop machines to near supercomputer speeds, total memory and storage is still a limitation on the science. A "full-up" personal computer will have 8 megabytes of memory and 100 megabytes of disk. A much more expensive high-end workstation may have as much as 128 megabytes of memory and 1000 megabytes of disk storage. Our Cray-2 supercomputer has 1000 megabytes of memory and our central on-line disk storage has a capacity of 100,000 megabytes. In order to provide the end user with this dynamic range in speed, memory, and storage of several orders of magnitude, a high bandwidth national network is essential.

The network connection originates in the building in which the researcher is working. Here Ethernet (1 megabit/sec) has become the standard for computational scientists. Many campuses are providing connectivity between buildings with higher bandwidth (such as the 80 megabit/sec Proteon backbone on our campus). Host computers designated as "campus gateways" are then hooked together by the many regional and NSF supercomputer centers networks shown in Fig. 1. Finally, the NSFNET backbone (1.5 megabits/sec) hooks

2 Those interested in this approach should contact: Professor Klaus Schulten, Beckman Institute for Advanced Science and Technology, 405 North Mathews Avenue, Urbana, IL 61801

the regional nets together with the national supercomputer centers. By early 1989, the entire national network had "jelled," and service is now quite reliable.

Clearly, we have not finished the job. The backbone should have a greater bandwidth than the "tributaries" flowing into it. Therefore, plans are under way to take some legs of the backbone to about 45 megabits/sec during 1990. Other government agencies, such as the National Aeronautics and Space Administration (NASA), the Department of Energy (DOE), the Department of Defense (DOD), and the National Institutes of Health (NIH) have networks and major data repositories that need to be merged with NSFNET. All this is in the implementation stage. By the mid to late 1990s, a national fiber optic "superhighway information system" will be in place to link all these sites at bandwidths in excess of 1000 megabits/sec.

During the past few years, much attention on this critical issue has been focused by the bill introduced in Congress by Senator Albert Gore, Jr., to create such a national network. In September 1989, the Executive Branch (Office of Science and Technology Policy) joined this effort with its plan to establish the National Research and Education Network (NREN). This network would extend NSFNET to researchers at universities, national laboratories, private companies involved in government-supported research and education, and facilities such as supercomputers, experimental remote instruments, data bases, and research libraries.

Going from 1200 bits/sec (modems) in 1985 to 1,000,000,000 bits/sec (fiber) in 1995 will qualitatively change the ways in which humans interact with computers. To attempt to see into this future, NCSA has installed an Ultranet (800 megabit/sec) network in our central facility. This connects a high-resolution frame buffer to our Cray-2 and Cray X-MP supercomputers. We are experimenting with selected scientists (including protein scientists) to see how such a "personal supercomputer" can allow us to tackle interactive simulations at a new level. This bandwidth is becoming an industry standard (called the High Speed Channel (HSC)), which will start to become commercially available on a wide range of nonsupercomputers in the years ahead.

The net effect of such high bandwidth is that it eliminates distance as a consideration in computation. As a result, this will alter not only remote computing but also scientific collaboration. Essentially, all researchers and their scientific instruments will be tightly linked in a National Collaboratory (Lederberg and Uncapher, 1989). Again, a very synergetic interaction between computer scientists and computational scientists will produce the new research paradigms for working in this digital world.

### *Supernodes*

As mentioned earlier, specialized centers are currently on the national network and more will arise in the future. These include not only general purpose supercomputers, but also novel architecture computers, high-end visualization facilities, massive file storage capacity, and specialized data centers. I will refer to any such nationally shared network node as a "supernode" (Press, 1981). They make sense whenever it is more cost effective to place specialized capabilities centrally rather than to distribute them. This approach is much the same as that used in astronomy, where there are many small telescopes available at universities (for routine work or for training observers) and a few national observatories such as the Space Telescope or Very Large Array radio telescope (for frontier research allocated by peer review). One cannot do the same science with 1000 pairs of binoculars as can be done with a single large aperture telescope, even though the combined light-gathering power of the binoculars is equal to that of the large instrument!

The modern supercomputer achieves its speed through vector and parallel processing. Vector processing allows the computer to process arrays of numbers ("vector") instead of single numbers ("scalar"). Parallel processing allows multiple tasks to occur simul-

taneously. The vector compilers of today have had more than a decade to develop in a highly competitive marketplace, so the user does not have to do much in the way of special programming to take advantage of this feature. Parallelism is much more complex. First, there are many types of hardware implementation of the idea of parallelism. Second, the "smart compilers" for these hardware designs are just emerging into the marketplace. Thus, it is more difficult today for the end user to take advantage of parallelism than it is of vector architectures. This is rapidly changing; the NSF national centers are playing a key role in this process by providing access to different computer architectures. This will allow application codes in many fields to be redeveloped with new algorithms that are most efficient on each architecture.

To give the reader some idea of what to expect in the way of speed and memory increases during the next five years, Fig. 6 shows the proposed upgrade path of NCSA's supercomputers through 1995. The units on the left are in units (of memory and speed) of the Cray 1 supercomputer of 1980. To obtain this increase in performance for the end user, I am assuming that the user's code has been rewritten to take advantage of both the parallel and vector hardware. Currently the top Cray Research, Inc., supercomputers have either four (Cray-2) or eight (Cray Y-MP) processors. Future machines from Cray Research, Inc. and other vendors will have 16 (1992) and later 64 (1994) processors. Furthermore, the speed of each processor in 1995 will be at least six times faster than those available today.

Massive parallelism is just emerging as an alternative approach to supercomputing in the marketplace. NCSA took delivery in 1989 of a 32,000 processor Connection Machine (CM) from Thinking Machines, Inc. The speed of this machine is comparable to (or in some cases faster than) current supercomputers that possess only a few processors because, even though each CM processor is very slow compared to a conventional supercomputer processor, the massive number of processors overcomes the slower speed. In addition, the CM is aircooled rather than having the complex liquid cooling required for conventional design supercomputers. The software environment on such radical architectures is, of course, not as highly "battle tested" as on decade-old architectures.

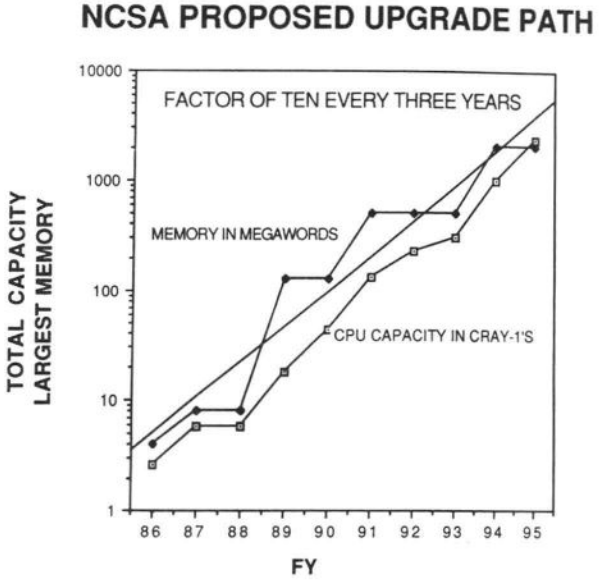



**Fig. 6.** Proposed upgrade path for supercomputers at NCSA. Units at left are in terms of a 1 megaword Cray 1 supercomputer of 1980.

Balancing this lack of application software is the fact that today's user has much more powerful code development environments than 10 years ago. Thus, I expect the 1990s to see a real race between supercomputers built around a few powerful processors and those built around hundreds to thousands of slower processors. The end user will once again be the winner as a wider array of capabilities becomes available.

As the unrelenting exponential increase of supercomputer power continues, the need for novel approaches to storage is becoming acute. Centers such as NCSA use mainframes to act as file servers for the hundreds to thousands of remote users. The on-line capacity is around 100 gigabytes and the off-line capacity is over a terabyte (tera = 1000 giga). These figures will grow by at least an order of magnitude by 1995. In addition to this, one needs to have random access to individual numbers in massive data sets. For this purpose, laser disk jukeboxes or optical tape with capacities of a terabyte or more are coming into use. At

the desktop, CD-ROMs, read-write laser disks, and high-density videotape are being used to provide massive local storage. Although visualization was a response to how to scan large amounts of data, the storage of thousands of images (each 1 megabyte) for animations is itself becoming a new driver of mass storage needs. In this field, as in desktops, networks, supercomputers, and algorithms, the next five years will see a great deal of experimentation with new technologies.

With the enormous increase in available information comes an increased need for sophisticated human interface systems to allow the user to search this information. The Computer and Genomic Systems Laboratory (CGSL) at the University of Arizona, led by Bruce Schatz, is implementing an advanced "digital library" interface system using data related to molecular biology, including the scientific literature itself. While at Bell Communications Research, Schatz developed a prototype digital library system (Schatz, 1985) from which he coined the word telesophy or "wisdom at a distance." The telesophy system provides a mechanism whereby the entire search or selected portions of the search can be stored as an additional information object, complete with links to the collected information. Thus, the insight of the individual performing the search is selectively preserved for other users of the telesophy system. NCSA is working with Schatz and the CGSL to construct an initially limited telesophy system, but with a continually expanding scope to include various scientific disciplines and with a goal of incorporating this important technology into the interdisciplinary environment.

Finally, the most important aspect of supernodes may lie in human factors. Because a large staff (NCSA has over 200 staff members) is aggregated around the specialized resources of supernodes, there is a critical mass effect. This allows visitors to the supernodes to learn about many new technologies simultaneously, using them in a working environment on their own research problems. National workshops, training sessions, vendor presentations, and interdisciplinary seminars are almost a daily occurrence.

For instance, NCSA has recently joined 20 other interdisciplinary research teams in the new Beckman Institute. These teams of faculty, postdocs, graduate students, and nonacademic staff not only cover genome sequencing, molecular recognition, protein dynamics, and tunneling microscopy, but also many other areas in the physical, biological, engineering, and computer sciences. Included in this mix are also major industrial partners to NCSA, including Kodak, Amoco, Eli Lilly & Co., Motorola, FMC Corp., Dow Chemical, and Caterpillar. By having a constant stream of national and international visitors in a state-of-the-art computational facility, we are able to identify many of the emerging trends in computational science. This, in turn, allows NCSA to provide better education and training to our many remote users.

The bottom line is that the next five years are going to be ones of great change in computational science. The critical factor is to keep our minds open and flexible, so that we can all take advantage of these new technologies and methodologies and apply them to our own scientific specialities. I believe that protein sciences will be one of the fields of research most affected. I look forward with a keen sense of anticipation to the new discoveries that will occur.

## Acknowledgments

I would like to thank the members of the NCSA Software Development Group and networking groups for providing the figures, and many NCSA staff and UIUC faculty for comments on the manuscript. Radha Nandkumar and Janus Wehmer were particularly helpful during the final editing. NCSA is supported by the National Science Foundation, the State of Illinois, the University of Illinois, and corporate partners.

## References

Lederberg, J., and K. Uncapher, cochairs, *Towards a National Collaboratory.* Report of an Invita-

tional Workshop (Rockefeller University, 1989).

Office of Science and Technology Policy, *The Federal High Performance Computing Program* (Executive Office of the President, Washington, DC, 1989).

Press, W.H., *Prospectus for Computational Physics.* Report by the Subcommittee on Computational Facilities for Theoretical Research to the Advisory Committee for Physics (Division of Physics, National Science Foundation, Washington, DC, 1981).

Schatz, B.R., *Telesophy Project Report #1* (Bell Communications Research, 1985).

Senate Bill 1067 "National High Performance Computer Technology Act," *Congressional Record* **135,** no. 64 (1989).

U.S. Congress, Office of Technology Assessment, *High Performance Computing and Networking for Science Background Paper.* OTA-BP-CIT-59 (U.S. Government Printing Office, Washington, DC, 1989).

Wolynes, P.G., in *Spin Glasses and Biology* (World Scientific Publishing Co., Inc., Teaneck, NJ) in press.