Congressional Testimony

Subcommittee on Research and Science Education
Committee on Science and Technology
U.S. House of Representatives

Hearing on "The State of Research Infrastructure at U.S. Universities"
Tuesday, 23 February 2010

Thom H. Dunning, Jr.
Director, National Center for Supercomputing Applications and Institute for Advanced Computing Applications and Technologies, University of Illinois at Urbana-Champaign, Urbana, Illinois

What Is Cyberinfrastructure?

Cyberinfrastructure*, n., cyberinfrastructure consists of computing systems, data sources and data storage systems, visualization environments, and support staff, all linked by high speed networks to make discoveries and innovations not otherwise possible.*

Over the past quarter century, computing has become an integral part of the fabric of experimental and theoretical science. All but the simplest laboratory experiments are performed under computer control, the data is analyzed using software running on a personal computer or small compute cluster, and the results compared with the latest theories through computational simulations on high performance computers. The use of computing technology is now spreading to the observational sciences, which are being

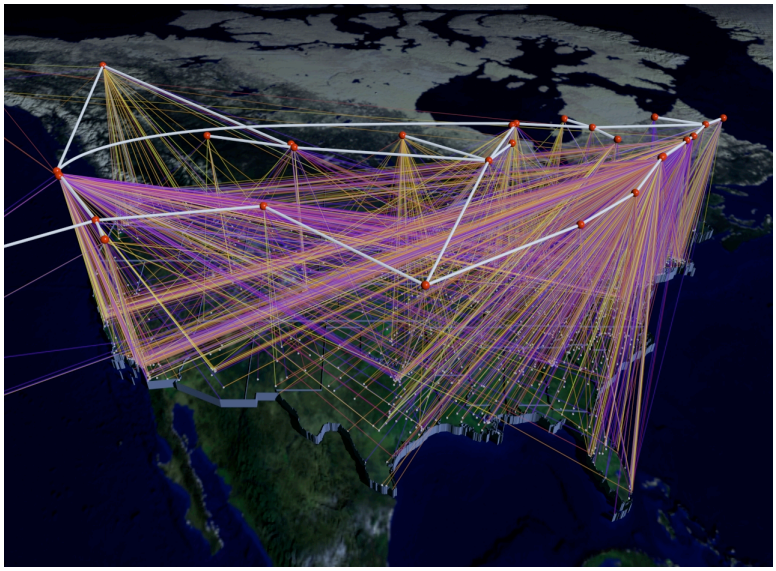

Figure 1. Cyberinfrastructure is a networked collection of computer systems, data sources and stores, and visualization systems linked by a software infrastructure that integrates these systems into a unique and powerful research capability.

revolutionized by the advent of powerful new sensors that can detect and record a wide range of physical, chemical and biological phenomena—from massive digital detectors in a new generation of telescopes to sensor arrays for characterizing ecological and geological areas and new advanced sequencing instruments for genomics research.

Research Advances Enabled by Cyberinfrastructure

Three major modes of scientific discovery are enabled by cyberinfrastructure: computational modeling and simulation, data-driven discovery, and, increasingly, the coupling of these two modes. To address the questions posed by the Subcommittee, I will discuss the cyberinfrastructure needs of these three modes of scientific discovery and then provide an analysis of the status of the existing cyberinfrastructure. To begin, let us briefly review the science and engineering advances made possible by cyberinfrastructure.

Computational Modeling and Simulation. In computational modeling and simulation, scientists develop a mathematical model of the phenomena of interest, e.g., the chemical and physical processes involved in an internal combustion engine or the processes involved in the prediction of weather, and then use high performance computers to solve the resulting equations. For most phenomena of interest, the equations are very complex and, so, the power of computational modeling and simulation grows with increases in computing power. As computing systems have progressed from the megaflops era in the 1970s to the petaflops era of today, our ability to accurately simulate a broad range of biological, chemical, physical and, even, social phenomena has grown dramatically.

- The Southern California Earthquake Center seeks to develop a predictive understanding of earthquake processes aimed at providing society with improved understanding of seismic hazards. In partnership with earthquake engineers, SCEC researchers are developing the ability to conduct end-to-end simulations ("rupture to rafters") to extend this improved understanding of seismic hazards to an improved understanding of earthquake risks and risk mitigation strategies.
- Researchers at the University of Illinois at Urbana-Champaign are using computational simulations to obtain a detailed understanding of the functioning of the ribosome, the large cellular machine responsible for synthesizing proteins in our cells, as well understanding the mechanism used by the poliovirus to gain entry into our cells. The former will enhance our fundamental understanding of cell biology, while the latter may lead to the development of better anti-viral drugs.
- A team from Michigan State University and the University of California, San Diego are studying the formation of the first galaxies. Based on a fundamental understanding of the physical processes and the initial conditions that led to the formation of the first stars, powerful numerical simulations are helping astrophysicists understand how and when the very first sources of light formed.

All of these simulations are numerical- and data-intensive and can only be performed on the most powerful computers available.

Data-driven Discovery. In data-driven discovery, scientists gather information from various data sources, e.g., a large digitally-enabled telescope, an array of environmental sensors, or "gangs" of genome sequencers, and then analyze the resulting mass of data using

sophisticated mathematical procedures seeking patterns, information and understanding. Data-driven discovery requires an extensive cyberinfrastructure that supports data collection and transport to storage sites, followed by data cataloging, integration and analysis (including visualization). Often, the cataloged data becomes a resource for a large research community. Depending on the quantities of data involved as well as the mathematical demands of the analysis, data-driven discovery may require extensive computing resources as well as large data storage facilities.

- The Ocean Observatory Initiative is constructing an integrated observatory network to provide the oceanographic research and education community with: (*i*) a cabled network of monitoring devices on the sea floor spanning important geological and oceanographic features, (*ii*) an array of relocatable deep-sea buoys that can be deployed in harsh environments, and (*iii*) construction of new facilities or enhancements to existing facilities leading to an expanded network of coastal observatories. The OOI will provide earth and ocean scientists with unique opportunities to study multiple, interrelated processes over timescales ranging from seconds to decades; to conduct comparative studies of regional processes and spatial characteristics; and to map whole-earth and basin scale structures.[1]

- The Large Synoptic Survey Telescope (LSST) is unlike other ground-based telescopes. It is a wide-field survey telescope and camera that can image the entire sky in just three nights, providing a time history of celestial events. Using an 8.4-meter ground-based telescope, the LSST will, for the first time, produce a wide-field astronomical survey of our universe. Its 3 gigapixel camera—the world's largest digital camera—will provide digital imaging of faint astronomical objects. The LSST will provide unprecedented three-dimensional maps of the mass distribution in the universe, in addition to the traditional images of luminous stars and galaxies. These maps will be used to better understand the nature of the mysterious dark energy that is driving the accelerating expansion of the universe. In addition, the LSST will also provide a comprehensive census of our solar system, including potentially hazardous near-Earth asteroids

Data-driven Computational Modeling and Simulation. There are increasing opportunities for linking data-driven discovery with computational modeling and simulation. For example, in the NSF-funded LEAD project (Linked Environments for Atmospheric Discovery[2]), one of the goals is to gather and analyze the data from a distributed array of Doppler radars to determine, *in real time*, when atmospheric conditions are ripe for the formation of a tornado and then launch computational simulations to determine the likely path and intensity of the tornado. Such opportunities will grow in the future as sources of sensed data become more widespread.

Development of a National Cyberinfrastructure

In recognition of the increasing importance of research cyberinfrastructure, the National Science Foundation recently issued a Dear Colleague Letter on "Cyberinfrastructure Framework for 21st Century Science and Engineering." This letter stated that it was

[1] See: http://www.oceanleadership.org/programs-and-partnerships/ocean-observing/.

[2] See: https://portal.leadproject.org/gridsphere/gridsphere.

imperative for NSF to develop a long term vision for the nation's cyberinfrastructure that covered the following critical areas:

1. Cyberinfrastructure for:
    a. High end computational, data, visualization and sensor-based systems and the associated user support for transformative science.
    b. NSF's large-scale collaborative research facilities and projects, integrated with that of other federal agencies and international organizations.
2. Linkage of this cyberinfrastructure into campuses (including government and businesses) accompanied by programs that support integrated, widely dispersed, broadly based activities and resources.
3. Education and outreach to help develop computational science- and technology-savvy researchers and workforce.

This letter was signed by all of the Assistant Directors at NSF as well as the directors of many major NSF programs.

The development of a national cyberinfrastructure for research poses many unique challenges for NSF. Cyberinfrastructure is very different from physical infrastructure such as a laboratory building. Computing and related technologies are still rapidly advancing—computing power doubles every two years, disk capacity increases even more rapidly, 60% per year. The software that ties all of the infrastructure elements together to create a unique research capability has to be revised in response to these changes in technology. Finally, the use of cyberinfrastructure is still in its infancy—high quality support staff are needed to ensure that the U.S. research community can take advantage of the new capabilities provided by cyberinfrastructure. This close coupling of hardware, software, and expertise with a rapidly changing technology base is unparalleled in other types of infrastructure.

Cyberinfrastructure must also be funded through different mechanisms. Infrastructure must be sustained over long periods of time to be useful to researchers, and it cannot be sustained through a series of short term, loosely integrated projects. Like an interstate highway, cyberinfrastructure must provide a smooth, continuous path from one point to another. On the other hand, cyberinfrastructure must also evolve as computing technology advances; otherwise, it will rapidly become outdated. So, there must be flexibility in how the funding is used in long term cyberinfrastructure projects. Finally, cyberinfrastructure is expensive, both in terms of the hardware that must be deployed, the software that must be developed and maintained, and the support staff that are critical for its efficient functioning. It is important to avoid duplication and leverage existing capabilities and resources whenever possible.

The NSF-wide Advisory Committee for Cyberinfrastructure has begun work on the development of the new cyberinfrastructure framework outlined in the Dear Colleague letter,[3] establishing six Task Forces:

Campus Bridging Data

[3] See: https://nsf.sharepointspace.com/acci_public/default.aspx.

Grand Challenges	High Performance Computing
Software and Tools	Work Force Development

The Task Forces involve distinguished scientists and engineers from across the nation as well as NSF program officers. Although the Task Forces are in the early stages of their work, they have already held a number of meetings and teleconferences to explore and discuss new concepts and strategies for developing a comprehensive national cyberinfrastructure. I am participating in three of these Task Forces: Grand Challenges, Software and Tools, and High Performance Computing and have colleagues who are involved in the other three Task Forces. This testimony provides a prologue to the work of the six NSF Task Forces.

Before moving on, I should note that NSF is not the only federal agency that supports cyberinfrastructure in the nation's universities. The Office of Science in the U.S. Department of Energy (DOE-SC) also funds cyberinfrastructure for university researchers. DOE-SC has a well defined, long term plan to provide computational, data and communications resources for laboratory and academic researchers based on the identified needs of its major research programs. However, with the exception of the INCITE program,[4] DOE-SC's cyberinfrastructure is closely tied to its mission needs, serving only those laboratories and universities deemed critical to that mission. The National Institutes of Health (NIH) supports a number of cyberinfrastructure-related software development efforts in its biomedical research programs but, by and large, depends on agencies such as NSF as well as the academic institutions that it supports to provide much of its cyberinfrastructure, especially the hardware. However, biomedical research is approaching a tipping point—the amount of data being accumulated in NIH's research programs will soon far exceed that which can be stored, managed and analyzed by the other agencies and institutions. NIH has several strategic planning activities underway to identify the best path forward. Whatever the outcome of these planning activities, meeting the growing computing and data needs of NIH's intramural and extramural research programs will surely require substantial increases in NIH's cyberinfrastructure investments.

High Performance Computing

As noted earlier, advances in computational modeling and simulation are driven by increases in computing power. Over the past few decades, increases in computing power have largely been driven by Moore's Law, with a doubling in computing power occurring every 18-24 months. Thus, the end of the 1980s saw the deployment of computers capable of performing a billion arithmetic operations per second.[5] Ten years later, computing technology had advanced to the point that it was possible to perform a trillion arithmetic operations per second. In 2008, computers capable of a quadrillion operations per second were deployed. It is expected that exascale computers, 1,000 times more powerful than petascale computers, will arrive in another 8 years, although many hardware and software challenges must first be overcome.

The National Science Foundation (NSF) and the Office of Science in the U.S. Department of Energy (DOE-SC) have committed to providing high performance computing resources for

[4] See: http://www.er.doe.gov/ascr/incite/index.html.

[5] A typical arithmetic operation is the multiplication of two 14-digit numbers to yield a 14-digit result.

open scientific and engineering research, including for researchers who are funded by other government agencies. DOE-SC is funding several major computing centers in support of its energy and environmental missions as well as its broader national science mission: its flagship facility at Lawrence Berkeley National Laboratory and its leadership computing facilities at Oak Ridge National Laboratory and Argonne National Laboratory. NSF funds large national computing facilities at the Texas Advanced Computing Center and University of Tennessee/Oak Ridge National Laboratory and its largest national facility at the University of Illinois at Urbana-Champaign. Although I am familiar with DOE's computing program, I will only discuss NSF's program here since NSF's programs are most relevant to the Subcommittee's charge. However, DOE-SC's contributions to the national cyberinfrastructure should be kept in mind.

Cyberinfrastructure for High Performance Computing. NSF's high performance computing plan for 2006-2010 was outlined in the document "*Cyberinfrastructure Vision for 21st Century Discovery*" (March 2007). The report recognized the need, first articulated in the Branscomb report,[6] for a broad range of computing resources, from leadership-class national computing resources to university high performance computers and the compute/data clusters and workstations used by small research groups—the so-called Branscomb pyramid.[7] The report stated NSF's intent to fund the highest performance computing systems, the so-called Track 1 and Track 2 systems, as national resources. It envisioned that, in the 2006-2010 time frame, Track 2 systems would provide 500+ teraflops (TF) of *peak* computing power and a Track 1 system would provide a *sustained* performance approaching 1 petaflop (PF) on a broad range of science and engineering applications.[8]

NSF awarded funding for Track 2 systems to the Texas Advanced Computing Center (TACC) in 2006 (Sun system with a peak performance of 579 TF) and the University of Tennessee/Oak Ridge National Laboratory in 2007 (Cray system with a peak performance of 1,028 TF). NSF announced the award of a Track 2 system to Pittsburgh Supercomputing Center in 2008. Unfortunately, the downturn in the economy led to the demise of the selected computer vendor, Silicon Graphics, Inc., which was acquired by Rackable Systems. Rackable Systems subsequently changed its name to SGI but cancelled the on-going contract negotiations with PSC. So, a third Track 2 system has not been deployed, despite clear evidence of a need for additional computing resources in the national allocation process run by NSF.

To complement the Track 2 systems, NSF has also deployed a number of experimental and specialized computing systems to serve the nation's scientists and engineers. These include

---

[6] "From Desktop to Teraflop: Exploiting the U.S. Lead in High Performance Computing," NSF Blue Ribbon Panel on High Performance Computing, Lewis Branscomb (chairman), NSF 93-205, August 1993.

[7] NSF supports the acquisition of computer systems at the lower levels of the Branscomb pyramid through many other programs, e.g., the Major Research Instrumentation (MRI) program. See: http://nsf.gov/pubs/2010/nsf10529/nsf10529.htm.

[8] The peak performance of a computer system is the theoretical limit of its computing capability; it can never be achieved. The sustained performance of a computer is the performance that is actually achieved on a given science or engineering application. Although peak performance is used as a proxy for sustained performance, the correlation can be very weak.

the many-core system deployed at the University of Illinois at Urbana-Champaign and another under development at the Georgia Institute of Technology, the data system being deployed at the San Diego Supercomputing Center, the experimental grid test-bed system at Indiana University, and the visualization systems at the University of Tennessee/Oak Ridge National Laboratory and the Texas Advanced Computing Center.

In August 2007, NSF announced that it had selected the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications (NCSA), IBM Corporation, and the Great Lakes Consortium for Petascale Computation to develop and deploy the Track 1 system called Blue Waters[9] by July 1, 2011. Blue Waters is based on the most advanced technologies under development at IBM. These technologies are embodied in PERCS (Productive, Easy-to-Use, Reliable Computing System), which IBM is developing with funding from DARPA's High Productivity Computing Systems (HPCS) program. Blue Waters will be the first production deployment of PERCS and will be a truly extraordinary resource for science and engineering research.

Blue Waters will have more than 300,000 compute cores, 1 petabyte of main memory, 10 petabytes of user disk storage, and 500 petabytes of archival storage. It will have an innovative low latency, high bandwidth communications network that will facilitate scaling to large numbers of compute cores, and an I/O subsystem that will enable the solution of the most challenging data-intensive problems. With a peak performance of approximately 10 petaflops, performance analyses indicate that Blue Waters will sustain 1 petaflops or more on a broad range of science and engineering applications.

The breakthroughs enabled by the extraordinary computing capabilities of Blue Waters will revolutionize many areas of science and engineering. In the past two years, NSF has awarded allocations of time or provisional allocations of time to eighteen (18) research teams from thirty (30) institutions across the country, with more to follow in future years. Research to be carried out on Blue Waters covers all areas of science and engineering from astronomy through biology, chemistry and materials science to geosciences and social and behavioral sciences.

The Blue Waters Project is based on a 24-year partnership between the state of Illinois, the University of Illinois at Urbana-Champaign, and the National Science Foundation. To ensure the success of the Blue Waters Project, the state of Illinois agreed to provide a new state-of-the-art, energy efficient facility to house Blue Waters. In addition, the University of Illinois at Urbana-Champaign is making substantial investments in the development of software for Blue Waters—collaborating with IBM and the Great Lakes Consortium to: (*i*) enhance the systems software for Blue Waters, (*ii*) develop software and tools to facilitate the development of science and engineering applications for Blue Waters, and (*iii*) aid scientists and engineers in rewriting their applications to obtain maximum performance on Blue Waters. In addition, previous investments by the state of Illinois in I-WIRE,[10] a high performance communications infrastructure connecting the major research universities and laboratories in Illinois, provides the transport mechanism for connecting Blue Waters to national research and education networks.

---

[9] See: http://www.ncsa.illinois.edu/BlueWaters.

[10] See: http://www.iwire.org/.

Status of High Performance Computing Cyberinfrastructure. I will discuss the status of computer hardware and software for high performance computing separately as the issues are distinct, if interconnected.

Computer Hardware. NSF has been successful in deploying new computing systems that are delivering extraordinary value for the U.S. research community—*the first system delivered to TACC exceeded the total computing capacity of NSF's TeraGrid by a factor of more than 5*. However, the focus of these acquisitions was on the delivery of raw computing cycles and the funding available to provide support for the users of these new high performance computer systems was limited. This is unfortunate because this approach favors those scientists and engineers who are already using supercomputers and need little assistance, while our experience at NCSA and that at many other centers indicates there is a growing need for high performance computing resources in almost all fields of science and engineering. Without adequate user support, it will be difficult for these new researchers to make effective use of the available resources. High quality support staff is one of the most valuable resources in NSF's supercomputing centers and a fully funded user support program is needed.

Both the Track 1 and Track 2 awards were made through open competitions that included the existing centers as well as many new entrants. The outcome of these competitions is that the two Track 2 awards went to new centers—the Texas Advanced Computing Center and University of Tennessee/Oak Ridge National Laboratory. This is not necessarily bad, although it represents a lost of significant capability at San Diego Supercomputer Center and Pittsburgh Supercomputing Center. At this point the long term impact of the loss of funding on SDSC and PSC is unknown, but the potential loss of expertise at these sites is of great concern to the computational science and engineering research community.

It should also be noted that the prospect of continual competitions has a corrosive effect on the staff at the centers—it is not only difficult to hire quality staff with funding that only lasts for 4-5 years, but enormous amounts of staff time have to be dedicated to preparing for the competitions, rather than assisting researchers. The advantages of competition must be carefully balanced against those of stability in NSF's supercomputing centers program.

The above problems have been extensively discussed by the Task Force on High Performance Computing. It is clear that stability and sustainability are critical if NSF's supercomputing centers are to attract high quality staff who can advance the use of high performance computing across the frontiers of science and engineering. Increased stability of the supercomputing centers that NSF supports, coupled with a rigorous review process to ensure operational quality, will certainly be one of the major recommendations from the Task Force. For additional thoughts on this topic, see the published comments by Larry Smarr[11] and Sid Karin,[12] the founding directors of NCSA and SDSC, respectively.

---

[11] "The Good, the Bad and the Ugly: Reflections on the NSF Supercomputer Center Program," Larry Smarr, HPCWire, January 4, 2010 (http://www.hpcwire.com/features/The-Good-the-Bad-and-the-Ugly-Reflections-on-the-NSF-Supercomputer-Center-Program-80658282.html).

[12] "Thoughts, Observations, Beliefs & Opinions About the NSF Supercomputer Centers," Sidney Karin, HPCWire, January 28, 2010 (http://www.hpcwire.com/features/Thoughts-Observations-Beliefs-Opinions-About-the-NSF-Supercomputer-Centers-82972987.html).

Computer Software. During my two years as Assistant Director for Scientific Simulation in DOE's Office of Science, I played a central role in crafting its *Scientific Discovery through Advanced Computing* (SciDAC) program. This program highlighted the intimate connection between hardware and software and sought to advance computational modeling and simulation through balanced investments in these two areas. Experiences from this program, as well as DOE's ASCI program clearly show that advancing our ability to model complex natural systems requires as much, if not more, investment in software than in hardware.

This problem is actually more acute now than when the SciDAC program was initiated. Since 2004, because of rapidly increasing thermal loads, the speed of a single compute core has not increased. Instead, computer vendors are adding additional cores to the chips and running the chips at lower speeds (to reduce the heat load). As a result, most laptops now use at least dual-core chips and quad-core chips are found in large compute servers, with eight-core chips now available from Intel and IBM. This trend has two major impacts:

1. *Science and Engineering Applications*. In the future, increases in the performance of computational modeling and simulation codes will only be achieved through the use of larger and larger number of processors. Although this "scalability" problem has been with us for nearly twenty years, for much of that time its impact was not felt because of the dramatic increases in the performance of single cores—a factor of two orders of magnitude from 1989 to 2004. With single core performance now stalled, computational scientists and engineers must confront the scalability problem head on.

   The need for ever more scalability has increased the difficulty of developing science and engineering applications for high performance computers. At the heart of the problem is algorithms that scale well to large numbers of compute cores. This problem can only be solved through inspired research. But, even given the appropriate algorithms developing science and engineering applications for computers with hundreds of thousands of compute cores, hundreds of terabytes of memory and tens of petabytes of disk storage is challenging. The software must be written, debugged, optimized and, to the extent possible, made resilient to computer faults (e.g., the loss of a compute core)—none of which is easy or straightforward. Progress will require the creation of new software development tools or the revision of existing tools (compilers, debuggers, libraries, performance analysis tools, etc.) and integration of these tools into a robust, easy-to-use application development environment.

2. *Computing System Software*. Although computer companies provide the base computing system software for high performance computers, enhancements to this base software can greatly facilitate operation, control and use of the system. This problem is becoming more acute as the computer systems become larger and more complex. Recently, a large international group of computer and computational scientists has come together to discuss plans for the development of software for petascale and exascale computers.[13] They are exploring how laboratories, universities, and vendors can work together to coordinate the development of a robust, full featured software stack for petascale and beyond computers.

[13] See http://www.exascale.org/iesp/Main_Page.

The development of high performance computing software—science and engineering applications and computing systems software—is a topic that is being heavily discussed in several NSF Task Forces (Grand Challenges, Software and Tools, High Performance Computing, and Data). What is clear is that the current approach to developing a high performance computing software stack is too fragmented. The Task Forces have noted the need for long term, multi-level efforts in high performance computing software that involves all of NSF's directorates and the Office of Cyberinfrastructure. A partnership between OCI and the Computer & Information Science & Engineering directorate would help create software to manage, control and operate petascale and beyond computers as well as the new tools and software development environment needed to develop science and engineering applications for these computers. A partnership between OCI and the other directorates at NSF would foster the development of a new generation of science and engineering applications that can take full advantage of the power of petascale and beyond computers and realize the promise of these extraordinary resources for advancing science and engineering. Such partnerships already exist, e.g., the *Accelerating Discovery in Science and Engineering through Petascale Simulations and Analysis* (PetaApps, NSF 08-592) program, but they could be substantially strengthened.

High Performance Computer Vendors. There is one last concern that deserves to be mentioned—the dwindling number of supercomputer vendors in the U.S. Just a few years ago, five companies were involved in the development and deployment of supercomputers: IBM, Cray, Sun, SGI and Hewlett-Packard. Sun has now been subsumed by Oracle and SGI has been taken over by Rackable Systems. Although the long term consequences of these actions are not yet known, it is unlikely that Oracle and Rackable Systems/SGI will have as strong an interest in supercomputing as the original companies. Of the remaining companies, only IBM and Cray are actively involved in research and development on supercomputing. Although you would have to talk with these companies to better understand the issues surrounding this situation, it is clear than the supercomputing industry in the U.S. is not as healthy as it was just a few years ago.

Advanced Information Systems

One of the most exciting research advances in science and engineering in the past decade is the digitization of observational science. Fields as disparate as astronomy, biology and environmental science are being revolutionized by the use of digital technologies: digital detectors like those in digital cameras in astronomy, highly automated sequencers in biology, and sensor arrays in environmental science. Data-driven discovery requires sophisticated, advanced information systems to collect, transport, store, manage, integrate and analyze these increasingly large amounts of invaluable data. The knowledge gained from data-driven discovery is already transforming our understanding of many natural phenomena and the future is full of promise.

National Observatories. National astronomy observatories are major investments in the NSF research portfolio. At the leading edge of this portfolio are the latest additions to the NSF's list of approved major research equipment and facilities: the Atacama Large Millimeter Array[14] (ALMA) and the Advanced Technology Solar Telescope[15] (ATST). In

[14] See: http://www.almaobservatory.org/.

addition, two other observatories are in the planning phases: the Giant Segmented Mirror Telescope[16] (GSMT), which will operate in the ultraviolet to the mid-infrared with unprecedented resolution and sensitivity, and the Large-aperture Synoptic Survey Telescope[17] (LSST), which will be able to image faint astronomical objects across the sky, including objects that change or move.

NCSA is heavily involved in the LSST project and has been designated as the main storage and distribution site for all of the data produced by the telescope's 3.2 gigapixel camera. The data challenges to be faced by the LSST are typical of next generation optical telescopes, although the data-processing needs of the Square Kilometer Array (SKA) radio-telescope will dwarf those of the LSST. The LSST will produce more than 15 terabytes of data per night, yielding several petabytes of data per year, and 200 petabytes over its lifetime. This data rate, when combined with the need for real-time analysis to identify and characterize changing or moving objects as well as traditional data mining on petabyte-size data sets, requires a new approach to data management, automated processing, and analysis. Although the telescope will not see first light until 2014, NCSA is already working with other partners in the LSST project to design the cyberinfrastructure needed to meet these challenges.

Several national-scale environmental observatories are also major initiatives in the current NSF research and development portfolio. These are represented by the Ocean Observatory Initiative[18] (OOI), which is leading the way in this space, along with the National Ecological Observatory Network[19] (NEON), and the WATERS Network.[20] Ecological observatories have been in existence for many years with one of the oldest large-scale observatories being the Long-Term Ecological Research Network,[21] although the grand challenges being addressed and the level of integration required for the new observatories far exceeds those of earlier observatories.

Environmental science often depends upon observations from multiple observatories not only of the same type but also complementary observatories. For instance researching the effects of climate change on a terrestrial species might include temperature, rainfall and other traditional measurements from the region being studied, but it might also include ocean temperature, and tidal and current flow data that may directly or indirectly influence the region, and it may also include weather patterns and pollution counts, all of which may be derived from observatories geographically far away that are owned and operated by other organizations. The ability to interact with and integrate data from multiple observatories that cross scientific, geographical, and administrative domains is an increasing requirement for environmental scientists today and presents a number of additional challenges with respect to coordination, standardization, and long term support for deployed cyberinfrastructure.

---

[15] See: http://atst.nso.edu/.

[16] See: http://www.gsmt.noao.edu/.

[17] See: http://www.lsst.org/lsst.

[18] See: http://www.oceanleadership.org/programs-and-partnerships/ocean-observing/ooi/.

[19] See: http://www.neoninc.org/.

[20] See: http://www.watersnet.org/.

[21] See: http://www.lternet.edu/.

Environmental observatories share many of the same general needs with other science domains including data storage and management, application codes, workflow systems to coordinate their research activities, and collaboration tools. However, it is the challenge of supporting potentially thousands of highly variable *in situ* sensors along with the need to manage and share them across vast geographical distances and administrative boundaries that makes environmental observatories unique.

The proposed Genome 10K project[22] is an example of the future of genomic research. The authors of this project, which includes scientists from across the world, are proposing to dramatically increase the number of vertebrate genomes available to the research community. This is made possible by a dramatic drop in sequencing costs coupled with a corresponding increase in computing capability. The Genome 10K Community of Scientists propose to assemble and sequence a collection of some 16,203 representative vertebrate species spanning evolutionary diversity across living mammals, birds, non-avian reptiles, amphibians, and fishes. This will allow scientists, for the first time ever, to carry out a comprehensive studies of vertebrate evolution. Just as computers enabled the assembly and annotation of the human genome, supercomputers will be required to manage and analyze massive quantities of genomic data to achieve the goals of the Genome 10K project.

Status of Cyberinfrastructure for Data-driven Discovery. The development of cyberinfrastructure for data-driven discovery is in its infancy. Within NSF, most of the activity in this area is being driven by large Major Research Equipment & Facilities Construction (MREFC) projects. Each of these projects is developing the cyberinfrastructure needed to accomplish its mission, relying to some extent on the cyberinfrastructure developed in other projects but often redeveloping cyberinfrastructure capabilities in slightly different guises. Since one of the major issues associated with cyberinfrastructure is the ongoing support and maintenance costs associated with the software, sharing cyberinfrastructure software, wherever feasible, will help keep these costs under control.

More recently, NSF has created major programs that are focused largely on the development of the cyberinfrastructure needed to support data-driven discovery. These include the iPlant Collaborative,[23] a project aimed at developing cyberinfrastructure to address a number of grand challenges in plant biology (Genotype to Phentotype in Complex Environments, Tree of Life for Plant Sciences, etc.), and DataNet (NSF 07-601), which consists of several projects designed to explore different approaches to organizing, managing and preserving the data being created in scientific and engineering research.

One of the major cyberinfrastructure requirements for data-driven discovery is the availability of the required data storage capacity, computing resources and associated software. Although these needs could often be met by augmenting the resources available at the NSF-funded supercomputing centers, most major data-driven discovery projects, which usually have lifetimes measured in decades, are reluctant to use the centers because of their uncertain future (current Track 2 grants are only for 4 years and funding for the Track 1 system expires in 2016). This is a lost opportunity for leveraging the expertise at and cost efficiency of the supercomputing centers.

---

[22] "Genome 10K: A Proposal to Obtain Whole-Genome Sequences for 10,000 Vertebrate Species," Journal of Heredity, November 6, 2009.

[23] See http://www.iplantcollaborative.org/.

Networking

To first order, the cyberinfrastructure most needed by universities to participate in or benefit from NSF's high performance computing and data-driven discovery projects is adequate network bandwidth linking them to the relevant project sites. The nation's major research universities are partners in Internet2, which provides a national high performance network. In addition, the National LambdaRail, which is also owned by the U.S. research and education community, provides a testbed for research in the development and use of communication technologies. However, this does not mean that all universities and colleges have access to network bandwidth adequate for their participation in or interaction with the big computing and data projects, an imbalance that will become more acute as the data volumes increase.

As comfortable as the situation may be now,[24] at least for the nation's major research universities, the volume of data that will be generated over the next few years in high performance computing and data-driven discovery will far outstrip the capacities of the current networks. For example, many simulations on Blue Waters will generate multiple terabyte data sets with the total amount of data generated in a given project being measured in petabytes. Although NCSA can provide connectivity to Chicago at 100-400 gigabits per second (Gbps), the national networks passing through Chicago (or any other U.S. city for that matter) do not have the capacity to deliver these data streams to the researchers' home institutions. Separate from the capacity issue, the underlying communication architecture, services and networking technologies required by data intensive science are very different from those that support common consumer services. Common carriers have shown little interest in meeting the specialized requirements of scientific research communities.

In this regard it is worthwhile to note the DOE-SC's ESnet is a welcome exception. ESnet connects more than 40 sites across the nation, including all of DOE-SC's major experimental and computing facilities. DOE-SC's new Science Data Network, which is a part of ESnet, provides services that are specifically targeted for data-intensive science. The SDN circuits provide a wealth of services that are invaluable to scientists who need reliable, high performance, end-to-end connections between two or more sites. ESnet received funding under the American Recovery and Renewal Act to develop and deploy a 100 Gbps network linking its open supercomputing centers in California, Illinois and Tennessee. This is the first step toward DOE-SC's vision of a 1 terabit per second (Tbps) network linking its major facilities.

Although communications bandwidth is critical to participating in high performance computing and data-driven discovery, the TeraGrid's Campus Champions program[25] has shown that access to local expertise is also critical. This program supports individuals on university campuses who are knowledgeable about the TeraGrid and who can help faculty and students apply for and make use of the resources and services available through the

[24] Some scientists note that the current "favorable" situation is deceptive. Because of bandwidth limitations, they note that many scientists are simply avoiding research practices that would stress the current networks.

[25] See: https://www.teragrid.org/web/eot/campus_champions.

TeraGrid. Such programs are likely to be just as important for data-driven discovery as for high performance computing.

Status of Networking. NSF was one of the pioneers in establishing a national networking infrastructure, e.g., NSFnet and Mosaic (the first web browser, which was created at NCSA). However, its networking infrastructure support programs were eliminated several years ago. So, the nation's scientists and engineers must rely on commercial providers, research and education network providers such as Internet2 and National LambdaRail, and state governments for their communications needs. To date, these entities have been able to provide the bandwidth and connectivity needed by researchers.[24] However, with the major new data-intensive research resources coming on line, this will no longer be adequate.

Another major problem is that, to date, there has been little focus on improving end-to-end networking capabilities, i.e., providing high performance connections between the researcher's desktop or local compute/data cluster and large computing and data sites. Even if it appears that there is adequate network bandwidth between these two end points, a bottleneck, often, but not always, on the researcher's campus dramatically limits the network performance. We need to have a better understanding of the issues affecting end-to-end performance to enable researchers to interact with their ongoing research activities at the major facilities.

There are steps that NSF could take to ensure that researchers in U.S. universities have the networking capacity and policies needed to support their research. NSF could begin by developing a high performance network connecting all of their major research facilities, observatories, and supercomputing centers, interconnecting this network with those serving other major federal research facilities, e.g., ESnet, as needed by the academic research community. There are many advantages that will accrue from connecting NSF's large experimental and observational facilities with its computing and data facilities, especially if the future of these centers were secure. In addition, NSF could undertake pilot projects to obtain a better understanding of the problems limiting high performance end-to-end connections between researchers/small research groups and its major research facilities. This would require close collaboration between groups providing national networking resources and campuses providing the "last mile" connection.

Education

I would be remiss if I did not include a section on education in responding to the Subcommittee's request for information on the state of cyberinfrastructure at U.S. universities. Although not a part of the cyberinfrastructure per se, our ability to advance science and engineering using the national cyberinfrastructure requires a new generation of scientists and engineers who can contribute to and understand the use of the basic technologies involved in cyberinfrastructure and computational science and engineering and who can collaborate with colleagues in other fields to take full advantage of the extraordinary capabilities provided by this infrastructure. We need to define the core competencies important for the next generation of scientists and engineers, followed by the development of implementation plan(s) to affect the needed curriculum and course changes.

The curriculum and course changes required to educate the next generation of research leaders is not obvious. Many schools have established graduate programs in computational

science and engineering that supplement study in a discipline with courses in computer science and engineering and applied mathematics; see, e.g., the Graduate Program in Computational Science and Engineering at the University of Illinois at Urbana-Champaign.[26] Such programs are invaluable in preparing students for future careers in computing- and data-intensive fields. But are they sufficient? And what about undergraduate education? At the rate that analog science is becoming digital science, what do we need to teach *all* undergraduates in science and engineering about computing and related technologies to prepare them for life and work in the 21st century. Through its investments in research and education, NSF can serve as a catalyst for this transformation.

In the Blue Waters Project, we are pursuing this goal through the *Virtual School of Computational Science and Engineering*,[27] headed by Professor Sharon Glotzer at the University of Michigan. The *Virtual School* brings together faculty across the universities in the Great Lakes Consortium for Petascale Computation to address the unique opportunities and challenges associated with petascale computing and petascale computing-enabled science and engineering. The *Virtual School* supports the creation and integration of courses and curricula that are tailored to the educational needs of 21st Century scientists and engineers, delivered using 21st century instructional technologies. Although the *Virtual School* is initially targeting graduate-level education, efforts in undergraduate education will follow.

[26] See: http://www.cse.illinois.edu/.

[27] See: http://www.greatlakesconsortium.org/education/VirtualSchool/.