NSF'S CYBERINFRASTRUCTURE VISION FOR 21ST CENTURY DISCOVERY

NSF Cyberinfrastructure Council

National Science Foundation
January 20, 2006
Version 5.0

ACRONYMS

CCSDS	Consultative Committee for Space Data Standards
CI	Cyberinfrastructure
CODATA	National Committee on Data for Science and Technology
CPU	Central Processing Unit
CSNET	Computer Science Network
DARPA	Defense Advanced Research Projects Agency
DOD	Department of Defense
DOE	Department of Energy
ETF	Extensible Terascale Facility
FLOPS	Floating point operations/sec
HPC	High Performance Computing
HPCMOD	DOD's High-Performance Computing Modernization program
HPCS	DARPA's High Productivity Computing Systems program
HPCC	High-Performance Computing and Communications
GEON	Geosciences Network
GriPhyN	Grid Physics Network
ICPSR	Inter-university Consortium for Political and Social Research
ICSU	International Council for Science
ICTSI	International Council for Scientific and Technical Information
IRIS	Incorporated Research Institutions for Seismology
ISO	International Organization of Standardization
IT	Information Technology
ITR	Information Technology Research
IVDGL	International Virtual Data Grid Laboratory
MPI	Message Passing Interface
NARA	National Archives and Record Administration
NASA	National Aeronautics and Space Administration
NIH	National Institutes of Health
NITRD	Networking and Information Technology Research and Development
NNSA	National Nuclear Security Administration
NSFNET	NSF Network
NRC	National Research Council
NSB	National Science Board
NSF	National Science Foundation
OAIS	Open Archival Information System
OS	Operating System
PACI	Partnership for Advanced Computational Infrastructure
PITAC	President's Information Technology Advisory Committee
RLG	Research Library Group
SciDAC	Scientific Discovery through Advanced Computing
SSP	Software Services Provider
TFLOPS	Teraflops: Trillion floating point operations/sec
WDC	World Data Center

Acronyms

TABLE OF CONTENTS

CHAPTER 1 Call to Action……..…….4

I. Drivers and Opportunities ……..……..……..……..4

II. Vision, Mission, and Principles ……..…….. 5

III. Goals and Strategies ……..…… ……..…….. 7

IV. Strategic Planning for Cyberinfrastructure Components…… ….9

CHAPTER 2 Strategic Plan for High Performance Computing (2006-2010)……..……10

I. What Does High Performance Computing Offer Science and Engineering? ..10

II. The Next Five Years: Creating a High Performance Computing Environment for Petascale Science and Engineering ..……..11

CHAPTER 3 Strategic Plan for Data, Data Analysis and Visualization (2006-2010)… …...16

I. A Wealth of Scientific Opportunities Afforded by Digital Data …… ……. …..16

II. Definitions ……..…… ……..…….. …….. …..17

III. Developing a Data Cyberinfrastructure in a Complex, Global Context ….. ….17

IV. Plan of Action ……..…… ……..…….. …….. .19

V. Conclusion…… …… 24

CHAPTER 4 Strategic Plan for Collaboratories, Observatories and Virtual Organizations (2006-2010) ……25

CHAPTER 5 Strategic Plan for Education and Workforce (2006-2010) ……..…. ……26

Appendix A Representative Reports and Workshops ……..……..27

Appendix B Chronology of NSF IT Investments……..30

Appendix C Management of Cyberinfrastructure…… ……..32

CHAPTER 1
CALL TO ACTION

I. DRIVERS AND OPPORTUNITIES

How does a protein fold? What happens to space-time when two black holes collide? What impact does species gene flow have on an ecological community? What are the key factors that drive climate change? Did one of the trillions of collisions at the Large Hadron Collider produce a Higgs boson, the dark matter particle or a black hole? Can we create an individualized model of each human being for targeted healthcare delivery? How does major technological change affect human behavior and structure complex social relationships? What answers will we find – to questions we have yet to ask – in the very large datasets that are being produced by telescopes, sensor networks, and other experimental facilities?

These questions – and many others – are only now coming within our ability to answer because of advances in computing and related information technology. Once used by a handful of elite researchers in a few research communities on select problems, advanced computing has become essential to future progress across the frontier of science and engineering. Coupled with continuing improvements in microprocessor speeds, converging advances in networking, software, visualization, data systems and collaboration platforms are changing the way research and education is accomplished.

Today's scientists and engineers need access to new information technology capabilities, such as distributed wired and wireless observing network complexes, and sophisticated simulation tools that permit exploration of phenomena that can never be observed or replicated by experiment. Computation offers new models of behavior and modes of scientific discovery that greatly extend the limited range of models that can be produced with mathematics alone, for example, chaotic behavior. Fewer and fewer researchers working at the frontiers of knowledge can carry out their work without cyberinfrastructure of one form or another.

While hardware performance has been growing exponentially – with gate density doubling every 18 months, storage capacity every 12 months, and network capability every 9 months – it has become clear that increasingly capable hardware is not the only requirement for computation-enabled discovery. Sophisticated software, visualization tools, middleware and scientific applications created and used by interdisciplinary teams are critical to turning flops, bytes and bits into scientific breakthroughs. It is the combined power of these capabilities that is necessary to advance the frontiers of science and engineering, to make seemingly intractable problems solvable and to pose profound new scientific questions.

The comprehensive infrastructure needed to capitalize on dramatic advances in information technology has been termed cyberinfrastructure. Cyberinfrastructure integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools. Investments in interdisciplinary teams and cyberinfrastructure professionals with expertise in

algorithm development, system operations, and applications development are also essential to exploit the full power of cyberinfrastructure to create, disseminate, and preserve scientific data, information, and knowledge.

For four decades, NSF has provided leadership in the scientific revolution made possible by information technology (Appendices A and B). Through investments ranging from supercomputing centers and the Internet to software and algorithm development, information technology has stimulated scientific breakthroughs across all science and engineering fields. Most recently, NSF's Information Technology Research (ITR) priority area sowed the seeds of broad and intensive collaboration between the computational, computer and domain research communities that sets the stage for this "Call to Action."

NSF is the only agency within the U.S. government that funds research and education across all disciplines of science and engineering. Over the past five years, NSF has held community workshops, commissioned blue-ribbon panels and carried out extensive internal planning (Appendix A.) Thus, it is strategically placed to leverage, coordinate and transition cyberinfrastructure advances in one field to all fields of research.

Other Federal agencies, the Administration and Congress, the private sector, and other nations are aware of the growing importance of cyberinfrastructure to progress in science and engineering. Other Federal agencies have planned improved capabilities for specific disciplines, and in some cases to address interdisciplinary challenges. Other countries have also been making significant progress in scientific cyberinfrastructure. Thus, the U.S. must engage in and actively benefit from cyberinfrastructure developments around the world.

Not only is the time ripe for a coordinated investment in cyberinfrastructure, progress at the science and engineering frontiers depends upon it. Our communities are in place and are poised to respond to such an investment.

Working with the science and engineering research and education communities and partnering with other key stakeholders, NSF is ready to lead.

II. VISION, MISSION, AND PRINCIPLES

A. Vision

NSF will play a leadership role in the development and support of a comprehensive cyberinfrastructure essential to 21st century advances in science and engineering research and education.

B. Mission

NSF's mission for cyberinfrastructure (CI) is to:

- Develop a human-centered CI that is driven by science and engineering research and education opportunities;

- Provide the science and engineering communities with access to world-class CI tools and services, including those focused on: high performance computing; data, data analysis and visualization; collaboratories, observatories and virtual organizations; and, education and workforce development;
- Promote a CI that serves as an agent for broadening participation and strengthening the Nation's workforce in all areas of science and engineering;
- Provide a sustainable CI that is secure, efficient, reliable, accessible, usable, and interoperable, and which evolves as an essential national infrastructure for conducting science and engineering research and education; and
- Create a stable CI environment that enables the research and education communities to contribute to the agency's statutory mission.

C. Principles

The following principles will guide NSF's actions.

- Science and engineering research and education are foundational drivers of CI.
- NSF has a unique leadership role in formulating and implementing a national CI agenda focused on advancing science and engineering.
- Inclusive strategic planning is required to effectively address CI needs across a broad spectrum of organizations, institutions, communities and individuals, with input to the process provided through public comments, workshops, funded studies, advisory committees, merit review and open competitions.
- Strategic investments in CI resources and services are essential to continued U.S. leadership in science and engineering.
- The integration and sharing of cyberinfrastructure assets deployed and supported at national, regional, local, community, and campus levels represent the most effective way of constructing a comprehensive CI ecosystem suited to meeting future needs.
- National and international partnerships, public and private, that integrate CI users and providers and benefit NSF's research and education communities are also essential for enabling next-generation science and engineering.
- Existing strengths, including research programs and CI facilities, serve as a foundation upon which to build a CI designed to meet the needs of the broad science and engineering community.
- Merit review is essential for ensuring that the best ideas are pursued in all areas of CI funding.
- Regular evaluation and assessment tailored to individual projects is essential for ensuring accountability to all stakeholders.
- A collaborative CI governance structure that includes representatives who contribute to basic CI research, development and deployment, as well as those who use CI, is essential to ensure that CI is responsive to community needs and empowers research at the frontier.

III. GOALS AND STRATEGIES

NSF's vision and mission statements need well-defined goals and strategies to turn them into reality. The goals underlying these statements are provided below, with each goal followed by a brief description of the strategy to achieve the goal.

Across the CI landscape, NSF will:

- ***Provide communities addressing the most computationally challenging problems with access to a world-class high performance computing (HPC) environment through NSF acquisition and through exchange-of-service agreements with other entities, where possible.***

NSF's investment strategy in the provision of HPC resources and services will be linked to careful requirements analyses of the computational needs of research and education communities. Our investments will be coordinated with those of other agencies in order to maximize access to these capabilities and to provide a range of representative high performance architectures.

- ***Broaden access to state-of-the-art computing resources, focusing especially on institutions with less capability and communities where computational science is an emerging activity.***

Building on the achievements of current CI service providers and other NSF investments, the agency will work to make necessary computing resources more broadly available, paying particular attention to emerging and underserved communities.

- ***Support the development and maintenance of robust systems software, programming tools, and applications needed to close the growing gap between peak performance and sustained performance on actual research codes, and to make the use of HPC systems, as well as novel architectures, easier and more accessible.***

NSF will build on research in computer science and other research areas to provide science and engineering applications and problem-solving environments that more effectively exploit innovative architectures and large-scale computing systems. NSF will continue and build upon its existing collaborations with other agencies in support of the development of HPC software and tools.

- ***Support the continued development, expansion, hardening and maintenance of middleware that permits the integration and sharing of digital resources by communities of researchers and educators, as well as new middleware and applications needed to take advantage of advances in connectivity made possible by Internet2, the National Lambda Rail and other emerging networks.***

These investments will build on the middleware products of current and former programs, and will leverage work in core computer science research and development efforts supported by NSF and other federal agencies.

- ***Support the development of the computing professionals, interdisciplinary teams and new organizational structures, such as virtual communities, needed to achieve the scientific breakthroughs made possible by advanced CI, paying particular attention to the opportunities to broaden the participation of underrepresented groups.***

NSF will continue to invest in understanding how participants in its research and education communities, as well as the scientific workforce, can use CI. For example, virtual organizations empower communities of users to interact, exchange information and access and share resources through tailored interfaces. Some of NSF's investments will focus on appropriate mechanisms or structures for use, while others will focus on how best to train future users of CI. NSF will take advantage of the emerging communities associated with CI that provide unique and special opportunities for broadening participation in the science and engineering enterprise.

- ***Support state-of-the-art innovation in data management and distribution systems, including digital libraries and educational environments that are expected to contribute to many of the scientific breakthroughs of the 21st century.***

NSF will foster communication between forefront data management and distribution systems, digital libraries and other education environments sponsored in its various directorates. NSF will ensure that its efforts take advantage of innovation in large data management and distribution activities sponsored by other agencies and international efforts as well. These developments will play a critical role in decisions that NSF makes about long-lived data.

- ***Support the design and development of the CI needed to realize the full scientific potential of NSF's investments in tools and large facilities, from observatories and accelerators to sensor networks and remote observing systems.***

NSF's large facilities and other tools investments require new types of CI such as wireless control of networks of sensors in hostile environments, rapid distribution of petascale data sets around the world, adaptive knowledge-based control and sampling systems, and innovative visualization systems for collaboration. NSF will ensure that these projects invest appropriately in CI capabilities and that its CI programs serve the needs of these projects.

- ***Support the development and maintenance of the increasingly sophisticated applications needed to achieve the scientific goals of research and education communities.***

The applications needed to produce cutting-edge science and engineering have become increasingly complex. They require teams, even communities, to develop and sustain wide and long-term applicability. NSF's investments in applications will involve its directorates, which support domain-specific science and engineering. Special attention will be paid to the cross-disciplinary nature of much of the work.

- ***Invest in the high-risk/high-gain basic research in computer science, computing and storage devices, mathematical algorithms and the human/CI interfaces that are critical to powering the future exponential growth in all aspects of computing, from hardware speed, storage, connectivity and scientific productivity.***

NSF's investments in operational CI must be coupled with vigorous research programs in the directorates that will ensure operational capabilities continue to expand and extend in the future. Important among these are activities to understand how humans adopt and use CI. NSF is especially placed to foster collaborations among computer scientists, social, behavioral and economic scientists, and other domain scientists and engineers to understand how humans can best use CI, both in research and education environments.

- ***Provide a framework that will sustain reliable, stable resources and enable the integration of new technologies and research developments with a minimum of disruption to users.***

NSF will minimize disruption to users by means of pre-planned arrangements for alternative CI availabilities during competitions, changeovers and upgrades to production operations and services, cooperative arrangements with other agencies, and additional evolving flexibility afforded by the realization of a comprehensive cyberinfrastructure created in an interoperable, open architecture format.

A strategy common to achieving all of these goals is partnering nationally and internationally, with other agencies, the private sector, and with universities to achieve a worldwide CI that is interoperable, flexible, efficient, evolving and broadly accessible. In particular, NSF will take a lead role in formulating and implementing a national CI strategy.

IV. STRATEGIC PLANNING FOR CYBERINFRASTRUCTURE COMPONENTS

To implement its cyberinfrastructure vision, NSF will develop detailed strategic plans for each of the following CI components:

- High Performance Computing;
- Data, Data Analysis, and Visualization;
- Collaboratories, Observatories, and Virtual Organizations; and,
- Education and Workforce Development.

Others may be added at a later date.

The strategies will be reviewed annually and will evolve over time, paced by the considerable rate of innovation in computing and the growing needs of the science and engineering community for state-of-the-art CI capabilities. Through their simultaneous implementation, NSF's vision will become reality.

++

CHAPTER 2
STRATEGIC PLAN FOR HIGH PERFORMANCE COMPUTING (2006-2010)

I. WHAT DOES HIGH PERFORMANCE COMPUTING OFFER SCIENCE AND ENGINEERING?

What are the three-dimensional structures of all of the proteins encoded by the human genome and how does structure influence their function in a human cell? What patterns of emergent behavior occur in models of very large societies? How do massive stars explode and produce the heaviest elements in the periodic table? What sort of abrupt transitions can occur in Earth's climate and ecosystem structure? How do these occur and under what circumstances? If we could design catalysts atom-by-atom, could we transform industrial synthesis? What strategies might be developed to optimize management of complex infrastructure systems? What kind of language processing can occur in large assemblages of neurons? Can we enable integrated planning and response to natural and man-made disasters that prevent or minimize the loss of life and property? These are just some of the important questions that researchers wish to answer using contemporary tools in a state-of-the-art High Performance Computing (HPC) environment.

With HPC tools, researchers study the properties of minerals at the extreme temperatures and pressures that occur deep within the Earth. They simulate the development of structure in the early Universe. They probe the structure of novel phases of matter such as the quark-gluon plasma. HPC capabilities enable the modeling of life cycles that capture interdependencies across diverse disciplines and multiple scales to create globally competitive manufacturing enterprise systems. And they examine the way proteins fold and vibrate after they are synthesized inside an organism. In fact, sophisticated numerical simulations permit scientists and engineers to perform a wide range of *in silico* experiments that would otherwise be too difficult, too expensive or impossible to perform in the laboratory.

HPC systems and services are also essential to the success of research conducted with sophisticated experimental tools. For example, without the waveforms produced by numerical simulation of black hole collisions and other astrophysical events, gravitational wave signals cannot be extracted from the data produced by the Laser Interferometer Gravitational Wave Observatory; high-resolution seismic inversions from the higher density of broad-band seismic observations furnished by the Earthscope project are necessary to determine shallow and deep Earth structure; simultaneous integrated computational and experimental testing is conducted on the Network for Earthquake Engineering Simulation to improve seismic design of buildings and bridges; and HPC is essential to extracting the signature of the Higgs boson and supersymmetric particles – two of the scientific drivers of the Large Hadron Collider – from the petabytes of data produced in the trillions of particle collisions.

Science and engineering research and education enabled by state-of-the-art HPC tools have a direct bearing on the Nation's competitiveness. If investments in HPC are to have long-term

impact on problems of national need, such as bioengineering, critical infrastructure protection (for example, the electric power grid), health care, manufacturing, nanotechnology, energy, and transportation, then HPC tools must deliver high performance capability to a wide range of science and engineering applications.

II. THE NEXT FIVE YEARS: CREATING A HIGH PERFORMANCE COMPUTING ENVIRONMENT FOR PETASCALE SCIENCE AND ENGINEERING

NSF's five-year HPC goal is to enable *petascale* science and engineering through the deployment and support of a world-class HPC environment comprising the most capable combination of HPC assets available to the academic community. The petascale HPC environment will enable investigations of computationally challenging problems that require computers operating at sustained speeds on actual research codes of 10^{15} floating point operations per second (petaflops) or that work with extremely large data sets on the order of 10^{15} bytes (petabytes).

Petascale HPC capabilities will permit researchers to perform simulations that are intrinsically multi-scale or that involve multiple simultaneous reactions, such as modeling the interplay between genes, microbes, and microbial communities and simulating the interactions between the ocean, atmosphere, cryosphere and biosphere in Earth systems models. In addition to addressing the most computationally challenging demands of science and engineering, new and improved HPC software services will make supercomputing platforms supported by NSF and other partner organizations more efficient, more accessible, and easier to use.

NSF will support the deployment of a well-engineered, scalable, HPC infrastructure designed to evolve as science and engineering research needs change. It will include a sufficient level of diversity, both in architecture and scale of deployed HPC systems, to realize the research and education goals of the broad science and engineering community.

The following principles will guide the agency's FY 2006 through FY 2010 investments.

- Science and engineering research and education priorities will drive HPC investments.
- Collaborative activities involving science and engineering researchers and private sector organizations are needed to ensure that HPC systems and services are optimally configured to support petascale scientific computing.
- Researchers and educators require access to reliable, robust, production-quality HPC resources and services.
- HPC-related research and development advances generated in the public and private sectors, both domestic and foreign, must be leveraged to enrich HPC capabilities.
- The development, implementation and annual update of an effective multi-year HPC. strategy is crucial to the timely introduction of research and development outcomes and innovations in HPC systems, software and services.

NSF's implementation plan to create a petascale environment includes the following three interrelated components:

1). Specification, Acquisition, Deployment and Operation of Science-Driven HPC Systems Architectures

An effective computing environment designed to meet the computational needs of a range of science and engineering applications will include a variety of computing systems with complementary performance capabilities. By 2010, the petascale computing environment available to the academic science and engineering community is likely to consist of: (i) a significant number of systems with peak performance in the 1-50 teraflops range, deployed and supported at the local level by individual campuses and other research organizations; (ii) multiple systems with peak performance of 100+ teraflops that support the work of thousands of researchers nationally; and, (iii) at least one system in the 1-10 petaflops range that supports a more limited number of projects demanding the highest levels of computing performance. All NSF-deployed systems will be appropriately balanced and will include core computational hardware, local storage of sufficient capacity, and appropriate data analysis and visualization capabilities. Chapters 3 and 4 in this document describe the complementary investments necessary to provide effective data analysis and visualization capabilities, and to integrate HPC resources into a comprehensive national CI environment to improve both accessibility and usability.

Over the FY 2006-2010 period, NSF will focus on HPC system acquisitions in the 100 teraflops to 10 petaflops range, where strategic investments on a national scale are necessary to ensure international leadership in science and engineering. Since different science and engineering codes may achieve optimal performance on different HPC architectures, it is likely that by 2010 the NSF-supported HPC environment will include both loosely-coupled and tightly coupled systems, with several different memory models.

To address the challenge of providing the research community with access to a range of HPC architectures within a constrained budget, a key element of NSF's strategy is to participate in resource-sharing with other federal agencies. A strengthened interagency partnership will focus, to the extent practicable, on ensuring shared access to federal leadership-class resources with different architectures, and on the coordination of investments in HPC system acquisition and operation. The Department of Energy's Office of Science and National Nuclear Security Administration have very active programs in leadership computing. The Department of Defense's (DOD's) High Performance Computing Modernization Office (HPCMOD) provisions HPC resources and services for the DOD science and engineering community, while NASA is deploying significant computing systems also of interest to NSF PIs. To capitalize on these common interests, NSF will work toward the creation of a Leadership Computing Council as proposed by Simon et al.[1], to include representatives from all federal agencies with a stake in science and engineering-focused HPC. As conceived, the Leadership Computing Council will make coordinated and collaborative investments in science-driven hardware architectures, will increase the diversity of architectures of leadership class systems available to researchers and educators around the country, will promote sharing of lessons learned, and will provide a richer HPC environment for the user communities supported by each agency.

Strong partnerships involving universities, industry and government are also critical to success. In addition to leveraging the promise of Phase III of the DARPA-sponsored High Productivity

[1] Simon *et al.,* "Science-Driven System Architecture: A New Process for Leadership Class Computing," Journal of the Earth Simulator, pages 1-9, Vol. 2, January 2005.

Computing Systems (HPCS) program[2] in which NSF is a mission partner, the agency will establish a discussion and collaboration forum for scientists and engineers - including computational and computer scientists and engineers - and HPC system vendors, to ensure that HPC systems are optimally configured to support state-of-the-art scientific computing. On the one hand, these discussions will keep NSF and the academic community informed about new products, product roadmap and technology challenges at various vendor organizations. On the other, they will provide HPC system vendors with insights into the major concerns and needs of the academic science and engineering community. These activities will lead to better alignment between applications and hardware both by influencing algorithm design and by influencing system integration.

NSF will also promote resource sharing between and among academic institutions to optimize the accessibility and use of HPC assets deployed and supported at the campus level. This will be accomplished through development of a shared governance structure that includes relevant HPC stakeholders.

2). Development and Maintenance of Supporting Software: New Design Tools, Performance Modeling Tools, Systems Software, and Fundamental Algorithms.

Many of the HPC software and service building blocks in scientific computing are common to a number of science and engineering applications. A supporting software and service infrastructure will accelerate the development of the scientific application codes needed to solve challenging scientific problems, and will help insulate these codes from the evolution of future generations of HPC hardware.

Supporting software services include the provision of intelligent development and problem-solving environments and tools. These tools are designed to provide improvements in ease of use, reusability of modules, and portable performance. Tools and services that are similar across different HPC platforms will greatly reduce the time-to-solution of computationally-intensive research problems by permitting local development of research codes that can then be rapidly transferred to larger production environments or shared with colleagues. Applications scientists and engineers will also benefit from the development of new tools and approaches to debugging, performance analysis, and performance optimization.

Specific applications depend on a broad class of numerical and non-numerical algorithms that are widely used by many applications; for example, linear algebra, fast spectral transforms, optimization algorithms, multi-grid methods, adaptive mesh refinement, symplectic integrators, and sorting and indexing routines. To date, improved or new algorithms have been important contributors to performance improvements in science and engineering applications, the development of multi-grid solvers for elliptic partial differential equations being a prime example. Innovations in algorithms will have a significant impact on the performance of applications software. The development of algorithms for different architectural environments is an essential component of the effort to develop portable, scalable, applications software. Other important software services include libraries for communications services, such as MPI and OpenMP.

[2] The DARPA High Productivity Computing Systems is focused on providing a new generation of economically viable high productivity computing systems. HPCS program researchers have initiated a fundamental reassessment of how performance, programmability, portability, robustness and ultimately, productivity in the HPC domain are defined and measured.

The development and deployment of operating systems and compilers that scale to hundreds of thousands of processors are also necessary. They must provide effective fault-tolerance and must effectively insulate users from parallelization, latency management and thread management issues. To test new developments at large scales, operating systems and kernel researchers and developers must have access to the infrastructure necessary to test their developments at scale.

NSF will support Software Services Providers (SSPs) to develop this supporting software infrastructure. SSPs will be *individually and collectively* responsible for: applied research and development of supporting technologies; harvesting promising supporting software technologies from the research communities; performing scalability/reliability tests to explore software viability; developing, hardening and maintaining software where necessary; and facilitating the transition of commercially viable software into the private sector. SSPs will also support general software engineering consulting services for science and engineering applications, and will provide software engineering consulting support to individual researchers and to research and education teams as necessary.

SSPs will be responsible for ensuring software interoperability with other components of the cyberinfrastructure software stack, such as those generated to provide Data, Data Analysis and Visualization services, and Collaboratories, Observatories and Virtual Organization capabilities – see Chapters 3 and 4 in this document. This will be accomplished through the creation and utilization of appropriate software test harnesses and will ensure that sufficient configuration controls are in place to support the range of HPC platforms used by the research and education community. The applications community will identify needed improvements in supporting software and will provide input and feedback on the quality of services provided by SSPs.

To guide the evolution of the SSP program, NSF will establish an HPC Software Services Council that includes representatives from academe, federal agencies and private sector organizations, including 3rd party and system vendors. The HPC Software Services Council will provide input on the strengths, weaknesses, opportunities and gaps in the software services currently available to the science and engineering research and education communities.

To minimize duplication of effort and to optimize the value of HPC services provided to the science and engineering community, NSF's investments will be coordinated with those of other agencies. DOE currently invests in software infrastructure centers through the Scientific Discovery through Advanced Computing (SciDAC) program, while DARPA's investments in the HPCS program contribute significant systems software and hardware innovations. NSF will seek to leverage and add value to ongoing DOE and DARPA efforts in this area.

3). Development and Maintenance of Portable, Scalable Applications Software

Today's microprocessor-based terascale computers place considerable demands on our ability to manage parallelism, and to deliver large fractions of peak performance. As the agency seeks to create a petascale computing environment, it will embrace the challenge of developing or converting key application codes to run effectively on new and evolving system architectures.

Over the FY 2006 through 2010 period, NSF will make significant new investments in the development, hardening, enhancement and maintenance of scalable applications software, including community models, to exploit the full potential of current terascale and future

petascale systems architectures. The creation of well-engineered, easy-to-use software will reduce the complexity and time-to-solution of today's challenging scientific applications. NSF will promote the incorporation of sound software engineering approaches in existing widely-used research codes and in the development of new research codes. Multidisciplinary teams of researchers will work together to create, modify and optimize applications for current and future systems using performance modeling tools and simulators.

Since the nature and genesis of science and engineering codes varies across the research landscape, a successful programmatic effort in this area will weave together several strands. A new activity will be designed to take applications that have the potential to be widely used within a community or communities, to harden these applications based on modern software engineering practices, to develop versions for the range of architectures that scientists wish to use them on, to optimize them for modern HPC architectures, and to provide user support.

++

CHAPTER 3
STRATEGIC PLAN FOR DATA, DATA ANALYSIS, AND VISUALIZATION (2006-2010)

I. A WEALTH OF SCIENTIFIC OPPORTUNITIES AFFORDED BY DIGITAL DATA

Science and engineering research and education have become increasingly data-intensive, as a result of the proliferation of digital technologies and pervasive networks through which data are collected, generated, shared and analyzed. Worldwide, scientists and engineers are producing, accessing, analyzing, integrating and storing terabytes of digital data daily through experimentation, observation and simulation. Moreover, the dynamic integration of data generated through observation and simulation is enabling the development of new scientific methods that adapt intelligently to evolving conditions to reveal new understanding. The enormous growth in the availability and utility of scientific data is increasing scholarly research productivity, accelerating the transformation of research outcomes into products and services, and enhancing the effectiveness of learning across the spectrum of human endeavor.

New scientific opportunities are emerging from increasingly effective data organization, access and usage. Together with the growing availability and capability of tools to mine, analyze and visualize data, the emerging data cyberinfrastructure is revealing new knowledge and fundamental insights. For example, analyses of DNA sequence data are providing remarkable insights into the origins of man, are revolutionizing our understanding of the major kingdoms of life, and are revealing stunning and previously unknown complexity in microbial communities. Sky surveys are changing our understanding of the earliest conditions of the universe and providing comprehensive views of phenomena ranging from black holes to supernovae. Researchers are monitoring socio-economic dynamics over space and time to advance our understanding of individual and group behavior and their relationship to social, economic and political structures. Using combinatorial methods, scientists and engineers are generating libraries of new materials and compounds for health and engineering, and environmental scientists and engineers are acquiring and analyzing streaming data from massive sensor networks to understand the dynamics of complex ecosystems.

In this dynamic research and education environment, science and engineering data are constantly being collected, created, deposited, accessed, analyzed and expanded in the pursuit of new knowledge. In the future, U.S. international leadership in science and engineering will increasingly depend upon our ability to leverage this reservoir of scientific data captured in digital form, and to transform these data into information and knowledge aided by sophisticated data mining, integration, analysis and visualization tools.

This chapter sets forth a framework in which NSF will work with its partners in science and engineering – public and private sector organizations both foreign and domestic representing data producers, scientists, engineers, managers and users alike – to address data acquisition, access, usage, stewardship and management challenges in a comprehensive way.

II. DEFINITIONS

A. Data, Metadata and Ontologies

In this document, "data" and "digital data" are used interchangeably to refer to data and information stored in digital form and accessed electronically.

- *Data*. For the purposes of this document, data are any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.
- *Metadata*. Metadata are a subset of data, and are data about data. Metadata summarize data content, context, structure, inter-relationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections.
- *Ontology*. An ontology is the systematic description of a given phenomenon, often includes a controlled vocabulary and relationships, captures nuances in meaning and enables knowledge sharing and reuse.

B. Data Collections

This document adopts the definition of data collection types provided in the NSB report on Long-Lived Digital Data Collections[3], where data collections are characterized as being one of three functional types:

- *Research Collections.* Authors are individual investigators and investigator teams. Research collections are usually maintained to serve immediate group participants only for the life of a project, and are typically subjected to limited processing or curation. Data may not conform to any data standards.
- *Resource Collections.* Resource Collections are authored by a community of investigators, often within a domain of science or engineering, and are often developed with community-level standards. Budgets are often intermediate in size. Lifetime is between the mid- and long-term.
- *Reference Collections.* Reference collections are authored by and serve large segments of the science and engineering community, and conform to robust, well-established, comprehensive standards, which often lead to a universal standard. Budgets are large and often derived from diverse sources with a view to indefinite support.

Boundaries between the types are not rigid and collections originally established as research collections may evolve over time to become resource and/or reference collections. In this document, the term data collection is construed to include one or more databases and their relevant technological implementation. Data collections are managed by organizations and individuals with the necessary expertise to structure them and to support their effective use.

III. DEVELOPING A COHERENT DATA CYBERINFRASTRUCTURE IN A COMPLEX, GLOBAL CONTEXT

Since data and data collections are owned or managed by a wide range of communities, organizations and individuals around the world, NSF must work in an evolving environment

[3] Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century, NSB-05-40

constantly being shaped by developing international and national policies and treaties, community-specific policies and approaches, institutional-level programs and initiatives, individual practices, and continually advancing technological capabilities.

At the international level, a number of nations and international organizations have already recognized the broad societal, economic, and scientific benefits that result from open access to science and engineering digital data. In 2004 more than thirty nations, including the United States, declared their joint commitment to work toward the establishment of common access regimes for digital research data generated through public funding[4]. Since the international exchange of scientific data, information and knowledge promises to significantly increase the scope and scale of research and its corresponding impact, these nations are working together to define the implementation steps necessary to enable the global science and engineering system.

The U.S. community is engaged through the National Committee on Data for Science and Technology (CODATA). CODATA is working with its international partners, including the International Council for Science (ICSU), the International Council for Scientific and Technical Information (ICTSI), the World Data Centers (WDCs) and others, to create a Global Information Commons for Science. As currently conceived, this online "open-access knowledge space" will: promote the promise of easy access to and use of scientific data and information; promote wider adoption of successful methods and models for providing open availability on a sustainable basis; facilitate reuse of publicly-funded scientific data and information, as well as cooperative sharing of research materials and tools among researchers; and, encourage and coordinate the efforts of many stakeholders in the world's diverse science and engineering community to achieve these objectives.

A number of international science and engineering communities have also been developing data management and curation approaches for reference and resource collections. For example, the international Consultative Committee for Space Data Standards (CCSDS) defined an archive reference model and service categories for the intermediate and long-term storage of digital data relevant to space missions. This effort produced the Open Archival Information System (OAIS), now adopted as the "de facto" standard for building digital archives, and evidence that a community-focused activity can have much broader impact than originally intended. In another example, the Inter-University Consortium for Political and Social Research (ICPSR) - a membership-based organization with over 500 member colleges and universities around the world - maintains and provides access to a vast archive of social science data. ICPSR serves as a content management organization, preserving relevant social science data and migrating them to new storage media as technology changes, and also provides user support services. ICPSR recently announced plans to establish an international standard for social science documentation. Similar activities in other communities are also underway. Clearly, NSF must maintain a presence in, support, and add value to these ongoing international discussions and activities.

Activities on an international scale are complemented by activities within nation states. In the United States, a number of organizations and communities of practice are exploring mechanisms to establish common approaches to digital data access, management and curation. For example, the Research Library Group (RLG – a not for profit membership organization representing libraries, archives and museums) and the U.S. National Archives and

[4] http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html

Records Administration (NARA – a sister agency whose mission is to provide direction and assistance to Federal agencies on records management) are producing certification requirements for establishing and selecting reliable digital information repositories. RLG and NARA intend their results to be standardized via the International Organization of Standardization (ISO) Archiving Series, and may impact all data collections types. The NIH National Center for Biotechnology Information plays an important role in the management of genome data at the national level, supporting public databases, developing software tools for analyzing data, and disseminating biomedical information.

At the institutional level, colleges and universities are developing approaches to digital data archiving, curation, and analysis. They are sharing best practices to develop digital libraries that collect, preserve, index and share research and education material produced by faculty and other individuals within their organizations. The technological implementations of these systems are often open-source and support interoperability among their adopters. University-based research libraries and research librarians are positioned to make significant contributions in this area, where standard mechanisms for access and maintenance of scientific digital data may be derived from existing library standards developed for print material. These efforts are particularly important to NSF as the agency considers the implications of not just making all data generated with NSF funding broadly accessible, but of also promoting the responsible organization and management of these data such that they are widely usable.

IV. PLAN OF ACTION

Motivated by a vision in which science and engineering digital data are routinely deposited in well-documented form, are regularly and easily consulted and analyzed by specialists and non-specialists alike, are openly accessible while suitably protected, and are reliably preserved, NSF's five-year goal is twofold:

- To catalyze the development of a system of science and engineering data collections that is open, extensible and evolvable; and
- To support development of a new generation of tools and services facilitating data mining, integration, analysis, and visualization essential to turning data into new knowledge and understanding.

The resulting *national digital data framework* will consist of a range of data collections and managing organizations, networked together in a flexible technical architecture using standard, open protocols and interfaces, and designed to contribute to the emerging global information commons. It will be simultaneously local, regional, national and global in nature, and will evolve as science and engineering research and education needs change and as new science and engineering opportunities arise. Widely accessible tools and services will permit scientists and engineers to access and manipulate these data to advance the science and engineering frontier.

In print form, the preservation process is handled through a system of libraries and other repositories throughout the country and around the globe. Two features of this print-based system make it robust. First, the diversity of business models deriving support from a variety of sources means that no single entity bears sole responsibility for preservation, and the system is resilient to changes in any particular sector. Second, there is overlap in the collections, and redundancy of content reduces the potential for catastrophic loss of information.

The *national data framework* is envisioned to provide an equally robust and diverse system for digital data management and access. It will: promote interoperability between data collections supported and managed by a range of organizations and organization types; provide for appropriate protection and reliable long-term preservation of digital data; deliver computational performance, data reliability and movement through shared tools, technologies and services; and accommodate individual community preferences. The agency will also develop a suite of coherent data policies that emphasize open access and effective organization and management of digital data, while respecting the data needs and requirements within science and engineering domains.

The following principles will guide the agency's FY 2006 through FY 2010 investments.

- Science and engineering research and education opportunities and priorities will motivate NSF investments in data cyberinfrastructure.
- Science and engineering data generated with NSF funding will be readily accessible and easily usable, and will be appropriately, responsibly and reliably preserved.
- Broad community engagement is essential in the prioritization and evaluation of the utility of scientific data collections, including the possible evolution between research, resource and reference collection types.
- Continual exploitation of data in the creation of new knowledge requires that investigators have access to the tools and services necessary to locate and access relevant data, and understand its structure sufficiently to be able to interpret and (re)analyze what they find.
- The establishment of strong, reciprocal, international, interagency and public-private partnerships is essential to ensure all stakeholders are engaged in the stewardship of valuable data assets.
- Mechanisms will be created to share data stewardship best practices between nations, communities, organizations and individuals.
- In light of legal, ethical and national security concerns associated with certain types of data, mechanisms essential to the development of both statistical and technical ways to protect privacy and confidentiality will be supported.

A. A Coherent Organizational Framework - Data Collections and Managing Organizations

To date, challenges associated with effective stewardship and preservation of scientific data have been more tractable when addressed through communities of practice that may derive support from a range of sources. For example, NSF supports the Incorporated Research Institutions for Seismology (IRIS) consortium to manage seismology data. Jointly with NIH and DOE, the agency supports the Protein Data Bank to manage data on the three-dimensional structures of proteins and nucleic acids. Multiple agencies support the University Consortium for Atmospheric Research, an organization that has provided access to atmospheric and oceanographic data sets, simulations, and outcomes extending back to the 1930s through the National Center for Atmospheric Research.

Existing collections and managing organization models reflect differences in culture and practice within the science and engineering community. As community proxies, data collections and their managing organizations can provide a focus for the development and dissemination of appropriate standards for data and metadata content and format, guided by an appropriate community-defined governance approach. This is not a static process, as new disciplinary

fields and approaches, data types, organizational models and information strategies inexorably emerge. This is discussed in detail in the Long-Lived Digital Data Collections report of the National Science Board.

Since data are held by many Federal agencies, commercial and non-profit organizations, and international entities, NSF will foster the establishment of interagency, public-private and international consortia charged with providing stewardship for digital data collections to promote interoperability across data collections. The agency will work with the broad community of science and engineering producers, managers, scientists and users to develop a common conceptual framework. A full range of mechanisms will be used to identify and build upon common ground across domain communities and managing organizations, engaging all stakeholders. Activities will include: the support of new projects; development and implementation of evaluation and assessment criteria that, amongst other things, reveal lessons learned across communities; support of community and inter-community workshops; and the development of strong partnerships with other stakeholder organizations. Stakeholders in these activities include data authors, data managers, data scientists and engineers, and data users representing a diverse range of communities and organizations, including universities and research libraries, government agencies, content management organizations and data centers, and industry.

To identify and promote lessons learned across managing organizations, NSF will continue to promote the coalescence of appropriate collections with overlapping interests, approaches, and services. This reduces data-driven fragmentation of science and engineering domains. Progress is already being made in some areas. For example, NSF has been working with the environmental science and engineering community to promote collaboration across disciplines ranging from ecology and hydrology to environmental engineering. This has resulted in the emergence of common cyberinfrastructure elements and new interdisciplinary science and engineering opportunities.

B. Developing A Flexible Technological Architecture

From a technological perspective, the *national data framework* must provide for reliable preservation, access, analysis, interoperability, and data movement, possibly using a web or grid services distributed environment. The architecture must use standard open protocols and interfaces to enable the broadest use by multiple communities. It must facilitate user access, analysis and visualization of data, addressing issues such as authentication, authorization and other security concerns, and data acquisition, mining, integration, analysis and visualization. It must also support complex workflows enabling data discovery. Such an architecture can be visualized as a number of layers providing different capabilities to the user, including data management, analysis, collaboration tools, and community portals. The connections among these layers must be transparent to the end user, and services must be available as modular units responsive to individual or community needs. The system is likely to be implemented as a series of distributed applications and operations supported by a number of organizations and institutions distributed throughout the country. It must provide for the replication of data resources to reduce the potential for catastrophic loss of digital information through repeated cycles of systems migration and all other causes since, unlike printed records, the media on which digital data are stored and the structures of the data are relatively fragile.

High quality metadata, which summarize data content, context, structure, inter-relationships, and provenance (information on history and origins), are critical to successful information

management, annotation, integration and analysis. Metadata take on an increasingly important role when addressing issues associated with the combination of data from experiments, observations and simulations. In these cases, product data sets require metadata that describe, for example, relevant collection techniques, simulation codes or pointers to archived copies of simulation codes, and codes used to process, aggregate or transform data. These metadata are essential to create new knowledge and to meet the reproducibility imperative of modern science. Metadata are often associated with data via markup languages, representing a consensus around a controlled vocabulary to describe phenomena of interest to the community, and allowing detailed annotations of data to be embedded within a data set. Because there is often little awareness of markup language development activities within science and engineering communities, energy is expended reinventing what could be adopted or adapted from elsewhere. Scientists and engineers therefore need access to tools and services that help ensure that metadata are automatically captured or created in real-time.

Effective data analysis tools apply computational techniques to extract new knowledge through a better understanding of the data, its redundancies and relationships, by filtering extraneous information and by revealing previously unseen patterns. For example, the Large Hadron Collider at CERN generates such massive data sets that the detection of both expected events, such as the Higgs boson, and unexpected phenomena requires the development of new algorithms, both to manage data and to analyze it. Algorithms and their implementations must be developed for statistical sampling, for visualization, to enable the storage, movement and preservation of enormous quantities of data, and to address other unforeseen problems certain to arise.

Scientific visualization, including not just static images but also animation and interaction, leads to better analysis and enhanced understanding. Currently, many visualization systems are domain or application-specific and require a certain commitment to understand or learn to use. Making visualization services more transparent to the user lowers the threshold of usability and accessibility, and makes it possible for a wider range of users to explore or use a data collection. Analysis of data streams also introduces problems in data visualization and may require new approaches for representing massive, heterogeneous data streams.

Deriving knowledge from large data sets presents specific scaling problems due to the sheer number of items, dimensions, sources, users, and disparate user communities. The human ability to process visual information can augment analysis, especially when analytic results are presented in iterative and interactive ways. Visual analytics, the science of analytical reasoning enabled by interactive visual interfaces, can be used to synthesize the information content and derive insight from massive, dynamic, ambiguous, and even conflicting data. Suitable fully interactive visualizations help us absorb vast amounts of data directly, to enhance our ability to interpret and analyze otherwise overwhelming data. Researchers can thus detect the expected and discover the unexpected, uncovering hidden associations and deriving knowledge from information. As an added benefit, their insights are more easily and effectively communicated to others.

Creating and deploying visualization services requires new frameworks for distributed applications. In common with other cyberinfrastructure components, visualization requires easy-to-use, modular, extensible applications that capitalize on the reuse of existing technology. Today's successful analysis and visualization applications use a pipeline, component-based system on a single machine or across a small number of machines. Extending to the broader distributed, heterogeneous cyberinfrastructure system will require new interfaces and work in

fundamental graphics and visualization algorithms that can be run across remote and distributed settings.

To address this range of needs for data tools and services, NSF will work with the broad community to identify and prioritize needs. In making investments, NSF will complement private sector efforts, for example, those producing sophisticated indexing and search tools and packaging them as data services. NSF will support projects to: conduct applied research and development of promising, interoperable data tools and services; perform scalability/reliability tests to explore tool viability; develop, harden and maintain software tools and services where necessary; and, harvest promising research outcomes to facilitate the transition of commercially-viable software into the private sector. Data tools created and distributed through these projects will include not only access and ease-of-use tools, but tools to assist with data input, tools that maintain or enforce formatting standards, and tools that make it easy to include or create metadata in real time. Clearinghouses and registries from which all metadata, ontology, and markup language standards are provided, publicized, and disseminated must be developed and supported, together with the tools for their implementation. Data accessibility and usability will also be improved with the support of means for automating cross-ontology translation. Collectively, these projects will be responsible for ensuring software interoperability with other components of the cyberinfrastructure, such as those generated to provide High Performance Computing capabilities and to enable the creation of Collaboratories, Observatories and Virtual Organizations.

The user community will work with tool providers as active collaborators to determine requirements and to serve as early users. Scientists, educators, students and other end users think of ways to use data and tools that the developers didn't consider, finding problems and testing recovery paths by triggering unanticipated behavior. Most importantly, an engaged set of users and testers will also demonstrate the scientific value of data collections. The value of repositories and their standards-based input and output tools arises from the way in which they enable discoveries. Testing and feedback are necessary to meet the challenges presented by current datasets that will only increase in size, often by orders of magnitude, in the future.

Finally, in addition to promoting the *use* of standards, tool and service developers will also promote the *stability* of standards. Continual changes to structure, access methods, and user interfaces, mitigate against ease of use, and against interoperability. Instead of altering a standard for a current need, developers will adjust their implementation of that need to fit within the standard. This is especially important for resource-limited research and education communities.

C. Developing and Implementing Coherent Data Policies

In setting priorities and making funding decisions, NSF is in a powerful position to influence data policy and management at research institutions. NSF's policy position on data is straightforward: all science and engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected and preserved. Through a suite of coherent policies designed to recognize different data needs and requirements within communities, NSF will promote open access to well-managed data recognizing that this is essential to continued U.S. leadership in science and engineering.

In addition to addressing the technological challenges inherent in the creation of a *national data framework*, NSF's data policies will be redesigned to overcome existing sociological and cultural barriers to data sharing and access. Two actions are critical. NSF will conduct an inventory of existing policies, to bring them into accord across programs and to ensure coherence. This will lead to the development of a suite of harmonized policy statements supporting data open access and usability. NSF's actions will promote a change in culture such that the collection and deposition of all appropriate digital data and associated metadata become a matter of routine for investigators in all fields. This change will be encouraged through an NSF-wide requirement for data management plans in all proposals. These plans will be considered in the merit review process, and will be actively monitored post-award.

Policy and management issues in data handling occur at every level, and there is an urgent need for rational agency, national and international strategies for sustainable access, organization and use. Discussions at the interagency level on issues associated with data policies and practices will be supported by a new interagency working group on digital data recently proposed by NSF under the auspices of the Committee on Science of the National Science and Technology Council. This group will consider not only data challenges and opportunities discussed throughout this chapter, but especially the issues of cross-agency funding and access, the provision and preservation of data to and for other agencies, and monitoring agreements as agency imperatives change with time. Formal policies must be developed to include data quality and security, ethical and legal requirements, and technical and semantic interoperability issues, throughout the complete process from collection and generation to analysis and dissemination.

As already noted, many large science and engineering projects are international in scope, where national laws and international agreements directly affect data access and sharing practices. Differences arise over privacy and confidentiality, from cultural attitudes to ownership and use, in attitudes to intellectual property protection and its limits and exceptions, and because of national security concerns. Means by which to find common ground within the international community must continue to be explored.

V. CONCLUSION

NSF is in a unique position to influence the science and engineering communities throughout the country, emphasizing the needs and importance of standardized straightforward access to and use of digital data collections. NSF will promote data management practices and the development of new tools and new algorithms that encourage appropriate levels of openness and sharing. NSF's merit review process will allow communities to select which of their collections should be long-lived, and to trade data preservation costs against new research costs in the way that most benefits their discipline. NSF's strong working relations with the community, with other agencies, and with its international partners, will engender the widespread existence of collaboratively supported, community-run data collections. By adopting persuasive strategies during the review and funding of proposals and projects, while respecting the diverse and disparate communities it serves, NSF has the opportunity to stimulate a cultural change that will greatly accelerate the pace of discovery. Twenty-first century science and engineering research and education deserve no less.

++

CHAPTER 4
STRATEGIC PLAN FOR COLLABORATORIES, OBSERVATORIES AND VIRTUAL ORGANIZATIONS (2006-2010)

Under development.

Collaboratories, observatories and virtual organizations – IT-enabled knowledge environments made possible by contemporary communication and networking technologies - allow scientists and engineers to pursue their research and education goals without regard to geographical location. In these environments, individuals will be able to access experimental and computational tools, interact with their colleagues, and share data, information and knowledge. This chapter will focus on the tools and services necessary to create these highly-interactive, widely-accessible environments to promote progress in science and engineering.

CHAPTER 5
STRATEGIC PLAN FOR EDUCATION & WORKFORCE (2006-2010)

Under development.

NSF recognizes that cyberinfrastructure will have a profound impact on the practice of science and engineering research and education, enabling individuals, groups and organizations to advance science and engineering in ways that revolutionize *what they can do*, *how they do it*, and *who can participate*. To harness the full power of cyberinfrastructure and the promise it portends for discovery, learning and innovation across and within all areas of science and engineering requires focused investments in the preparation of a science and engineering workforce with the knowledge and requisite skills needed to create, advance and exploit cyberinfrastructure over the long-term. This chapter will describe NSF's approach to doing so.

APPENDIX A: REPRESENTATIVE REPORTS AND WORKSHOPS

Building a Cyberinfrastructure for the Biological Sciences; workshop held July 14-15, 2003; information available at http://research.calit2.net/cibio/archived/CIBIO_FINAL.pdf and http://research.calit2.net/cibio/report.htm

CHE Cyber Chemistry Workshop; workshop held October 3-5, 2004; information available at http://bioeng.berkeley.edu/faculty/cyber_workshop

Commission on Cyberinfrastructure for the Humanities and Social Sciences; sponsored by the American Council of Learned Societies; seven public information-gathering events held in 2004; report in preparation; information available at http://www.acls.org/cyberinfrastructure/cyber.htm

Community Climate System Model Strategic Business Plan (2003), 28pp; information available at http://www.ccsm.ucar.edu/management/busplan2004-2008.pdf

Community Climate System Model Science Plan 2004-2008 (2003)," 76pp; information available at http://www.ccsm.ucar.edu/management/sciplan2004-2008.pdf

Computation as a Tool for Discovery in Physics; report by the Steering Committee on Computational Physics: information available at http://www.nsf.gov/pubs/2002/nsf02176/start.htm

Cyberinfrastructure for the Atmospheric Sciences in the 21st Century; workshop held June 2004; information available at http://netstats.ucar.edu/cyrdas/report/cyrdas_report_final.pdf

Cyberinfrastructure for Engineering Research and Education; workshop held June 5 – 6, 2003; information available at http://www.nsf.gov/eng/general/Workshop/cyberinfrastructure/index.jsp

Cyberinfrastructure for Environmental Research and Education (2003*)*; workshop held October 30 – November 1, 2002; information available at http://www.ncar.ucar.edu/cyber/cyberreport.pdf

CyberInfrastructure (CI) for the Integrated Solid Earth Sciences (ISES) (June 2003); *workshop held on March 28-29, 2003;*, June 2003; information available at http://tectonics.geo.ku.edu/ises-ci/reports/ISES-CI_backup.pdf

Cyberinfrastructure and the Social Sciences (2005); workshop held March 15-17, 2005; information available at http://www.sdsc.edu/sbe/

Cyberinfrastructure needs for environmental observatories; information available at http://www.orionprogram.org/office/NSFCyberWkshp.html

Cyberlearning Workshop Series; workshops held Fall 2004 – Spring 2005 by the Computing Research Association (CRA) and the International Society of the Learning Sciences (ISLS); information available at http://www.cra.org/Activities/workshops/cyberlearning

Data Management for Marine Geology and Geophysics: Tools for Archiving, Analysis, and Visualization (2001); information available at http://humm.whoi.edu/DBMWorkshop/data_mgt_report.hi.pdf

Environmental Cyberinfrastucture Needs For Distributed Sensor Networks; workshop held August 12-14, 2003; information available at http://www.lternet.edu/sensor_report

Federal Plan for High-End Computing (2004); 72 pp; available at: http://www.ostp.gov/nstc/html/HECRTF-FINAL_051004.pdf

Geoinformatics: Building Cyberinfrastructure for the Earth Sciences (2004)*;* workshop held May 14 – 15, 2003; Kansas Geological Survey Report 2004-48; information available at http://www.geoinformatics.info

Geoscience Education and Cyberinfrastructure, Digital Library for Earth System Education, (2004); workshop held April 19-20, 2004; information available at http://www.dlese.org/documents/reports/GeoEd-CI.pdf

Getting Up to Speed: The Future of Supercomputing (2004). 308pp; available at: http://www.nap.edu/books/0309095026/html/ or http://www.sc.doe.gov/ascr/Supercomputing%20Prepub-Nov9v4.pdf)

High-Performance Computing Requirements for the Computational Solid Earth Sciences (2005)*;* 96 pp; available at: http://www.geo-prose.com/computational_SES.html.

Identifying Major Scientific Challenges in the Mathematical and Physical Sciences and their CyberInfrastructure Needs, workshop held April 21, 2004; information available at http://www.nsf.gov/attachments/100811/public/CyberscienceFinal4.pdf

Improving the effectiveness of U.S. Climate modeling, Commission on Geosciences, Environment and Resources (2001). National Academy Press, Washington, D.C., 144pp; information available at http://www.nap.edu/books/0309072573/html/

An Information Technology Infrastructure Plan to Advance Ocean Sciences (2002). 80 pp. available at http://www.geo-prose.com/oiti/index.html

Materials Research Cyberscience enabled by Cyberinfrastructure; workshop held June 17 – 19, 2004; information available at http://www.nsf.gov/mps/dmr/csci.pdf

Multi-disciplinary Workshop at the Interface of Cyber infrastructure, and Operations Research, with "Grand Challenges" in Enterprise-wide Applications in Design, Manufacturing and Services; workshop held August 31 - September 1, 2004; information available at https://engineering.purdue.edu/PRECISE/CI-OR/index.html

Multiscale Mathematics Initiative: A Roadmap; workshops held May 3-5, July 20-22, September 21-23, 2004; information available at www.sc.doe.gov/ascr/mics/amr/Multiscale%20Math%20Workshop%203%20-%20Report%20latest%20edition.pdf

NIH/NSF Spring 2005 Workshop on Visualization Research Challenges; workshop held on May 2-3, 2005; information available at http://www.sci.utah.edu/vrc2005/index.html

An Operations Cyberinfrastructure: Using Cyberinfrastructure and Operations Research to Improve Productivity in American Enterprises"; workshop held August 30 – 31, 2004; information available at http://www.optimization-online.org/OCI/OCI.doc; http://www.optimization-online.org/OCI/OCI.pdf

Planning for Cyberinfrastructure Software (2005); workshop held October 5 – 6, 2004; information available at www.nsf.gov/cise/sci/ci_workshop/index.jsp

Preparing for the Revolution: Information Technology and the Future of the Research University (2002); NRC Policy and Global Affairs, 80 pages; information available at http://www.nap.edu/catalog/10545.html

Polar Science and Advanced Networking: workshop held on April 24 - 26, 2003; sponsored by OPP/CISE; information available at http://www.polar.umcs.maine.edu

Recurring Surveys: Issues and Opportunities: workshop held March 28-29, 2003; information available at www.nsf.gov/sbe/ses/mms/nsf04_211a.pdf (2004)

Research Opportunities in CyberEngineering/CyberInfrastructure; workshop held April 22 - 23, 2004; information available at http://thor.cae.drexel.edu/~workshop/

Revolutionizing Science and Engineering Through Cyberinfrastructure: report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure; Daniel E. Atkins (Chair), January 2003; information available at http://www.nsf.gov/cise/sci/reports/atkins.pdf

Roadmap for the Revitalization of High-End Computing (200?); available at http://www.hpcc.gov/hecrtf-outreach/20040112_cra_hecrtf_report.pdf

Science-Based Case for Large-Scale Simulation; workshop held June 24-25, 2003; information available at http://www.pnl.gov/scales/docs/volume1_72dpi.pdf; http://www.pnl.gov/scales/docs/SCaLeS_v2_draft_toc.pdf

Supplement to the President's Budget for FY 2006; Report by the Subcommittee on Networking and Information Technology Research and Development (NITRD), February 2005; information available at http://www.nitrd.gov

Trends in IT Infrastructure in the Ocean Sciences (2004); workshop held May 21-23, 2003; information available at http://www.geo-prose.com/oceans_iti_trends/oceans_iti_trends_rpt.pdf

APPENDIX B: CHRONOLOGY OF NSF IT INVESTMENTS

NSF's early investments in what has now become known as cyberinfrastructure date back almost to the agency's inception. In the 1960's and 1970's, the agency supported a number of campus-based computing facilities. As computational methodologies became increasingly essential to the research endeavor, the science and engineering community began to call for NSF investments in specialized, higher capability computing facilities that would meet the computational needs of the broad national community. As a consequence, NSF's Supercomputer Centers program was initiated in 1985 through the agency's support of five academic-based supercomputer centers.

During the 1980's, academic-based networking activities also flourished. Networking technologies were expected to improve the effectiveness and efficiency of researchers and educators, providing enhanced, easier access to computer resources and more effective transfer and sharing of information and knowledge. After demonstrating the potential of CSNET in linking computer science departments, NSF moved on to develop the high-speed backbone, called NSFNET, with the five supercomputer centers supported under the Supercomputer Centers program and the National Center for Atmospheric Research becoming the first nodes on the backbone. NSF support also encouraged the development of regional networks to connect with the backbone NSFNET, thereby speeding the adoption of networking technologies on campuses around the country. In 1995, in partnership with MCI, NSF catalyzed support of the vBNS permitting advanced networking research and the development of novel scientific applications. A few years later, we established the NSF Middleware Initiative, focused on the development of advanced networking services to serve the evolving needs of the science and engineering community.

In the early to mid-1990's, informed by both the Branscomb and the Hayes Reports, NSF consolidated its support of national computing facilities in the establishment of the Partnerships for Advanced Computational Infrastructure (PACI) program. Two partnerships were established in 1997, together involving nearly 100 partner institutions across the country in efforts to make more efficient use of high-end computing in all areas of science and engineering. The partnerships have been instrumental in fostering the maturation of cyberinfrastructure and its widespread adoption by the academic research and education community, and by industry.

Also in the early 1990's, NSF as part of the U.S. High-Performance Computing and Communications (HPCC) program, began to support larger-scale research and education-focused projects pursuing what became known as "grand challenges." These HPCC projects joined scientists and engineers, computer scientists and state-of-the-art cyberinfrastructure technologies to tackle important problems in science and engineering whose solution could be advanced by applying cyberinfrastructure techniques and resources. First coined by the HPCC program, the term "grand challenge" has been widely adopted in many science and engineering fields to signify an overarching goal that requires a large-scale, concerted effort.

During the 1990's, the penetration of increasingly affordable computing and networking technologies on campuses was also leading to the creation of what would become mission-critical, domain-specific cyberinfrastructure. For example, in the mid 1990's the earthquake engineering community began to define what would become the Network for Earthquake Engineering Simulation, one of many significant cyberinfrastructure projects in NSF's portfolio today.

In 1999, the President's Information Technology Advisory Committee (PITAC) released the seminal report *ITR-Investing in our Future*, prompting new and complementary NSF investments in CI projects, such the Grid Physics Network (GriPhyN) and international Virtual Data Grid Laboratory (iVDGL) and the Geosciences Network, known as GEON. Informed by the PITAC report, NSF also created an MREFC project entitled Terascale Computing Systems that began its construction phase in FY 2000 and ultimately created the Extensible Terascale Facility – now popularly known as the Teragrid. Teragrid entered its production phase in October 2004 and represents one of the largest, fastest, most comprehensive distributed cyberinfrastructures for science and engineering research and education.

In 2001, NSF charged an Advisory Committee for Cyberinfrastructure under the leadership of Dr. Dan Atkins, to evaluate the effectiveness of PACI and to make recommendations for future NSF investments in cyberinfrastructure. The Atkins Committee, as it became popularly known, recommended support for the two Partnership lead sites through the end of their original PACI cooperative agreements. In October 2004, following merit review, the National Science Board (NSB) endorsed funding of those sites through the end of FY 2007.

Through 2005, in addition to the groups already cited, a number of prestigious groups have made recommendations that continue to inform the agency's cyberinfrastructure planning including the High-End Computing Revitalization Task Force, the PITAC Subcommittee on Computational Science, and the NRC Committee on the Future of Supercomputing.

APPENDIX C: MANAGEMENT OF CYBERINFRASTRUCTURE

NSF has nurtured the growth of what is now called cyberinfrastructure for a number of decades. In recent years, the Directorate for Computer and Information Science and Engineering (CISE) has been responsible for the provision of national supercomputing infrastructure for the academic community. In addition, the Directorate was instrumental in the creation of what ultimately became known as the Internet. During this incubation period, the management of CI was best provided by those also responsible for the research and development of related CI technologies.

Over the years, the penetration and impact of computing and networking on campuses has been extensive, and has led to the creation of many disciplinary-specific or community-specific CI projects and activities. Today, CI projects are supported by all NSF Directorates and Offices. Because of the growing scope of investment and variability in needs among users in the broad science and engineering community, it has become clear that effective CI development and deployment now requires the collective leadership of NSF senior management. This leadership will be provided by a Cyberinfrastructure Council chaired by the NSF Director and comprised of the NSF Deputy Director, the Assistant Directors of NSF's Directorates (BIO, CISE, GEO, EHR, ENG, MPS, and SBE) and the Heads of the Office of International Science and Engineering, Office of Polar Programs, and the recently established Office of Cyberinfrastructure (OCI). The Cyberinfrastructure Council has been meeting regularly since May 2005, and OCI was established in the Office of the Director on July 22, 2005.

CISE will continue to be responsible for a broad range of programs that address the Administration's priorities for fundamental research and education in computing, representing more than 85% of the overall federal investment in university-based basic research.