CYBERINFRASTRUCTURE FRAMEWORK FOR 21st CENTURY SCIENCE AND ENGINEERING (CIF21)

Goal

Develop and deploy comprehensive, integrated, sustainable, and secure cyberinfrastructure (CI) to accelerate research and education and new functional capabilities in computational and data-intensive science and engineering, thereby transforming our ability to effectively address and solve the many complex problems facing science and society.

Description and Rationale

Widespread use of a comprehensive CI framework has the potential to revolutionize every science and engineering discipline as well as education. Computing power, data volumes, software, and network capacities are all on exponential growth paths. Highly diverse, multidisciplinary collaborations and partnerships are growing dramatically, greatly enabled by new and emerging technologies, spanning multiple agencies and international domains to address complex grand challenge problems. Scientific discovery is being advanced by linking computational facilities and instruments to build highly-capable simulation models, sophisticated algorithms, software, and other tools and services. CIF21 will enable new approaches to research and education – supporting new modalities such as distributed collaborative networks, allowing researchers to more easily adapt to changes in the research and education process, and providing an integrated framework for people, instruments, and tools to address complex problems and conduct multidisciplinary research. CIF21 will consist of secure, geographically distributed, and connected CI: advanced computing facilities, scientific instruments, software environments, advanced networks, data storage capabilities, and the critically important human capital and expertise.

NSF has a long history of providing leadership for CI and computational science for the U.S. academic science and engineering community, including high performance computing (HPC) systems and the networks and tools to provide access to such capabilities and systems. *Revolutionizing Science and Engineering through Cyberinfrastructure* (2003) set the stage for aggressive efforts led by NSF to build transformative CI. The creation of the Office of Cyberinfrastructure (OCI) provided a focal point for enhancing CI and moving into other important areas such as data interoperability and virtual organizations. The NSF Cyber-enabled Discovery and Innovation (CDI) investments developed new approaches to research and education across the NSF disciplines and in numerous interdisciplinary areas.

NSF has now determined that, in order to realize the potential of CI and computational science to accelerate the progress of science and engineering, it must elevate its coordination and leadership in these areas. The NSF Advisory Committee for Cyberinfrastructure (ACCI) established six task forces in FY 2010 to address critical emerging needs and opportunities, and initiated detailed planning for the development of a framework for expanding the influence of CI. Simultaneously, NSF established an NSF-wide working group to discuss, plan, and coordinate CI programs and activities. These parallel activities led to the development of CIF21, which is designed to leverage previous CI and ongoing CI investments for transformative results across science and engineering. The portfolio has four interconnected components that underpin efforts to realize the potential of computational and data-enabled science:

- Data-Enabled Science
- Community Research Networks
- New Computational Infrastructure
- Access and Connections to Cyberinfrastructure Facilities

In addition to specific activities in each of these four areas, NSF's CIF21 investments will benefit two key domains of high priority for the Nation:

- Matter by Design, which will build on existing strengths in nanotechnology, nanomanufacturing, materials science, mathematical and statistical science, chemistry, engineering, software applications, and investments in programs such as PetaApps and CDI; and
- Research activities in energy, environment, and society, as presented in NSF's Science, Engineering, and Education for Sustainability (SEES) portfolio.

NSF is developing and supporting a systematic and purposeful approach to the creation, development, deployment, and maintenance of cyberinfrastructure. NSF has adopted a "spiral" development approach (for data, software, hardware, etc) that utilizes 3-5-year periods (spirals) leading to new and successively more sophisticated generations of comprehensive cyberinfrastructure to explore and support science and engineering. Establishing effective cyberinfrastructure to handle the volume and range of data in today's world, coupled with the need for more advanced computational expertise and capabilities, requires focus, planning, and long-term coordination. CIF21 will consolidate, coordinate, and leverage a set of CI programs and efforts across NSF to create and establish meaningful infrastructure and develop a level of integration and interoperability of data and tools that is unprecedented. A roadmap for this spiral development path is in process of being finalized and will be shared with stakeholders.

Complex problems in areas of national importance will benefit from creating research and education communities that coalesce around computational and data-enabled challenges and make use of CI and CI-connected infrastructure of all types. CIF21 will establish a broad foundation in computational and data-enabled science and engineering (CDS&E) that supports both disciplinary and interdisciplinary research environments.

Details on each of the four CIF21 components are provided below.

Data-Enabled Science

Data are being generated at prodigious rates across science and engineering, leading to new insights, innovation and discovery. They are already radically transforming science and society. Data-enabled science refers to any science that depends on data; data-intensive science is a subset of data-enabled science and refers to science that uses computational methods to analyze and manipulate data. Because generation, use, curation, and reuse of data are critical components of all science and engineering, CIF21 emphasizes immediate and long-term data support and infrastructure and the development of data-intensive computational algorithms and mathematical and statistical methods including data analytic tools, interoperability, and repositories. CIF21 efforts to facilitate and advance data-intensive scientific and engineering research will be highly responsive to the differing needs of specific research communities, in the context of the overall goal of facilitating the collection, analysis, and retention of data critical to NSF-related research domains. Data acquired or created through simulations, modeling, and analytics provide a basic resource for future research and the opportunity to increase the understanding and involvement of students and the general public in science and engineering.

Within the Data-Enabled Science component of CIF21, three efforts will be initiated in FY 2012. These are data services, data analysis, and data-intensive science.

The data services program will focus on establishing data services, including providing reliable digital preservation, access, integration, and curation capabilities for science and engineering data over a decades-long timeline and serving as component elements of an interoperable data preservation and access network. The program will build on and leverage prior investments in high-end computing, networking, software, algorithms, digital libraries, and domain-specific data systems. Efforts will include

community engagement to develop standards, open access policies, meta-data systems, and ontologies; education and training of a knowledgeable workforce in data; new institutional paradigms for archiving and curation on a national grid; and new global data partnerships able to catalyze and deploy advances quickly. This program will enable transparent access, control, analysis, and synthesis of data and information, while maintaining data integrity and ensuring appropriate security and privacy.

The data analysis program will focus on data analysis efforts and tools that support data mining, manipulation, modeling, simulation, visualization and decision-support systems and will also continuously anticipate and adapt to changes in technologies and in user needs and expectations.

The data-intensive sciences program will support data-intensive scientific and engineering that requires intensive disciplinary efforts and fosters wide ranging, broad programs that build from multiple data domains and areas of expertise to cross disciplinary boundaries and create new algorithms and policies for areas such as data sharing and open access.

Community Research Networks

New cyberinfrastructure tools and changes in the research process have enabled community research networks to address complex, multi-disciplinary problems of societal concern such as competitiveness, security, economic development, and well-being. Community research networks enable people and organizations to perform everyday research functions more effectively by building on and integrating diverse resources, knowledge, and abilities. NSF has a long history of investing in community research networks such as the iPlant collaborative, the Southern California Earthquake Center, the SRS Data Enclave, and the nanoHUB. Cyberinfrastructure links these combinations of people, organizations, instrumentation, physical facilities, computers, data, and software, but few scientists know how to select and assemble these components into a functioning community research network. Focused investments in sociotechnical analyses advance understanding of how to develop virtual organizations, and under what conditions they can foster innovation in science, engineering and education. Such investments are necessary to harness the full potential and promise offered by virtual organizations.

Two efforts are planned for FY 2012 within the Community Research Networks component of CIF21.

The first program will support the establishment of new multidisciplinary research communities to exploit existing and developing computational and data-enabled capabilities to attack scientific challenges that require groups or communities of researchers. To better support and optimize research community networks, appropriate resources must be created for each of three kinds of collaborations: small, mid-level, and large. Small research community networks include two or three researchers working in one or at most two organizations. Mid-level research community networks involve four to seven researchers across three to five organizations. Large research community networks involve eight or more researchers across more than five organizations. Mid and large-scale networks will connect already functioning research groups, potentially including centers.

The second program will focus on advanced research on community research networks. Although the terms “community research networks,” “social networks,” “collaboration,” “virtual organizations,” and “multi-disciplinary research” are used frequently, the underlying processes and optimal structures are not systematically well researched. Relatively little is known about how they unfold, their various structures and forms of leadership, how to provide them with conducive environments, their role in promoting the advancement of science and in addressing societal concerns, and, finally, the role of cyberinfrastructure in supporting and fostering them.

The community research networks that are envisioned as part of the Research Coordination Networks and SEES investments, where highly interdisciplinary, global interactions are essential, will provide a unique

arena for research on scientific networks. Other research efforts include a portfolio of research community networks, approaches to address complex, multidisciplinary distributed grand challenge problems, and offering new ways to collaborate and support transformative research. Advances in this area will allow scientists to collaborate more broadly, more rapidly, and across more dimensions than ever before.

New Computational Infrastructure

A new vision for computational resources and services, from HPC, clouds, clusters, and data centers to focused special-purpose resources and incorporation of sustained software at all levels, all protected and embedded in a rich and robust cybersecure environment, will provide the foundation for supporting innovation and discovery in computational and data-enabled science and engineering. Discipline-specific, as well as cross-discipline software institutes, novel computing platforms, multi-disciplinary data centers, and major computational resources will be expanded to address and support long-term research requirements. Realizing this vision requires attention to the sustainability and extensibility of software, data, and algorithms as well as efforts that ensure robustness while also providing opportunities for upgrades and for introduction of new capabilities.

Two efforts are planned for FY 2012 within the New Computational Infrastructure component of CIF21.

The first program will focus on establishing new, innovative computational and data-enabled resources that will leverage and expand the existing HPC sustainable program. This includes innovative computing environments (e.g., GPGPU, clouds) as well as innovative data sharing and archiving systems and approaches such as shared distributed file systems and services. Other resources needed to advance science and engineering will be developed including a focus on leveraging existing and planned computational capabilities with a plan toward interoperability. And once the computational or data resource becomes operational, it will be linked and integrated with other CI resources (such as eXtreme Digital (XD)) to significantly expand the national and global cyberinfrastructure ecosystem. These activities will also provide an important step towards more deeply integrating campus and national CI (including projects funded through NSF's Major Research Equipment and Facilities Construction (MREFC) account), making it easier to address scaling problems as well providing deeper connectivity between researchers at different sites and in different communities.

The second effort will leverage and expand the activities of the existing Software Infrastructure for Sustained Innovation (SI2) program. This effort will focus on the development of new software tools and services across multiple science disciplines. It also will focus on software as a service; application interfaces, workflows, middleware, testing, evaluation, deployment models and sustainability. It should be noted that a significant efforts in education will required, from undergraduates to postdocs to faculty as CDS&E is essentially new ground for the academic community. While such activities will benefit individual research groups, emphasis will be placed on discipline-specific activities that impact communities, providing services that also may integrate with the national CI investments such as MREFCs, XD, and SI2.

Access and Connections to Cyberinfrastructure Facilities

Many NSF research communities are already organized for conducting research around major pieces of infrastructure. The Global Environment for Network Innovations (GENI) is exploring network architectures and models to support next generation science and research; these new models will inform the development and deployment of leading edge network connections and access. Improved access and connections to facilities and scientific instruments and resources will enable computational communities built around emerging national data- and compute-intensive facilities, such as the National Ecological Observatory Network (NEON), Ocean Observatories Initiative (OOI), EarthScope, Network for Earthquake Engineering Simulation (NEES), and iPlant. Effective use of networks of remote instruments

(e.g., Arctic Observing Network and Polenet) and access to large databases by remote users are essential and require research and development for user-control and interactive remote steering. Research in this topical area also includes work on connectivity to widely distributed sensors, diverse data collections, and geographically remote instruments, where the challenge is not simply one of adequate bandwidth, but of providing at-speed secure connectivity to campuses and labs for researchers and students.

Within the Access and Connections to Cyberinfrastructure Facilities component of CIF21, two programs are planned for FY 2012.

The first program in this CIF21 component is a network connections and engineering program that combines both new and upgraded network connectivity with advancements in deployed networking technology drawn from academic research, commercial development, and engineering. This program has three elements: access to facilities and instruments, networking for the campus and researcher, and integration.

Access to facilities and instruments will address capacity requirements for existing and new facilities and instruments that are being driven by a non-linear increase in scientists' ability to capture data, (for example, through the Large Synoptic Survey Telescope).

Networking for the campus and researcher will focus on investments in the "last mile" connections and building upgrades (including end-to-end activities) to ensure researchers and students have adequate access from their office desktops and not just from centrally-located campus facilities.

Integration will support end-to-end networked cyberinfrastructure through integration activities, including transitioning successful research to development, deployment, and broad scale use.

The second program will expand the efforts in cybersecurity (including identity management) and will leverage the considerable research efforts NSF already supports. The program will focus on moving cybersecurity from innovation to practice and early deployment as part of NSF's support for CNCI.

Management, Assessment, and Funding

Activities in FY 2012 will include enhancing ongoing programs, integrating programs, and issuing new solicitations for CIF21. The portfolio of investments will continue to be led by a senior NSF leadership team and coordinated by an implementation group of senior managers. Advancing the CIF21 vision is a Foundation-wide priority that is reflected not only by the participation of the different units identified in the table below, but also by a management strategy that coordinates these diverse investments in a way that supports science by leveraging the benefits of centralized and scalable cyber-resources. This will be achieved by enlisting the participation of program officers across the Foundation who are members of the CIF21 Working Group (CIF21 WG) to help manage CIF21 solicitations and provide direct linkages to each directorate and office. The CIF21 WG and the senior NSF leadership team are developing performance metrics for each major component of CIF21 and a process to assess and evaluate the programs and efforts of CIF21; these will be finalized in FY11 and put into place before FY12 execution. The NSF Advisory Committee for Cyberinfrastructure (ACCI) will review the CIF21 program and provide external oversight.

Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21) Funding

(Dollars in Millions)

	BIO	CISE	ENG	GEO	MPS	SBE	OCI	OPP	IA	Total, NSF
Data-Enabled Science	$3.00	$7.00	$5.00	$7.00	$10.90	$4.00	$10.00	$3.00	-	$49.90
Community Research Networks	1.00	-	-	2.00	-	3.00	2.00	1.00	-	9.00
New Computational Infrastructure	1.00	9.00	3.00	5.00	9.10	-	5.00	-	11.00	43.10
Access and Connections to Cyberinfrastructure Facilities	1.00	-	1.00	2.00	-	5.00	6.00	-	-	15.00
Total, CIF21	**$6.00**	**$16.00**	**$9.00**	**$16.00**	**$20.00**	**$12.00**	**$23.00**	**$4.00**	**$11.00**	**$117.00**