Cyberinfrastructure Framework for 21st Century Science and Engineering

VISION

1. Introduction

Cyberinfrastructure is rapidly advancing science, and changing its conduct. From its vantage point atop Cerro Pachón in Chile, an unprecedented new telescope will reveal billions of previously unknown celestial objects and probe the deepest mysteries of the cosmos. An earlier project, the Sloan Digital Sky Survey, collected more data in its first weeks than had been collected in the history of astronomy; this new telescope will exceed that decade-long survey in its first week, discovering potentially millions of new events every night. Astronomers will not take turns peering through this telescope; instead, astronomers---and *all* interested researchers, educators or students---will witness events by accessing data. Just one of its events will present a rare opportunity to analyze a sudden supernova or gamma-ray burst, *simultaneously* through many different windows of optical, radio, cosmic ray, and gravitational wave astronomy—each in the domain of historically separate communities. In response, a diverse team of international scientists forms a spontaneous "Grand Challenge Community" to share data, software and preliminary analyses in real-time. Scientists analyze the event through computer simulations, incorporating expertise from astronomy, physics, mathematical and computer sciences, parallel computing, data analysis, high speed-networking, and visualization. Students and teachers in classrooms around the world participate with delight, observing the same event in their instantly updated science eTextbook, where a star that wasn't visible yesterday suddenly outshines an entire galaxy.

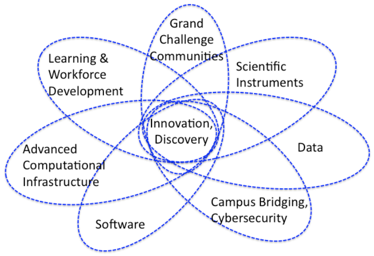
This example is by no means unique. Across the full range of NSF-supported fields increasingly sophisticated instrumentation and expanded computational resources are opening new windows onto phenomena from the universe to the human brain, from the largest scales to the smallest. Across all domains, data play the key role in a profound transformation of the culture and conduct of science and society. Scores of petabytes of data per year emerge from the Large Hadron Collider, used by thousands of scientists; similar data rates are seen in an individual biologist's laboratory. Citizens, scientists and educators alike now *communicate by sharing data*, not only raw data, but in the form of email, software, publications, reports, simulations and visualizations. Coupled with appropriate policy and infrastructure development, these kinds of networked activities can create new capabilities for collaborations at multiple scales, from individuals to communities, to address far more complex problems of science and society than previously possible. This revolution will transform research, practice, and education in science and engineering, as well as advance innovation in society.

This vision of the near future shows clearly the urgent need for a comprehensive, scalable, cyberinfrastructure that bridges diverse scientific communities and integrates high-performance computing, data, software, and facilities in a manner that brings theoretical, computational, experimental, and observational approaches together to advance the frontier. Seizing these opportunities and meeting these challenges is the fundamental purpose of NSF's Cyberinfrastructure Framework for the 21st Century (CIF21).

2. Vision, Goals and Strategies

CIF21 contributes to NSF's overall strategic objectives,[1] by supporting the creation of advanced cyberinfrastructure, including high performance computing systems, data storage systems and repositories, advanced instruments, and visualization systems, enabling researchers to gain new insights and investigate ever broader and more complex research challenges. CIF21's cross-community and multipronged approach will direct those investments along a new path toward a comprehensive, integrated, sustainable, and secure cyberinfrastructure (CI) that will accelerate research and education and new functional capabilities in computational and data-enabled science and engineering (CDS&E). CIF21 also addresses the engagement and education of students from groups traditionally underrepresented in science: African Americans, Hispanics, Native Americans and indigenous people, persons with disabilities, and women.

CIF21: cyberinfrastructure as an ecological system


Grand Challenge Communities
Learning & Workforce Development
Scientific Instruments
Innovation, Discovery
Advanced Computational Infrastructure
Data
Campus Bridging, Cybersecurity
Software

The overarching goals of CIF21 are:

1. Develop a deep symbiotic relationship between science and engineering users and developers of cyberinfrastructure to simultaneously advance ***new research practices and open transformative across opportunities*** all science and engineering fields.
2. Provide an ***integrated*** and ***scalable cyberinfrastructure*** that leverages existing and new components across all areas of CIF21 and establishes a national data infrastructure and services capability.
3. Ensure long-term ***sustainability*** for cyberinfrastructure, via ***community development***, ***learning and workforce development*** in CDS&E and ***transformation of practice***

To balance the need for both singular and integrated advances, CIF21 employs five strategies.

- First, NSF will foster scientific discovery and innovation and build communities capable of leveraging state-of-the-art CI to advance science. This strategy will support activities that gather and manage collective requirements, identify common CI needs across a broad range of scientific domains, and develop trust through community building.
- Second, infrastructure will be deployed in a coordinated way to ensure integration, balancing current and future needs.
- Third, developing advanced CI will require foundational research in both the core CI components and the science related to their use.
- Fourth, CIF21 will ensure the long-term sustainability of CI by supporting development of a trained diverse workforce able to maintain and leverage innovations.

[1] See "Empowering the Nation Through Discovery and Innovation: The National Science Foundation Strategic Plan for Fiscal Years (FY) 2011-2016" at http://www.nsf.gov/news/strategicplan/index.jsp

- Fifth, CIF21 will transform the conduct of science by influencing norms and incentives related to community management of CI resources, recognition of the role of CI development in scientific research, and the development of career paths for computational scientists.

3. The way forward

3.1 Scientific discovery, innovation and community building through CI

Scientific discovery and innovation through cyberinfrastructure, referred to as computational- and data-enabled science and engineering (CDS&E)[2], enables extensions of both theoretical science through computational modeling and simulation, and experimental and observational science through data-intensive computing. Complex transformational science and engineering problems often cannot be adequately addressed by small groups, without CI investments to facilitate the smooth functioning of intellectually diverse, geographically dispersed teams. Also, NSF will foster change enabling career paths for within the academic research enterprise as part of CDS&E.

In the spirit of CDS&E, CIF21 programs will use the requirements of scientists, engineers and educators to drive the development of new CI, balancing long-term goals against short-term needs. In particular, programs will focus on the use of complex and visionary end-to-end scientific use cases in different disciplines to drive innovation in cyberinfrastructure development and use, with particular emphasis on the involvement of early stage researchers. These efforts will be supported through community-building activities, for instance, by bringing together scientists to identify common cross-disciplinary requirements for CI to enable innovative science. These activities will develop lasting mechanisms for interaction and engagement with processes that funnel defined requirements into research funding streams.

3.2 Building Infrastructure

Fostering innovation requires cyberinfrastructure spanning multiple levels (national, community, campus) as well as multiple infrastructure lifecycle stages. CI should be managed efficiently and equitably, particularly in the allocation of resources (e.g. computing cycles) and with well-established mechanisms to ensure sustainability.

As articulated in the Advanced Computing Infrastructure Vision and Strategic Plan[3], CIF21 will continue to invest in high performance computing (HPC) hardware. It will also enhance support for services, including integration with campus and other national computational resources, including cloud systems and services.

For both data and software, infrastructure investments will focus on the full lifecycle. For data, this includes early stage challenges in acquisition, arising for example from innovations in scientific instruments, to later stage challenges in modeling and visualization. Instruments and facilities, from individual labs to MREFC-scale international facilities, will be considered integral to a national data capacity. Similarly, for software, support will be provided for creating new tools and services in priority areas identified by multi-disciplinary teams as well as standardization and maintenance of existing tools spanning multiple research domains. Policy will also be developed to enable greater sharing and

[2] For further description see http://www.nsf.gov/od/oci/taskforces/TaskForceReport_GrandChallenges.pdf

[3] http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf12051

discoverability of data, software, and scientific literature, accelerating discovery and innovation while enabling communities to work more effectively together to address complex problems.

To ensure a reliable and trusted end-to-end CI, CIF21's Campus Bridging portfolio will support using the campus environment to try out and experimentally deploy cybersecurity innovations as well as enhance access to CI by establishing sustainable models of pairing under-resourced campuses with more advanced neighbors. These investments will also foster integration of campus and national distributed computing environments.

3.3 Foundational Research

CIF21's foundational research in advanced cyberinfrastructure and its applications to all areas of science and engineering will support both technological and scientific advances together.

The need to handle the data that will be involved in individual and Grand Challenge-scale problems alike will require algorithms and software architectures capable of handling both small- and extreme-scale data systems. For software, foundational research investments will target programming paradigms, addressing the use of massively parallel computers, highly distributed computer systems (including private, public, and hybrid clouds), complex file systems (both parallel and distributed), new accelerator architectures and the potentially hybrid systems that will be built from them. These systems also require tools and services for gateways/portals/hubs as well as middleware for dynamic data-driven workflows. Investments will also target domain-specific programming to establish paradigms for verification, validation, uncertainty quantification, and provenance to ensure trustworthy and reproducible scientific findings. Software must also support collaborative science (technologies for teams, data, and computing).

The combination of improved software and enhanced data availability creates opportunities for foundational research in CDS&E targeting enhancements both at the platform and tool levels. At the platform level, NSF will support development of new algorithms to exploit massively parallel and distributed platforms for complex and/or data-intensive computational tasks. NSF will also support development of a wide array of advanced methods and algorithms in discretization, nonlinear solvers, sub-continuum models, statistical methods and theory, language processing, combinatorial computing, optimization, compressed sensing, uncertainty quantification, and integrated sensing-assimilation-simulation-prediction-control.

3.4 Workforce Development

CIF21 investments in workforce development will target the range from data and software literacy for the public, including policymakers, to training and new curricula for providers and developers of data and software services. CIF21 will also support advanced training and education for IT professionals in scientific fields, particularly those who can "harden," support, maintain, evolve, and ensure access to software and data. At all levels, from K-16 and beyond, CIF21 investments, partnered with others across the foundation, will support new data-enabled approaches to education, introducing scientific data from NSF programs and facilities directly into teaching, as illustrated in the introduction. Data-enabled approaches will also allow for much better evaluation and assessment of both student/teach performance and NSF programs alike.

In addition, CIF21 will support broadening participation by underrepresented groups, enhancing curricula, training and internship experiences at the undergraduate, graduate and postdoctoral levels for domain scientists, cyberinfrastructure developers and interdisciplinary students with the goal of producing a diverse scientific workforce capable of developing and using cyberinfrastructure.

In campus environments, education and training will take advantage of the broader campus cyberinfrastructure, explicitly exposing and leveraging those resources in the education process. Such goals are likely best achieved through courses designed for multiple sections of the CI workforce, bringing together domain scientists, computational, mathematical and statistical scientists, and campus IT professionals responsible for research into, delivery of, and support of production CI. Such efforts should also seek to build from these campus level Learning and Workforce Development (LWD) investments to a regional and national network of expertise in order to share knowledge and best practices, educational resources, advocacy networks, outreach and engagement programs, and forums that increase the awareness, access, engagement and inclusion of all students.

CI workforce development will leverage existing programs targeting both the IT workforce in general and those promoting diversity in STEM. Examples include Research Experiences for Undergraduates (REU), Graduate Research Fellowships (GRF), Postdoctoral Research Fellowships, and CAREER awards that are targeted toward training and research in CDS&E and the use of existing and future CI. CIF21 projects also further the NSF goal of preparing tomorrow's innovation workforce that is enriched by the assets of diverse participants from a range of groups and communities. This STEM workforce will engage diverse teams that can offer new ways to solve problems and provide unique perspectives to improve performance and outcomes. To this end, a key feature of projects will be a program strategy and plan for recruitment, mentoring, retention, and graduation of U.S. students (U.S. citizens, nationals, and permanent residents) in NSF-supported STEM fields, with specific efforts aimed at members of groups underrepresented in science and engineering, including women, minorities and persons with disabilities. Where appropriate, CIF21 projects should engage K-12 students through the development of appropriate curricula for CDS&E.

In addition, CIF21 will catalyze a reexamination of the university curriculum by introducing new approaches to teaching and research through CDS&E, in all disciplines, through Expeditions in Education and other new programs, as well as through Transforming Undergraduate Education in Science, Technology, Engineering, and Mathematics (TUES) and Integrative Graduate Education and Research Traineeship (IGERT). To the extent possible, these efforts will also be coordinated with those underway in other federal agencies and take into account programs promoted by the nation’s IT and science industries.

3.5 Transforming Practice

The success of CIF21’s technical and educational investments will rely, to a large extent, on evolution in the scientific community. Accordingly, CIF21 will aim to transform practice by targeting norms and incentives related to community management of CI resources, the role of CI development in scientific research, and career paths for computational scientists.

NSF is an important partner in cyberinfrastructure development; however the sustainability of CI relies in large part on its ongoing management by scientific communities. CIF21 will foster community management by identifying collective and sustainable funding models and policies. Such policies include those encouraging open, sharable data services as well as those addressing lifecycle management, such as support for data centers, repositories, open source models enhancing software reuse, and policies defining public-private partnerships to share the burden of long-term data and software stewardship.

Policies can also influence the role of CI research and development in scientists' careers. NSF will endeavor to influence norms of citation for new forms of publication and scientific expression, including data sets, so that researchers are able to ensure their work is citable, and others are able to discover and access it. The outcomes should target mechanisms for citation of software and datasets as distinct products of scholarship, promoting standards of academic credit and rigor for these CI components. CIF21 will also promote the development and use of metrics that measure software and data usage and their subsequent impact on science, engineering and education.

In summary, CIF21 is driven by science, engineering and education needs, providing opportunities for new discoveries and innovation enabled by new cyberinfrastructure. It lays a foundation for future innovations, providing a platform for communities to come together to address grand challenges through multi-disciplinary approaches; shared cyberinfrastructure; software across disciplines; data management, access, curation, standardization, sharing and policies; advanced computing infrastructure; computational- and data-enabled science and engineering research, and training a diverse workforce to address national needs. It addresses the infrastructure and research needs in all scientific and engineering domains through advances in foundational research and infrastructure deployment, while ensuring sustainability through workforce development and transforming the practice of science.