2.0 Vision for an Advanced Cyberinfrastructure Program

2.1 A Nascent Revolution

Scientists in many disciplines have begun revolutionizing their fields by using computers, digital data, and networks to replace and extend their traditional efforts. The calculations that can be performed and the information that can be archived and used are exploding. In the not-too-distant future, the contents of the historic scientific literature will fit on a rack of disks, and an office computer will provide more computing than all the supercomputing centers together today. The results of today's largest calculations and most sizable collections will take seconds to transmit using the fastest known network technologies. New technology-mediated, distributed work environments are emerging to relax constraints of distance and time. These new research environments are linking together research teams, digital data and information libraries, high-performance computational services, scientific instruments, and arrays of sensors. In many cases these emerging environments for knowledge work are essential, not optional, to the aspirations of research. We see glimpses of the future in some shifts in current research practice:

- The classic two approaches to scientific research, theoretical/analytical and experimental/observational, have been extended to *in silico* simulation and modeling to explore new possibilities and to achieve new precision.
- The enormous speedups of computers and networks have enabled simulations of far more complex systems and phenomena, as well as visualizing the results from many perspectives.
- Advanced computing is no longer restricted to a few research groups in a few fields such as weather prediction and high-energy physics, but pervades scientific and engineering research, including the biological, chemical, social, and environmental sciences, medicine, and nanotechnology.
- The primary access to the latest findings in a growing number of fields is through the Web, then through classic preprints and conferences, and lastly through refereed archival papers.
- Crucial data collections in the social, biological, and physical sciences are now online and remotely accessible – modern genome research would be impossible without such databases, and soon astronomical research will be similarly redefined through the National Virtual Observatory.
- Groups collaborate across institutions and time zones, sharing data, complementary expertise, ideas, and access to special facilities without travel.

The trends represented by these examples will only accelerate. In the future, we might expect researchers to

- Combine raw data and new models from many sources, and utilize the most up-to-date tools to analyze, visualize, and simulate complex interrelations.
- Collect and make widely available far more information (the outputs of all major observatories and astronomical satellites, satellite and land-based weather data, three-dimensional images of anthropologically important objects), leading to a qualitative change in the way research is done and the type of science that results.
- Work across traditional disciplinary boundaries: environmental scientists will take advantage of climate models, physicists will make direct use of astronomical observations, social scientists will analyze interactive behavior of scientists as well as others.
- Simulate more complex and exciting systems (cells and organisms rather than proteins and DNA; the entire earth system rather than air, water, land, and snow independently).
- Access the entire published record of science online.
- Make publications incorporating rich media (hypertext, video, photographic images).
- Visualize the results of complex data sets in new and exciting ways, and create techniques for understanding and acting on these observations.
- Work routinely with colleagues at distant institutions, even ones that are not traditionally considered research universities, and with junior scientists and students as genuine peers, despite differences in age, experience, race, or physical limitations.

2.2 Thresholds and Opportunities

Why act now? Currently observed activities and benefits represent just the beginnings of a revolution. Computers have been improving for decades, and some researchers have tried to do many of the activities listed above. We believe that several key thresholds have recently been reached in the use of IT, in part because NSF has made large and successful investments in a number of research areas, including networking, supercomputing, human interfaces, collaboration environments, and information management. There are many reasons:

- The Internet and the Web were invented to support the work of researchers, and their use permeates all of science and engineering. Broadband networks connect all research centers and enable the rapid communication of ideas, the sharing of resources, and remote access to data. The next generations of the net promise even greater benefits to the research community.
- Most modern researchers are fully conversant with and dependent on advanced computing for their daily activity, and have a thirst for more. Older scientists are learning to take advantage of the new technologies.

- Closed-form analytic solutions are available for a decreasing fraction of interesting research challenges; often only a numeric computation can produce useful results.
- Moore's law has led to simulations that begin to match the complexity of the real world, with fully three dimensional, time-dependent modeling with realistic physical models opening up a vast range of problems to qualitative attacks. They range from cosmology to protein folding – problems formerly considered far too complex to address directly.
- In an increasing fraction of cases, it is faster, cheaper, and more accurate to simulate a model than construct and observe a physical object.
- Increasingly ubiquitous networking and interoperability of information formats and access make high-quality remote collaboration feasible.
- Storing terabytes of information is common and inexpensive; archives containing hundreds or thousands of terabytes of data will be affordable and necessary for archiving scientific and engineering information.
- Computing power that was unavailable only a few years ago – trillions of operations a seconds – can now be found in a number of research organizations.
- Computational and visualization techniques have progressed enormously and provide as much scientific value as improved hardware.
- Most researchers would not be able to function without e-mail or access to the Web. They certainly would have fewer contacts with distant, especially international, scientists and be much less able to stay on the cutting edge of their field.

There are also significant risks and costs if we do not make a major move at this time:

- Absent coordination, researchers in different fields and at different sites will adopt different formats and representations of key information, which will make it forever difficult or impossible to combine or reconcile.
- Absent systematic archiving and curation of intermediate research results (as well as the polished and reduced publications), data gathered at great expense will be lost.
- Effective use of cyberinfrastructure can break down artificial disciplinary boundaries, while incompatible tools and structures can isolate scientific communities for years.
- Groups are building their own application and middleware software without awareness of comparable needs elsewhere, both within the NSF and across all of science. Much of this software will be of limited long-term value absent a consistent computer science perspective. Time and talent will be wasted that could have led to much better computing *and* much better science.

- Dramatic changes are coming in computing and application architectures; lack of consideration of work in other sciences and in the commercial world could render projects obsolete before they deliver.
- Much of the effort under way to use cyberinfrastructure for collaborative research is not giving adequate attention to sociological and culture barriers to technology adoption that may cause failure, even after large investments.

The time is ripe for NSF to accelerate the revolution for the benefit of society. A confluence of technology-push and science and engineering research-pull activities and possibilities makes this the right time. Researchers are ramping up their use of computing resources, starting to store enormous amounts of information, and sharing it. Distributed computing, large clusters, data farms, and broadband networks (typified by Internet2 [21], Grid[22], and Web Services[23] directions) have moved from research to practical use. We anticipate a phase change, where direct attention to this opportunity can have a highly desirable and nonlinear effect.

We envision an environment in which raw data and recent results are easily shared, not just within a research group or institution but also between scientific disciplines and locations. There is an exciting opportunity to share insights, software, and knowledge, to reduce wasteful re-creation and repetition. Key applications and software that are used to analyze and simulate phenomena in one field can be utilized broadly. This will only take place if all share standards and underlying technical infrastructures. Although many of the mechanisms to support the best scientific computing are becoming available through commercial channels, there continue to be special needs that the commercial sector is unlikely to meet directly because of the market size and technological risks.

Scientists must have easy access to the finest tools from the commercial and advanced research sectors, without dampening their creativity and ardor to do even better. Individual researchers expend too much effort, frequently with insufficient knowledgeable computing assistance, to create and re-create computing resources; to access, reformat, and save information; to protect the data and software assets. Much of this work could be done by computing experts and shared across the scientific research community. The ACP will encourage groups of scientists to undertake large coordinated information-intensive projects that can radically change the way they and their peers work, and that will support the sharing and long-term use of information that results from their work.

In summary then, the opportunity is here to create cyberinfrastructure that enables more ubiquitous, comprehensive knowledge environments that become functionally complete for specific research communities in terms of people, data, information, tools, and instruments and that include unprecedented capacity for computational, storage, and communication. Such environments enable teams to share and collaborate over time and over geographic, organizational, and disciplinary distance. They enable individuals working alone to have access to more and better information and facilities for discovery and learning. They can serve individuals, teams and organizations in ways that revolutionize *what they can do*, *how they do it*, and *who participates*.

Figure 2.1 illustrates the types of facilities and services to be provided in an integrated way by a cyberinfrastructure layer (shaded). This layer is built upon base technology for computation, storage, and communication. Cyberinfrastructure should be produced and managed in a way that enables research communities/projects to tailor efficient and effective application-specific, *but interoperable*, knowledge environments for research and education. Interoperability is important for facilitating multidisciplinary projects as the evolution of discovery dictates. The Panel has learned that new types of scientific organizations and supporting environments ("*laboratories without walls*") are essential to the aspirations of growing numbers of research communities/projects and that thus they have begun creating such environments under various names including *collaboratory, co-laboratory, grid community, e-science community,* and *virtual community*. The NSF through an ACP can now enable, encourage, and accelerate this nascent grass-roots revolution in ways that maximize common benefits, minimize redundant and ineffective investments, and avoid increasing barriers to interdisciplinary research.

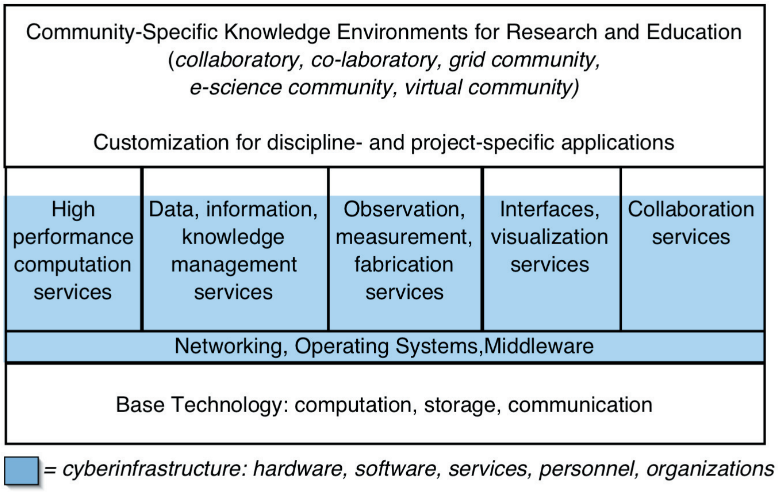
Community-Specific Knowledge Environments for Research and Education
(*collaboratory, co-laboratory, grid community,*
e-science community, virtual community)
Customization for discipline- and project-specific applications
High performance computation services
Data, information, knowledge management services
Observation, measurement, fabrication services
Interfaces, visualization services
Collaboration services
Networking, Operating Systems,Middleware
Base Technology: computation, storage, communication
🟦 = *cyberinfrastructure: hardware, software, services, personnel, organizations*

Figure 2.1 Integrated cyberinfrastructure services to enable new knowledge environments for research and education.

Achieving this vision challenges our fundamental understanding of computer and information science and engineering as well as parts of social science, and it will motivate and drive basic research in these areas. We envision radical improvements in cyberinfrastructure and its impact on all science and engineering over time, as work ripens at the intersection of fundamental technical and social *research* relevant to cyberinfrastructure, as well as the *application* of cyberinfrastructure to discovery and learning. Success in this venture has profound broad implications for research, education, commerce, and the social good.

2.3 Improving Information Technology Performance and Use

The vision of an ACP cannot be achieved by procuring existing commercial technologies alone. Of course, to the extent that commercial technologies and services are available off the shelf, they should be incorporated. But information technology is hardly mature; in fact, it is always evolving toward greater capabilities. Its applications are even less mature, and there are many opportunities to mold it to better meet the needs of end users. While possessing many commonalities with commercial technologies and applications in widespread use, science and engineering research have distinctive needs. These needs can often serve as technology drivers requiring extremes of processing and communication rates, storage capacities, the need for unanticipated access to data by many, and the longevity of data. Thus, research in new information technologies and applications utilizing those technologies often have important commercial spin-offs. This situation is illustrated by supercomputing, first applied to scientific and military applications and later to many commercial purposes.

The NSF mission includes advancing information technologies and their effective application to societal needs through basic and applied research in information technology. The ACP offers a significant opportunity for research into the more effective applications of information technology and opportunities for identifying and refining its supporting cyberinfrastructure. Just as supercomputing and numerical methods have been greatly advanced (and will continue to be advanced) by addressing the needs of the scientific and engineering communities, the ACP will be a significant driver for a diverse suite of technologies including collaborative technologies, massive interoperable distributed databases, digital libraries, and the preservation and mining of data. We expect (and the NSF should encourage) commercial spin-offs from this research, benefiting [21]commercial science and engineering research and development and other application areas.

The conduct of science and engineering is a social activity, pursued by individuals, collaborations, and formal organizations. Any enlightened application of information technology must take into account not only the mission of science and engineering research but also the organizations and processes adopted in seeking these missions. A major opportunity in the ACP is to rethink and redesign these organizations and processes to make best use of information technology. In fact, this is more than an opportunity; it is a requisite for success. Experience has shown that simply automating existing methodologies and practices is not the most effective use of technology; it is necessary to fundamentally rethink how research is conducted in light of new technological capabilities. Advanced cyberinfrastructure offers the potential to conduct new types of research in new ways. Doing this effectively requires holistic attention to mission, organization, processes, and technology. It creates the need to involve social scientists as well as natural scientists and technologists in a joint quest for better ways to conduct research.

2.4 Rationale for Government Investment

The ACP requires government investment in research and development of cyberinfrastructure technologies (principally software) for several reasons. First, the marketplace under invests in long time-horizon research. The cyberinfrastructure and application technologies are within the domain of NSF responsibility for government-funded research, and the ACP will maintain U.S. leadership in these technologies through research, experimentation, and commercialization. Second, infrastructure and applications suffer from a chicken-and-egg conundrum that infrastructure requires a diversity of successful applications for its commercial viability, while commercial applications target only widely deployed infrastructure. This ACP will follow the successful model of the Internet, with targeted and coordinated government investment in infrastructure and applications, experimentation and refinement in actual uses, and coordinated commercialization of both elements together. Third, while we expect many if not most of the technologies developed in this ACP to be of broad applicability, science and engineering research has special needs in functionality, performance, and scale that are unlikely to be fully served by commercial firms, at least not without government assistance.

2.5 Scope of the ACP

We propose a large and concerted new effort, not just a linear extension of the current investment level and resources. NSF must recognize that both the scope and the scale of shared cyberinfrastructure must be far broader and deeper than in the past. Cyberinfrastructure includes computing cycles, but also broadband networking, massive storage, and managed information. Even these are not sufficient. There must be leadership on shared standards, middleware, and basic applications for scientific computation. The individual disciplines must take the lead on defining certain specialized software and hardware configurations, but in a context that encourages them to give back results for the general good of the research enterprise and that facilitates innovative cross-disciplinary activities in the near term and in the distant future.

A major point is that cyberinfrastructure includes more than high - performance computing and connectivity. Not only is it focused on sharing and efficiency and making greater capabilities available across the science and engineering research communities, but it also serves other important goals such as facilitating new applications, allowing applications to interoperate across institutions and disciplines, ensuring that data and software acquired at great expense are preserved for future generations and easily available to all, and empowering enhanced collaboration over distance and across disciplines.

To succeed, NSF must institute a broad and deep program that supports the true needs of all the science and engineering missions within NSF by committing to make the fruits of cyberinfrastructure research and development (as well as related work from other agencies and companies) available in an integrated fashion to facilitate new approaches to scientific and engineering research. It must ensure that the exponentially growing data is collected, curated, managed, and archived for long-term access by scientists (and their IT applications) everywhere, to create and continually renovate a new “high end”, so that selected research projects can use centralized resources 100-1000 times faster and bigger than are available locally. The continuing geometrical improvements in computing speeds and storage and networking capacity mean that research groups and universities now have immediate access to far more resources than ever, but the recent limited national investment in high-end resources constrains the most aggressive research projects from achieving the next level of complexity and resolution.

National needs for advanced cyberinfrastructure will drive significant new research and development in computer and systems architecture. The NSF needs to take advantage of and participate in such efforts to continually improve research cyberinfrastructure; and to support research in areas of computing science that are likely to have largest impact. Science and engineering educators can also use the new infrastructure to educate the next generations of scientists using best techniques, spanning disciplinary boundaries, and democratizing participation. It can enhance international collaboration and resource sharing.

The ACP involves significant educational dimensions in terms of both needs and outcomes. The research community needs more broadly trained personnel with blended expertise in disciplinary science or engineering, mathematical and computational modeling, numerical methods, visualization, and sociotechnical understanding of grid or collaboratory organizations. Grid and collaboratory environments built on cyberinfrastructure can enable people to work routinely with colleagues at distant institutions, even ones that are not traditionally considered research universities, and with junior scientists and students as genuine peers, despite differences in age, experience, race, or physical limitations. These environments can contribute to science and engineering education by providing interesting resources, exciting experiences, and expert mentoring to students, faculty, and teachers anywhere there is access to the Web. The new tools, resources, extensions of human capability, and organizational structures emerging from these activities will eventually have beneficial effect on the future of education at all levels[24] and on knowledge-based institutions more generally.

The ACP also has great potential to empower people who, because of physical capabilities, location, or history, have been excluded from the frontiers of scientific and engineering research and education.

2.6 How Will Science and Engineering Research be Changed?

The vision of ACP is to use cyberinfrastructure to build more ubiquitous, comprehensive digital environments that become interactive and functionally complete for research communities in terms of people, data, information, tools, and instruments and that operate at unprecedented levels of computational, storage, and data transfer capacity. Increasingly, new types of scientific organizations and supporting environments for science are essential, not optional, to the aspirations of research communities and to broadening participation in those communities. They can serve individuals, teams and organizations in ways that revolutionize *what they can do*, *how they do it*, and *who participates*.

Early computational models of physical, mechanical, and biological systems were confined to basic representations of the most fundamental properties and processes. Results from such models, based upon a limited number of calculations painstakingly evaluated, provided new theories and explanations of behaviors either observed in nature or simulated with physical models. Later, increases in computing and networking capabilities bred a new generation of models containing substantially greater realism and the ability to approach scientific and engineering problems from a “systems” point of view. Further, and perhaps more significantly, these models have led to fundamental discoveries.

The ACP is expected to produce another significant step forward in scientific and engineering discovery, not only through investments in raw computing, storage, and networking resources, but also by creating an infrastructure of equipment, software tools, and personnel – appropriately administered – to facilitate the solution of complex, coupled problems involving massive data collection, computation, and analysis. Cyberinfrastructure as we envision it includes not only high-performance computation services, but also integrated services for knowledge management, observation and measurement, visualization, interaction, and collaboration.

While it is impractical and unnecessary to make detailed projections of the impact of ACP on all science and engineering disciplines, a few examples can illustrate how scientific and engineering research will be revolutionized and the benefits that will flow from those changes.

Atmospheric Science – In 1998, the National Research Council[25] noted that, although small- and intermediate-scale climate modeling in the United States is enjoying notable success, the highest-end research opportunities are limited in part by the lack of appropriate computing resources. Not surprising, the highest-end resources are most important for understanding the carbon cycle and other complex processes that govern the global climate system. The ACP will enable the development and execution of fully coupled Earth system models that will allow the simulation of climate for hundreds and thousands of years, down to grid spacing of 10 km, and that will include complete and fully linked representations of chemical, biological, and ecological processes in the atmosphere, hydrosphere, and lithosphere.

At the other end of the time spectrum, today's operational global and hemispheric weather forecast models utilize grid spacing of ~50 km, while limited-area regional/synoptic models operate on grids of 15-20 km spacing. Although such representations of the atmosphere are vastly better than those used even a decade ago, they remain inadequate for capturing nature's most intense and locally disruptive weather. Research now under way in the explicit prediction of individual thunderstorms and their wintertime counterparts, using grid spacing of order 1 kilometer, is showing considerable promise and could have a tremendous impact on aviation, communications, agriculture, and energy. However, the computational challenges are daunting. The ACP will enable research to create effective frameworks for both exploring small-scale atmospheric predictability and dealing with their associated massive amounts of observational data and model output. It will also enable the federation of the necessary multidisciplinary, multi-institutional, and geographically dispersed human expertise, archival data, and computational models.

Forestry – Tremendous progress is being made in the modeling of wildfires, with the explicit inclusion of fuels and chemical reactions and full two-way coupling with the atmosphere. The ACP computational resources will allow for a more complete representation of land-surface characteristics, fuel composition and consumption, and feedbacks, leading to more effective strategies for combating fires including the development of chemical agents whose impacts can be tested empirically, at very large scales, in a virtual world. We can also imagine the emergence of "rapid response collaboratories" that will eventually enable an actual forest fire to be modeled in real time based upon sensor data from the field and used to monitor and direct the process of fire fighting. Running in a predictive mode, these models demonstrate the capability to anticipate a "blowout" event in time to move a fire fighting crew to safety.

Ocean Science – The field of ocean sciences is poised to capitalize upon extraordinary opportunities for advancement, ranging from an understanding of the roles played by the world's oceans in climate and global change to the delicate balances that exist in coastal ecosystems. Computer ocean models now are capable of simulating detailed turbulent structures and transport processes in three dimensions. To understand and predict the full climate system will, for example, require facilities in the ACP capable of computational coupling to atmospheric models, and inclusion of the complex chemistry needed to understand the physics of carbon sequestration. Such efforts require computational, data, and networking resources orders of magnitude beyond those presently available. The coastal zone, fundamentally important to fisheries, defense, recreation, and human health, is a vastly complex environment affected by freshwater runoff, the introduction of large inputs of nitrogen and other nutrients, and the episodic release of pollutants. An accurate representation of these and other biogeochemical processes will allow for better stewardship of the coastal environment and provide frameworks for policy decisions affecting the nation's economy.

Environmental Science and Engineering – The previous three activities and many others are part of a growing collection of interdisciplinary and interorganizational activities in the area of environmental research and education, much of it nurtured by a cross-cutting Environmental Research and Education (ERE)[26] program at the NSF. This community has been among the leaders in exploring requirements for cyberinfrastructure supporting the necessary integration of environmental research and education focused on understanding fundamental processes involved in physical, biological, and human system interactions. Examples include research in the areas of ecosystem dynamics, cell function, atmospheric chemistry, biogeochemical cycles, political or economic institutional processes, coastal ocean processes, population biology and physiological ecology, Earth system history, solar influences, and the study of the interactions responsible for the ozone hole. This is an example of a community for which advanced cyberinfrastructure will have a high payoff.

Space Weather – Although terrestrial weather and climate are of considerable importance to society, space weather – or the conditions in space that arise from interactions between the Earth and sun – is growing rapidly in importance. Active space weather can, for example, disrupt surface and space-based power and communication infrastructures, benefiting not only commerce and the economy but also national defense. To date, the sun and Earth have been studied largely as individual, isolated systems. However, a fully coupled Earth-sun framework is essential for understanding the physics and societal impacts of space weather. This is a truly global research community, and the ACP will create a collaboratory of international science teams, hundreds of ground and space-borne instruments, predictive computational models, and historical data archives. This will improve fundamental understanding and operational space weather forecasting.

Computer Science and Engineering – The foundation of cyberinfrastructure is computer and information science and engineering – areas whose breadth and impact have expanded in the past two decades, and upon which numerous other disciplines depend for efficient and reliable processing, communication, security, management, storage, and visualization. The research challenges are varied, and enable revolutionary science and engineering. Unconventional architectures based upon new substrates (e.g., quantum and biological, including smart fabric and molectronics) offer promise for breaking the silicon CMOS barrier. Self-diagnosing and adaptive systems will be essential for managing the increasingly complex distributed hardware and software infrastructures. We also face challenges in security, scalability, fault tolerance, brokering, scheduling, and policy. Digital libraries, metadata standards, digital classification, and data mining are critical. Additionally, more effective languages, compilers, middleware, and integration – especially in utilizing distributed systems – are a key enabler. An ACP could revolutionize computer science and engineering research itself because, for example, of its inherent complexity and requirements for systemic integration, the opportunity for synergy between creating and applying new knowledge, and the need for a more integrated understanding of the technical and social dimensions of cyberinfrastructure applied to research and education.

Information Science and Digital Libraries – An information-driven digital society requires the collection, storage, organization, sharing, and synthesis of huge volumes of widely disparate information and the digitization of analog sensor data and information about physical objects. The digital library encompasses these functions, and research and development are needed for the infrastructures to mass-manipulate such information on global networks. Digital libraries also provide powerful tools for linking and relating different types of information, leading to new knowledge. These capabilities require new paradigms for information classification, representation (e.g., standards, protocols, formats, languages), manipulation, and visualization. The ACP will spearhead such new developments.

Biology/Bioinformatics – A new era of biology is dawning exemplified by the human genome project and the promise of new science affecting areas such as crop production and personalized medicine. The raw DNA sequence information deposited in public databases doubles every six months or so; its analysis has motivated development of the new field bioinformatics. Characterization of protein folding, for example, utilizing the 30,000 or so protein structures currently available in the public repository, would require hundreds of years on today's desktop computers. While this calculation can be completed in weeks on currently available massively parallel teraflops computers, under the envisioned ACP one can imagine the process being reduced to literally hours. Such work will improve our understanding of myriad biological functions and disease states and provide a framework for developing new therapies and disease and weather/climate resistant plants.

Medicine – Medical advances of great benefit to humanity are expected, ranging from telemedicine and drug therapies to non-invasive repair of damaged tissue. A significant breakthrough will be the creation of a functional, three-dimensional cyber human body. This capability will provide vast educational opportunities ranging from the performance of surgeries on virtual cadavers to physiology education of middle- and high-school students. Much like flight simulators, the virtual human also provides a framework for repeated experimentation under strictly controlled conditions, ranging from macroscale structures (like organs and the musculo-skeletal system) to individual cells. Among other benefits, a virtual human will significantly reduce the imbalance among schools in their facilities for studying human physiology.

Physics – Physics is pursuing major projects depending on advanced cyberinfrastructure. High-energy physics, for example, must have global-scale, high-performance grids and collaboratories to support the acquisition, distribution, storage, and collaborative evaluation of the massive data sets generated by the premiere instruments at CERN. Global scale collaboration will enable experimentation and also designing and constructing facilities and the experiments using them. This community is using cyberinfrastructure to support distributed learning for professional development and to allow faculty to remain active in teaching and mentoring at their home institutions while resident at CERN in Switzerland.

Astronomy – Traditionally, astronomers have analyzed observations of individual targets while assembling theories limited in their consideration of larger-scale interactions. Such individual observations are being replaced by whole-sky surveys of enormous detail and petabyte datasets, providing global views of phenomena ranging from black holes to supernovae, and identifying new objects so rare that only one or two may exist among billions of objects. The needed computational infrastructure does not exist but will be enabled by the ACP. This revolution in astronomy driven by cyberinfrastructure promises to enable a whole new level of understanding of the universe, its constituents, and their origins and evolution, touching on the issues ranging from the fundamental physics in the early universe to the abundance of Earth-like planets and the origins of life.

Engineering – The distinctions between science and engineering are blurring, as illustrated by an engineering component in all the areas in this section. One example of the impact on engineering practice is the understanding of turbulence. Thirty years ago, it was generally believed impossible to perform direct numerical simulation of turbulence (i.e., simulation from first principles, with explicit representation of chaotic motions). Today this has been done, and is revolutionizing the design of combustion engines, aircraft, and automobiles as well as the understanding of clouds and the spread of pollution. However, it is not currently possible to simulate turbulence in large volumes or at high speeds – a significant limitation affecting most of the interesting and relevant applications. Further, the massive datasets produced by direct turbulence simulations are difficult to analyze and visualize, thus thwarting efforts to move from the turbulence produced by a small bird to that produced in the wake of a jumbo jet. The ACP will make available the raw computational, data handling, and visualization resources needed to meet these challenges and thus to improve the manufacturing of large and small devices where turbulence is important.

Materials Science & Engineering – Computer simulations, enabled by the envisioned ACP, will make possible quantum mechanical calculations on nanoscale systems, which, in turn, will enable the fundamental principles governing the rational design of new materials for nanotechnology to be uncovered. Such simulations will, for example, contribute not only to the design but also to the rational synthesis of truly novel materials for IT and national security applications and of nanocatalytic materials for the chemical industry. Extrapolation of what is currently possible with simulations based on classical mechanics and atom-based force fields on current teraflops computers indicates that structural and dynamical properties of trillion-atom systems covering a length scale of a few microns will be possible on a petaflops computer. This will enable the study of systems ranging from nanoscale composite materials with realistic microstructures to biologically inspired self-assembled devices for medical applications. Further, multiscale quantum-atomistic-continuum simulations using the envisioned ACP may enable the integration of thousands of heterogeneous teraflops-scale physical models that will be needed for more fundamental component design and optimization in advanced engineering applications.

Social and Behavioral Sciences – As a relatively new user of cyberinfrastructure, the social and behavioral sciences are poised to make tremendous advances in a variety of areas ranging from cognition and linguistics to economic forecasting. Simulations of the interplay between concepts and perception in the course of analogy making have been created, and programs are under development for modeling the perception and creation of style in the world of letterforms. Devices that convert neural signals to speech are being studied, and new techniques based upon numerical simulation are accelerating the pace of mental and physical rehabilitation, particularly for cases of extreme physical trauma. New virtual organizations and practices made possible by cyberinfrastructure provide new areas of study within the social sciences.

2.7 Participation Beyond the NSF Community

NSF has both a unique breadth of scientific scope and responsibility for the health of the scientific research enterprise in the U.S., so NSF is ideally poised as a leader in cyberinfrastructure within the federal government. However, ACP cannot be fully effective if it is an NSF-only program: significant coordination with other federal agencies, universities, industry, and international programs is required. This will magnify the impact through interoperability and consistency across a larger universe of researchers and will also bring significant added resources to bear.

Other Research Sponsors – The NIH is spending billions of dollars annually on information technology infrastructure and its support and use in research, but in a way that may not lead to a common, interoperable cyberinfrastructure, nor infrastructure at the leading edge. NIH has recently initiated more coordination, in the spirit of an ACP, for example the Biomedical Informatics Research Network (BIRN)[15]. Similarly the Department of Energy (DOE) National Collaboratories Program[16], and the DOE program in Scientific Discovery through Advanced Computing (SciDAC)[17] are examples of growing investment in cyberinfrastructure that can supplement NSF investments.

Industry and Universities – Some of the capabilities needed in ACP are commercially available, industry may be interested in developing new technologies relevant to the science and engineering research community, and many technologies that are an outgrowth of the ACP research and development will be of interest to the commercial sector. Thus industry must be a partner in development and deployment in the ACP and will also be a beneficiary. ACP will also encourage co-investment by universities in advanced cyberinfrastructure on campuses and will provide models and experience with new tools and new organizational forms for knowledge creation and education in the digital age. It could directly complement, for example, a major three-year study now begun at the U.S. National Academies of Science, Engineering, and Medicine on information technology and the future of the research university[24]. It can catalyze and provide over-the-horizon visibility to other agencies, research labs, and education-at-large.

International – It is imperative that the ACP interoperate with cyberinfrastructure being developed and deployed in other countries. Science is international; and many other countries have expertise, data resources, computing systems, applications and systems software, and instruments (such as telescopes and particle accelerators) that need to be available more easily to international teams including American scientists (and vice versa). The high-energy physics community, for example, must have appropriate cyberinfrastructure to enable collaboration in experiments using the premier instruments at CERN in Switzerland.

Collaboration within and among disciplines is growing rapidly; in some cases hundreds of scientists are working on a single project across the globe. Cyberinfrastructure must support this type of collaboration in a reliable, flexible, and cost-effective manner. The activity in Europe and Asia in cyberinfrastructure has increased of late; it is mutually beneficial to established strong links with relevant international efforts and to co-fund significant collaborative international projects. Science is increasingly global, yet it is still difficult to fund joint international e-science projects that develop or require cyberinfrastructure.

Major scientific laboratories elsewhere have contributed significantly to advanced scientific computing, and continue to do so. (The Web was born at CERN, just as the browser was born at NCSA.) For example, the UK National Grid is part of their overall e-science effort, and the Netherlands National Grid has similar goals. The EU is considering a number of even broader Grid proposals.

A few examples of relevant international activities include the following:

- The UK recently launched an “e-Science” program[18] that has many of the characteristics of the ACP. The aims of this program include:

  - provide infrastructure and facilities needed for next major stages of international collaborative research in genomics and bioscience, particle physics, astronomy, earth science & climatology, engineering systems, and the social sciences;

  - contribute to the emergence of next generation open platform standards for global information utilities;

  - solve major challenges in processing, communication, and storage of very large volumes of valuable data;

  - provide *generic* solutions to needs of individual disciplines and applications; and

  - provide optimal international infrastructure.

Initial funding for the e-Science program is on the order of $200 million over three years, most of which is allocated to large applications projects and a quarter of which is devoted to developing the necessary software infrastructure. The latter efforts are collaborating closely with American and European projects that are developing middleware and in some cases even providing funding for those international groups. The e-Science funds are supplemented by infrastructure funding from previously existing programs that support both a very capable UK-wide research network (10 Gb/s backbone) and high-speed international links and high end computing resources. On the latter topic, in July the UK Science Research Council signed a contract with IBM with an overall cost of £53m (over $82M) for a computer system known as HPC(X).   The initial 3 teraflops configuration of HPC(X) will ramp up to 12 teraflops by 2006, with a teraflops rating based on LINPACK performance.

- The European Union has funded well over a dozen Grid projects as well as a high-speed European research network – GEANT[27]. GEANT reaches over 3,000 research and education institutions in 30-plus countries through 28 national and regional research and education networks. It is also quite fast: nine of its circuits operate at speeds of 10 Gbps, while eleven others run at 2.5 Gbps. GEANT has the dual roles of providing an infrastructure to support research in application domains and providing an infrastructure for (network) research itself.

- In the upcoming Sixth Framework Program[19], the EU has allocated 300M euros for further upgrading the GEANT network and for building large-scale Grid test-beds. A solicitation for proposals will be issued in the first half of 2003. In addition, there are a number of grid projects under way funded by individual countries. A partial list includes Canada, China, Denmark, India, Japan, Korea, Norway, Romania, Sweden, and Switzerland. Typical funding levels are tens of millions of dollars per project over several years.

- Other countries also have significant computing resources that are used for computational science. In the early 1980s U.S. academic researchers gained access to European computing facilities enabling larger-scale computational science research. In Japan, many universities and research laboratories have high-end facilities, and in March 2002 the Earth Simulator[20] system became operational. The Earth Simulator, currently the world's fastest computer system with a peak speed of 40 teraflops, was built by NEC for the Earth Simulator Research and Development Center, a collaborative organization of the National Space Development Agency of Japan, Japan Atomic Energy Research Institute, and Japan Marine Science and Technology Center. The Earth Simulator is targeted at analysis of global environmental problems through simulation of geophysical, climate, and weather-related phenomena.

At present 55% of the top 500 computer systems in the world (based on LINPACK ratings), representing 56% of the aggregate LINPACK flops, are outside the US. Of the top 100 systems, 33 are designated for academic use. Of those, only 9 are in the US, even if one includes the systems at NCAR[28] and at NERSC[29]. In areas such as high-end computing and high-speed network infrastructure, other countries are either in the lead or on a par with the US. However, late in 2002 Lawrence Livermore National Laboratory announced an order to IBM for delivery of a 100 teraflops machine in 2004 (for national security calculations) and delivery of a 360 teraflops machine in 2005 (mostly for open scientific applications). Many scientific investigations have international components and therefore ACP should make both U.S. and international resources available for shared international collaboration.

2.8 Educational Needs and Impact

A new interdisciplinary work force – The need for a new workforce – a new flavor of mixed science and technology professional – is emerging. These individuals have expertise in a particular domain science area, as well as considerable expertise in computer science and mathematics. Also needed in this interdisciplinary mix are professionals who are trained to understand and address the human factors dimensions of working across disciplines, cultures, and institutions using technology-mediated collaborative tools. Prior work on computer-supported collaborative work and social dimensions of collaboratories needs to be better codified, disseminated, and applied in the design and refinement of new knowledge environments for science based on cyberinfrastructure.

The term “computational science and engineering” (CSE) has emerged as a descriptor of broad multidisciplinary study that encompasses applications in science/engineering, applied mathematics, numerical analysis, and computer science. As noted by the Society for Industrial and Applied Mathematics (SIAM)[30]:

Computer models and computer simulations have become an important part of the research repertoire, supplementing (and in some cases replacing) experimentation. Going from application area to computational results requires domain expertise, mathematical modeling, numerical analysis, algorithm development, software implementation, program execution, analysis, validation and visualization of results. CSE involves all of this.

SIAM notes that “CSE is a legitimate and important academic enterprise even if it has yet to be formally recognized as such at some institutions. Although it includes elements from computer science, applied mathematics, engineering and science, CSE focuses on the integration of knowledge and methodologies from all of these disciplines, and as such is a subject which is distinct from any of them.”

The community surveyed in creating this report noted repeatedly that insufficient attention is being given to educating non-computer or domain science students in the concepts and tools of cyberinfrastructure. For example, graduate and higher-level undergraduate courses in computer science are designed for disciplinary majors, and non-majors wishing to take such courses are dissuaded by an onerous prerequisite structure. Further, even if such skills are attained, domain science courses often do not exercise them sufficiently, leading to atrophy of skills.

In response to this problem – while also recognizing the need to maintain strong, traditional disciplinary programs in science and engineering research – significant resources must be directed toward developing programs of study in the computational sciences at both the graduate and undergraduate levels. A survey of educational objectives, as well as sample programs and curricula, can be found at the SIAM Web site.[30]

Continuing education is also needed. Community-wide workshops are needed for science and engineering practitioners so as to function effectively in the rapidly evolving IT world. Such workshops and courses could be delivered via distance learning and would lower the entry threshold for those new to high-performance computation.

Impact on science and engineering education – The ACP requires the aforementioned innovation and reforms in education and can also be directly leveraged in science and engineering education. Grid and collaboratory environments built on cyberinfrastructure can enable people to work routinely with colleagues at distant institutions, even ones that are not traditionally considered research universities, and with junior scientists and students as genuine peers, despite differences in age, experience, race, or physical ability. These new environments can contribute to science and engineering education by providing interesting resources, exciting experiences, and expert mentoring to students, faculty, and teachers anywhere. By making access to reports, raw data, and instruments much easier, a far wider audience can be served. Since broadband networks are increasingly available in schools, videos and other complex effects can be viewed by students and teachers as well as by researchers.The new tools, resources, human capacity building, and organizational structures emerging from these activities will also eventually have even broader beneficial impact on the future of education at all levels, in almost all disciplines, and in all types of educational institutions.

Minority Serving Institutions – An important goal of the ACP must be to more effectively include Minority Serving Institutions (MSIs), which include Historically Black Colleges and Universities (HBCU), American Indian Tribal Colleges (AIT), and Hispanic Colleges and Universities (HCUs) and other underrepresented groups into mainstream scientific and engineering research and education. Few of these institutions were involved in discussions leading to the original NSF supercomputing centers, and collaboration efforts to date, though well intentioned and covering a spectrum of activities ranging from education/outreach/training to basic research, have for the most part fallen short of their goals for a variety of reasons. This failure is particularly troubling in light of the fact that, by 2035, it is estimated that one in five Americans will be Hispanic.

One of the most important barriers to engaging MSIs in research using cyberinfrastructure is the lack of adequate network connectivity – a problem especially acute for the Tribal Colleges because of their largely rural location and frequently impoverished localities (three of the five poorest counties in the United States are homes to Tribal Colleges). Further, such institutions lack the tools and infrastructure needed to participate in mainstream research. Although various initiatives (e.g., EOT-PACI[31], the Advanced Network with Minority Serving Institutions Initiative[32]) have shown promise, the principal audience has been IT staff rather than faculty and researchers. These and other limitations have perpetuated the so-called digital divide, reflected by a 20+ year gap in capability between mainstream institutions and many MSIs (based on statistics from the U.S. Department of Commerce, 46.1% of white non-Hispanic households have access to the Internet, compared with 23.6% for Hispanics).

Although this challenge is multifaceted, solutions need not be incrementally applied; indeed, it is eminently possible, through a significant infusion of both technology and education, to close the digital divide and establish meaningful research collaborations and educational initiatives. The PITAC[33] emphasized the importance of reaching MSIs, and we underscore it again here. The ACP therefore must support strategic IT planning for underserved communities.
In addition, opportunities for research collaboration must be more effectively communicated to both mainstream institutions and MSIs, and significant efforts must be directed toward engaging underserved communities directly, rather than as programmatic add-ons.

Experimental Program to Stimulate Competitive Research – A more encouraging story can be told about EPSCoR[34] (Experimental Program to Stimulate Competitive Research), which at present involves 21 states and the Commonwealth of Puerto Rico. A joint program of the NSF and several U.S. states and territories, EPSCoR promotes the development of science and technology resources through partnerships involving universities, industry, government, and the federal research and development enterprise. It operates on the principle that aiding researchers and institutions in securing federal R&D funding will develop a state's research infrastructure and advance economic growth, and its main goal is to maximize the potential inherent in a state's science and technology resources and use those resources as a foundation for economic growth.

Most EPSCoR states have taken significant steps to provide high-speed connectivity and engage researchers in collaborative cyber activities with major universities and national centers and laboratories, and these efforts should be continued and expanded. For example, the University of Kentucky spearheaded a project through which scientists and researchers in EPSCoR states can use Access Grid (AG) technology to bridge the digital divide caused by their geographic dispersion and limited funding. Six EPSCoR-grant states are implementing AG nodes, and the two newest EPSCoR states, Hawaii and New Mexico, have nodes as a result of their participation in the National Computational Science Alliance.

EPSCoR co-funding of mainline research grants, particularly in the Information Technology Research Program, has had a significant positive impact on research competitiveness of participating institutions (see http://www.ehr.nsf.gov/epscor/start.cfm). In the NSF Geosciences Directorate alone, EPSCoR co-funding has increased by a factor of 3 in the past few years. The ACP should embrace EPSCoR and continue to support what clearly is a very successful, high-impact program. Indeed, the EPSCoR model could be applied more specifically to MSIs, particularly with regard to high performance network connectivity.

Access by the wider public – By making access to reports, raw data, and instruments much easier, a far wider audience can be served. Although large teams and major financial investment are required to create comprehensive data repositories and specialized scientific facilities, individuals, even amateurs, working alone or in small groups, given access to such resources, can provide scientific discoveries. A good example is amateur astronomy, which significantly expands the reach of scientific observation.

Participation by the physically challenged – There are many ways to assist scientists and other users who have physical constraints through advanced cyberinfrastructure, as long as this opportunity is addressed from the beginning. Most of these resources are likely to be provided close to the individual rather than in a shared environment. Many of these supportive pieces of hardware and software will be generic, but there may be some tools specific to the scientific milieu. We have also identified a few functions that could most appropriately be implemented centrally. A few examples that should be considered for implementation within the ACP will illustrate this. Even people who are not challenged will still find some of these features of use – a common observation about assistive technologies, the so-called curb-cut effect (sidewalks with curb cuts are simply better sidewalks – they help bikers, skaters, and people pushing strollers – not just those confined to wheel chairs).

It takes massive computing to do a first-class conversion of speech to text. (Online and moderate accuracy conversion can be done by commercial software on a typical PC, but higher accuracy requires elaborate algorithms that repeatedly examine delayed inputs.) Such computing might be provided as a Grid service on shared multiprocessors and would make an excellent adjunct for collaborative environments such as the Access Grid[35]. By using a shared networked resource, the service would be available to hearing-impaired scientists wherever they are. (The service would also be valued for making seminars available for delayed use by everybody.)

Infrastructure services that can convert sounds to visual signals would help the hearing-impaired interact with experimental equipment. The inverse translation of control panels to sounds would be useful for the visually impaired. This specialized translation does not fit simply into the commercial Webpage-enablement paradigm and may be particularly important for control gauges and warning devices. There could be broad social benefit to providing standards and support software for infrastructure-connected apparatus. (Sighted people might benefit from audible alarms, and workers in crowded environments might prefer silent visual signals, so standardized conversions for laboratory equipment may find broader usage.)

A research challenge would be to extend "visualization" to provide information for the visually impaired. Tactile (haptic) exploration combined with audible signals may be a useful way to convey information about complex phenomena and mathematical surfaces. If successful approaches are found, they should be made available through the ACP.

Digital libraries, discipline-specific collections, and archives of the published literature will be key components of the cyberinfrastructure. It is difficult for people with motor or visual disabilities to point to many specific items, or to look at very long stretches of text. A variety of services would help them and would also speed the work of others. Some possible approaches are abstracting services and interfaces that encourage skipping or shifting focus.

2.10 Overall Finding and Recommendation

The Panel's overarching finding is that a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology; and pulled by the expanding complexity, scope, and scale of today's research challenges. The capacity of this technology has crossed thresholds that now make possible a comprehensive "cyberinfrastructure" on which to build new types of scientific and engineering knowledge environments and organizations and to pursue research in new ways and with increased efficacy. The cost of not doing this is high, both in opportunities lost and through increasing fragmentation and balkanization of the research communities.

Such environments and organizations, enabled by cyberinfrastructure, are increasingly required to address national and global priorities such as understanding global climate change, protecting our natural environment, applying genomics-proteomics to human health, maintaining national security, mastering the world of nanotechnology, and predicting and protecting against natural and human disasters, as well as to address some of our most fundamental intellectual questions such as the early formation of the universe and the fundamental character of matter.

As will be discussed in Section 5, there is already a significant base of effort and capability in the PACIs, which were created in response to the Hayes Report[36]. They run computing and data centers, create important middleware and scientific software, and coordinate activities with other scientists. Subject to appropriate review, we anticipate that they will play a continuing but evolving substantial role in the greatly enlarged activity we propose.

The Panel's overarching recommendation is that the National Science Foundation should establish and lead a large-scale, interagency, and internationally coordinated Advanced Cyberinfrastructure Program (ACP) to create, deploy, and apply cyberinfrastructure in ways that radically empower all scientific and engineering research and allied education. We estimate (details in Section 6) that sustained new NSF funding of $1billion per year is required to achieve critical mass and to leverage the necessary coordinated co-investment from other federal agencies, universities, industry, and international sources required to empower a revolution.

This Panel believes that the National Science Foundation has a once-in-a-generation opportunity to lead the revolution in science and engineering through coordinated development and expansive use of cyberinfrastructure.

The following sections of this report provide a further basis for this recommendation, our estimate of new funding required, and principles for the organization and management of the program. Appendixes provide additional details.