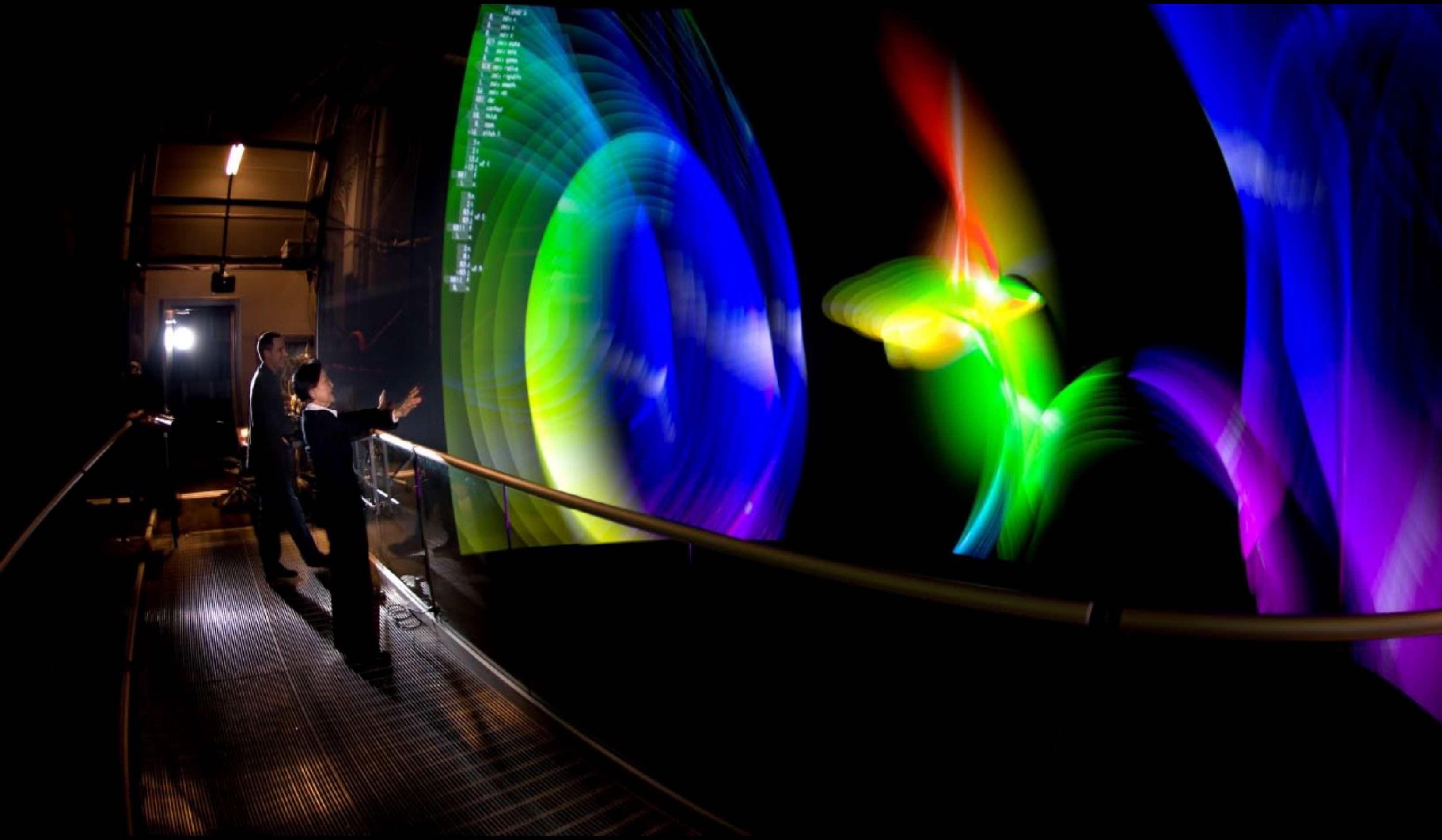
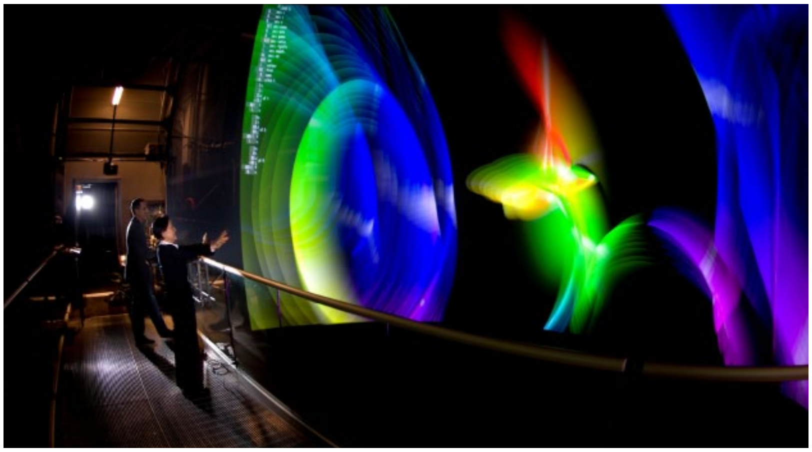
February 2012



Cyberinfrastructure for 21st Century Science and Engineering

Advanced Computing Infrastructure

Vision and Strategic Plan



COVER IMAGE:

As part of the Multimodal Representation of Quantum Mechanics: The Hydrogen Atom project, this image shows people on the bridge of the AlloSphere interacting with the hydrogen atom with spin.

Credit: Professor JoAnn Kuchera-Morin, Media Arts and Technology, UCSB; Professor Luca Peliti, University of Naples, Italy; Lance Putnam, Media Arts and Technology, UCSB; photo by Kevin Steele

CYBERINFRASTRUCTURE FOR 21ST CENTURY SCIENCE AND ENGINEERING (CIF21)

ADVANCED COMPUTING INFRASTRUCTURE STRATEGIC PLAN

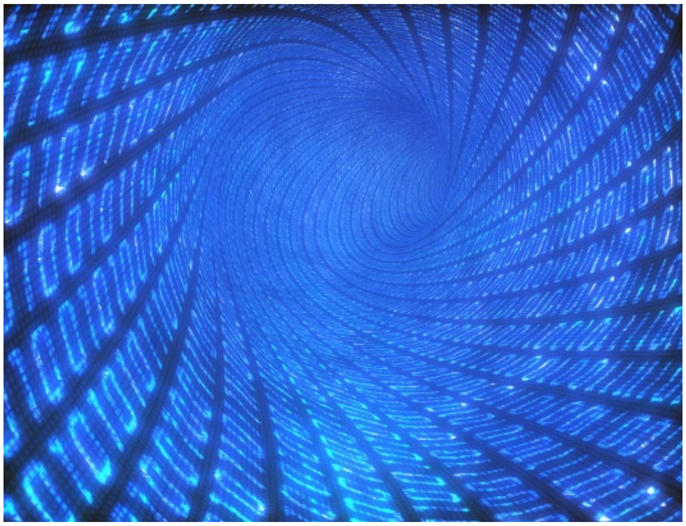
EXECUTIVE SUMMARY

Advanced Computing Infrastructure (ACI) is a key component of the Cyberinfrastructure for 21st Century Science and Engineering (CIF21) framework. While CIF21 addresses broadly the cyberinfrastructure needed by science, engineering, and education communities to tackle complex problems and issues, ACI specifically focuses on ensuring these communities have ready access to needed advanced computational capabilities. The CIF21 framework includes other complementary, but overlapping components: data, software, campus bridging and cybersecurity, learning and workforce development, grand challenge communities, computational and data-enabled science and engineering, and scientific instruments (see Figure 1, page 6, for more detail). Many of these components are beginning to be addressed by CIF21 programs in fiscal year (FY) 2012, and a process is now underway to develop strategic plans for each component.

The National Science Foundation (NSF) has been an international leader in high- performance computing deployment, application, research, and education for almost four decades. With the accelerating pace of advances in computing and related technologies, coupled with the exponential growth and complexity of data for the science, engineering, and education enterprise, NSF requires a new vision and strategy to advance and support a comprehensive advanced computing infrastructure that facilitates transformational ideas using new paradigms and approaches.

The ACI Strategic Plan outlined here seeks to position and support the entire spectrum of NSF-funded communities at the cutting edge of advanced computing technologies, hardware, and software. It also aims to promote a more complementary, comprehensive, and balanced portfolio of advanced computing infrastructure and programs for research and education to support multidisciplinary computational and data-enabled science and engineering that in turn support the entire scientific, engineering, and education community.

The vision and strategies articulated here are derived from numerous discussions within NSF and from input from experts in the community such as that from the six task force reports of the Advisory Committee for Cyberinfrastructure and from the various directorate advisory committees.



The exponential growth and complexity of data requires a new and qualitatively different approach to data storage, stewardship, management, cybersecurity, distribution and access.

Credit: Thinkstock

Smartphones, tablets, gaming systems and new sensors are changing business, education and research.

Credit: Thinkstock

ACI VISION:

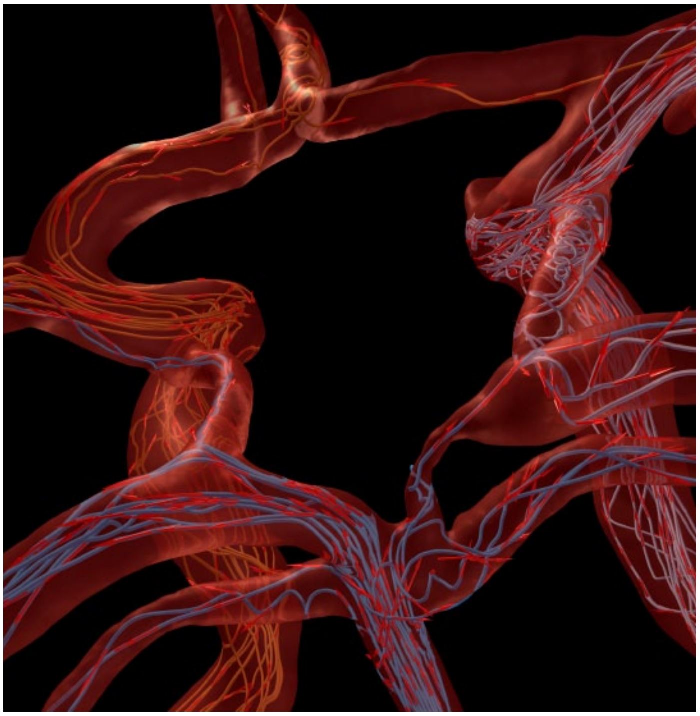
NSF will be a leader in creating and deploying a comprehensive portfolio of advanced computing infrastructure, programs, and other resources to facilitate cutting-edge foundational research in computational and data-enabled science and engineering (CDS&E) and their application to all disciplines. NSF will also build on its leadership role to promote human capital development and education in CDS&E to benefit all fields of science and engineering.

ACI STRATEGIES:

1. Foundational research to fully exploit parallelism and concurrency through innovations in computational models and languages, mathematics and statistics, algorithms, compilers, operating and run-time systems, middleware, software tools, application frameworks, virtual machines, and advanced hardware.
2. Applications research and development in use of high-end computing resources in partnerships with scientific domains, including new computational, mathematical and statistical modeling, simulation, visualization and analytic tools, aggressive domain-centric applications development, and deployment of scalable data management systems.
3. Building, testing, and deploying both sustainable and innovative resources into a collaborative ecosystem that encompasses integration/coordination with campus and regional systems, networks, cloud services, and/or data centers in partnerships with scientific domains.

Second-year mechanical engineering technology students Tim Brogan (left) and Ryan Strand used tablet PCs as part of a pneumatics and hydraulics course. They were part of a study to determine how use of educational technology might enhance learning, improve interaction and engagement with classmates and faculty, and decrease withdrawal rates from the course.

Credit: Michelle Cometa, University News, Rochester Institute of Technology



Human cranial arterial network includes 65 arteries accounting for every artery in the brain larger than 1 millimeter in diameter. Using color and scale to show magnitude, this visualization depicts the flow of blood in the Circle of Willis, a pattern of redundant circulation that maintains the brain's blood supply in case, part of the circle or a supply artery becomes restricted or blocked.

Credit: Greg Foss, Pittsburgh Supercomputing Center

4. Development of comprehensive education and workforce programs, from deep expertise in computational, mathematical and statistical simulation, modeling, and CDS&E to developing a technical workforce and enabling career paths in science, academia, government, and industry.
5. Development and evaluation of transformational and grand challenge community programs that support contemporary complex problem solving by engaging a comprehensive and integrated approach to science, utilizing high-end computing, data, networking, facilities, software, and multidisciplinary expertise across communities, other government agencies, and international partnerships.

INTRODUCTION AND BACKGROUND

Innovative information technologies are transforming the fabric of society and data is the new currency for science, education, government, and commerce. High-performance computing (HPC) has played a central role in establishing the importance of simulation and modeling as the third pillar of science (theory and experiment being the first two), and the growing importance of data is creating the fourth pillar.

Sustainable Harvest, a specialty coffee importer in Portland, Ore., is partnering to develop applications that will equip farmers in developing countries with tools to improve crop and harvest tracking and also give farmers access to educational videos and best practices for improving crop quality. These iPad applications increase traceability and transparency across the coffee supply chain.

Credit: Sustainable Harvest Coffee Importers

The continued growth in the number of cores per chip and accelerator-based hybrid systems requires expanded research and development efforts in new computer architectures, computational models, parallel programming languages, and software development for parallel and distributed systems. It also calls for increased attention to fault-tolerance (resiliency), new operating systems, and run-time systems. Power consumption is currently a key limitation for all sizes of computers. Similarly, memory bandwidth limitations and increased data movement require significantly greater effort in fundamental research in computer science, mathematical, and statistical sciences, engineering, and materials science. Multidisciplinary research and use of data require increased levels of research and development in advanced simulation methods, coupling of complex models, new algorithms, approaches to software and data integrity and resilience, new data analytic and statistical tools, and data management and sustainability. The growth of data-intensive science coupled with multidisciplinary collaboration requires additional effort in existing and new domain-centric applications and tools including software engineering, statistics, and mathematics, broadening use of computational science across all of NSF, and developing the entire CDS&E workforce.

The 2010 President's Council of Advisors on Science and Technology (PCAST) report, "Designing a Digital Future,"[1] points out that floating point operations per second (FLOPS) measurements are not definitive measures for success in HPC and that it is now important "to conduct basic research in hardware, in hardware/software systems, in algorithms, and in both systems and applications software." HPC must encompass the ability to efficiently manipulate and manage vast quantities of data. It must also simultaneously address innovations in software and algorithms, data analytics, statistical techniques, fundamental operating system research, file systems, and innovative domain-centric applications. The new ACI strategies directly address these issues raised in the PCAST report.

Commoditization of both hardware and software is creating an era of significant disruptions. One disruption is the changing nature and role of the private sector in the development of next generation computing and technologies. Advanced computing and data will not be driven by high-end science requirements, but instead by millions of $100 devices (e.g., computer games, cell phones, tablets). Microscale margins will drive manufacturers toward volume, and new technologies will be abandoned if they do not have a demonstrable share of the market necessary to recoup any development and production costs. A second disruption arises from the fact that the ubiquitous availability of a wide range of technologies will fundamentally change the development of many processes and workflows, including the type of algorithms and software that must be implemented for research and education. A third disruption is the emerging transformation of the institutions engaged in the higher education enterprise, as there is increasingly much less connection between researchers and the physical place of their institutions. This will lead to new models for data-intensive science that will be organized dynamically around research questions and domains, and will present new challenges to geographically centered

In this zoomed-in image of the Antennae Galaxies, the generation of super-bright, hot stars that formed when the denser centers of the two spirals first collided shine in white-blue.

Future stars are growing now, concealed in dark clouds into which optical telescopes cannot see. However, ALMA sees through the obscuring dust and traces of these stellar nurseries, many of which show the continuation of the cloud that has been lit pink by a previous generation of new stars. ALMA's millimeter/submillimeter wave test views shown here are represented in orange and yellows to contrast with the previous star birth generations. (Optical images from HST ACS/WFC.)

Credit: (NRAO/AUI/NSF); ALMA (ESO/NAOJ/NRAO); HST (NASA, ESA) and B. Whitmore [STScI]

[1] *PCAST, December 2010, "Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology."*

research efforts, including traditional campuses.

While supercomputers remain a key generator of data, the exponential increase in data from a growing, distributed set of diverse scientific instruments and sensor networks requires a new and qualitatively different approach to data storage, stewardship, management, cybersecurity, distribution, and access. Not only is the data much larger, more diverse, and more distributed, but the needs for data analysis require potentially different computational, mathematical, and statistical approaches and the collaborative nature of research has increased the need for more distributed access.

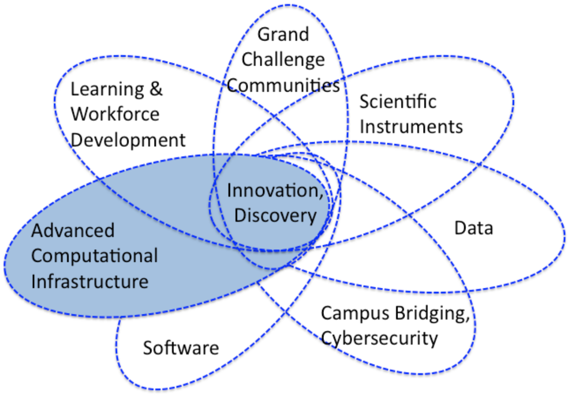
This new NSF ACI vision supports both computational and data-intensive research coming from simulations, scientific instruments, "cloud" computing, and sensors. It is critical that the newly developed ACI ecosystem accommodates traditional national centers as well as those on university campuses, and include supercomputers, local clusters, storage, and visualization systems that can support far more researchers than in the past. Advanced technologies and sustained research in HPC has created a ubiquitous need for advanced digital services across the landscape, from schools and campuses to research centers and industry. Therefore, NSF's vision for advanced computing also must expand to focus on the broader base of CDS&E across multiple domains.

Achieving the ACI vision will advance science and engineering research and education to serve the nation's needs for years to come.

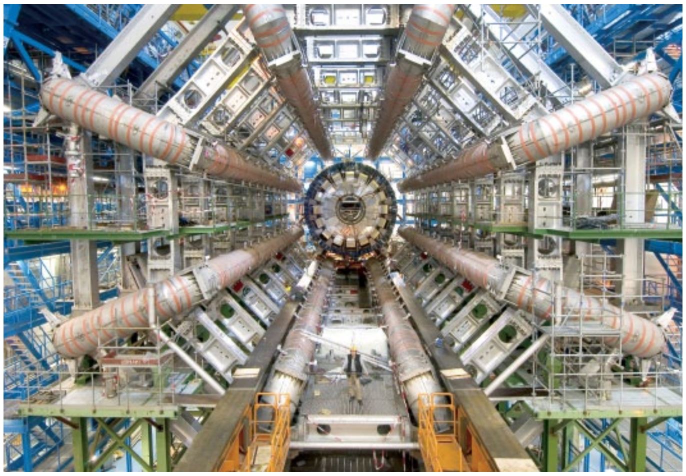
STRATEGIC DIRECTIONS

The NSF ACI strategies are part of the larger NSF CIF21 framework and are not separate or stand-alone efforts (see Figure 1). Although this document focuses on the ACI-specific strategies, it is important to note that the complete CIF21 planning involves an integrative approach to support complex problems and issues addressed by the science, engineering, and education communities. Implementation of the strategies for ACI complements and dovetails with other CIF21 components, including data, software, learning and workforce development, and cybersecurity, as well as with individual directorate and office research and education efforts. CDS&E and grand challenge communities' activities connect with ACI and all components of the CIF21 strategy. These activities are driven and enabled by a coherent approach to developing these components to meet the research and science requirements of the nation.

Cyberinfrastructure Framework for the 21st Century

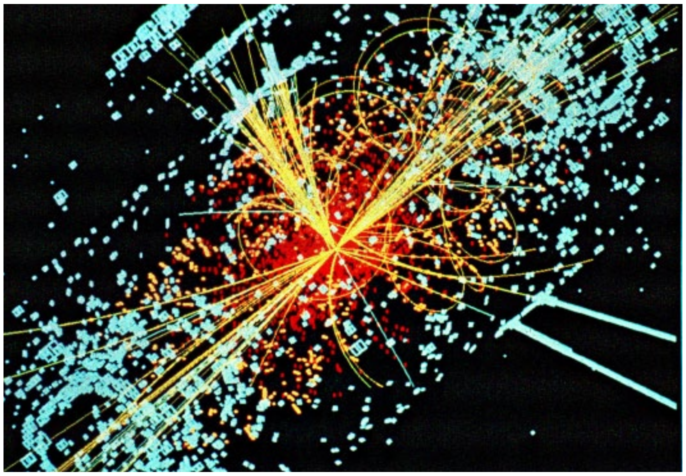

Grand Challenge Communities
Learning & Workforce Development
Scientific Instruments
Innovation, Discovery
Advanced Computational Infrastructure
Data
Campus Bridging, Cybersecurity
Software

1. Foundational research to fully exploit parallelism and concurrency through innovations in computational models and languages, mathematics, and statistics, algorithms, compilers, operating and run-time systems, middleware, software tools, application frameworks, virtual machines, and advanced hardware. This strategy encompasses:
   - Computational models to enable new and transformative ways of "thinking parallel," including new abstractions that account for parallelism and concurrency, and support reasoning about the correctness



The ATLAS detector at CERN.

Credit: CERN



An example of simulated data modeled for the Compact Muon Solenoid (CMS) particle detector on the Large Hadron Collider. Here, following a collision of two protons, a Higgs boson is produced that decays into two jets of hadrons and two electrons.

Credit: TACC

and parallel performance, consisting of communication, energy costs, resiliency, and security;

- Programming languages to enable effective expression of parallelism and concurrency at every scale, including new approaches to developing software, handling messaging and shared memory, and improving programming productivity on parallel and distributed systems;
- Disruptive rethinking of the canonical computing "stack" – applications, programming languages, compilers, run-time systems, virtual machine, operating systems, and architecture – in light of parallelism and resource-management challenges and to support optimization across all layers of the stack from software down to the architecture level;
- New algorithmic paradigms that promote reasoning about parallel performance and lead to provable performance guarantees, while allowing algorithms to be mapped onto diverse parallel and distributed environments, and optimizing resource usage including compute cycles, communication, input-output (I/O), memory hierarchies, and energy;
- Computer software architectures to enable resilient computation at a (or the) large scale, including new operating systems for multicore systems and cloud architectures; file systems and data stores for data-intensive computing; run-time systems to manage parallelism, synchronization, communication, scheduling, and energy usage; and compilers to manage debugging, predictability, power consumption, and security;
- Computer architectures that focus on efficient communication, including interconnection networks, fine-grain synchronization, parallel memory systems, and I/O;

Renci
East Carolina University
Center for Coastal Systems Informatics and Modeling
Renci
Renaissance Computing Institute
at East Carolina University
i.org

Elementary students from Dare County, N.C., measure wind velocity during a National Hurricane Week outreach program with the RENCI East Carolina Regional Engagement Center.

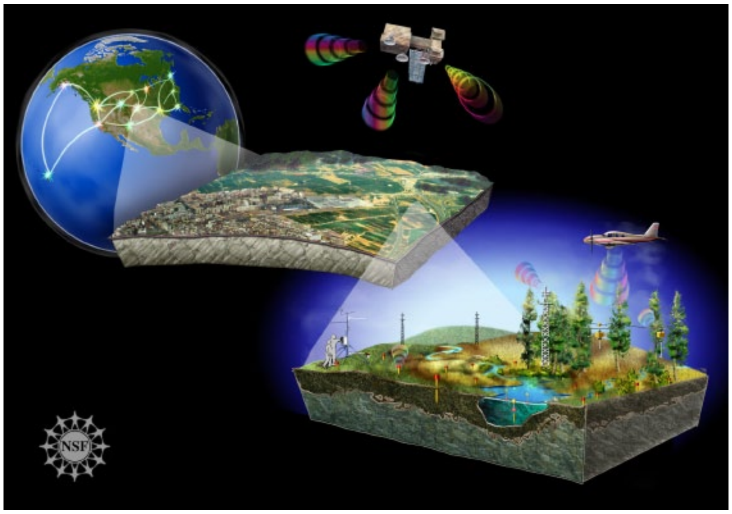
Credit: RENCI East Carolina Regional Engagement Center

Scientists studied the interaction of the Deep Water Horizon oil spill and microbes in Gulf of Mexico waters.

Credit: Luke McKay, University of Georgia

- Research into highly parallel and scalable application-specific and heterogeneous system architectures;
- Algorithms and software architectures capable of handling both small- and extreme-scale data systems and data analytics;
- Fundamental research in mathematical algorithms, statistical theory and methodologies to address the challenges with massive and distributed data.

2. Research and development in the use of high-end computing resources in partnerships with scientific domains, including new computational, mathematical, and statistical modeling, simulation, visualization, and analytic tools, aggressive domain-centric applications development, and deployment of scalable data management systems. This strategy encompasses:
   - A systematic exploration of next-generation science methods, algorithms, and applications in all disciplines, their computational needs, and their mapping on to potential future architectures and approaches to computing;
   - New algorithms to exploit massively parallel and distributed platforms, and for data-intensive computational tasks, as well as methods to decompose existing serial algorithms into faster combinations of serial/parallel/distributed computation;
   - Research into highly parallel and scalable application-specific and heterogeneous system architectures;
   - Focused investments in the development of algorithms, tools, and software that will support all disciplines, especially those that have not utilized parallelism and concurrency capabilities in the past, including science for statistical analysis, data mining, visualization, and simulation, as well as sophisticated



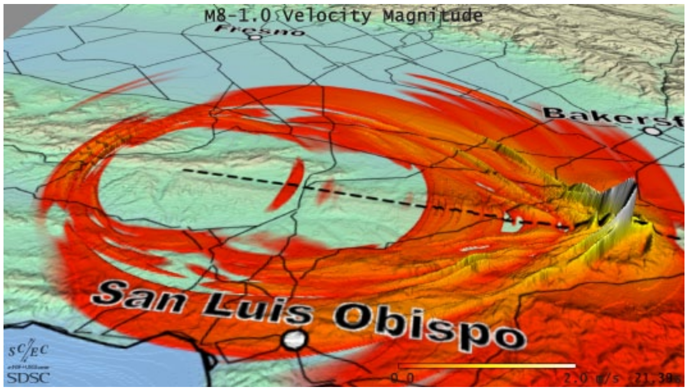
An artist's conception of the National Ecological Observatory Network (NEON) depicting its distributed sensor networks, experiments and aerial and satellite remote sensing capabilities, all linked via cyberinfrastructure into a single, scalable, integrated research platform for conducting continental-scale ecological research. NEON is one of several National Science Foundation Earth-observing systems.

Credit: Nicolle Rager Fuller, National Science Foundation

cyberinfrastructure for discipline-based scientists (such as biologists, geologists, social scientists, education researchers, and economists);

- Integrated end-to-end data pipeline management paradigms harnessing parallelism and concurrency, focused on the entire data path from generation to transmission, to storage, use, and maintenance, all the way to eventual archiving or destruction;
- Development of sustainable data services to provide data mining, statistical analyses, mathematical algorithms, and computational tools to a broad set of researchers, scientists, and educators, and thereby advancing research across a range of other areas including statistical, mathematical, and computational sciences, engineering, and education.

3. Building, testing, and deploying both sustainable and innovative resources into a collaborative ecosystem that encompasses integration/coordination with campus and regional systems, networks, cloud services, and/or data centers in partnerships with scientific domains. This strategy encompasses:
   - A more balanced and sustainable approach to NSF ACI facilities, including support for not only HPC hardware, but also for a broader culture of scientific computing assistance and integrated approaches that go beyond traditional HPC services, including integration with campus and other national computational resources and exploring the growing number and capabilities of cloud systems and services. This approach will also entail a close working relationship with campuses;
   - Development of HPC facilities to be supportive of major scientific data centers being established by scientific domains, and enable storage of legacy data and allow communities to access and integrate such data sets in ways that are currently not possible;
   - Development of capabilities that focus on ACI for the broader science and research communitythat include facilities that all researchers can use and support staff who would be trained and available for consultation, as well as strategic investments in domain-specific ACI centers;
   - Revise the current allocation process to accommodate a broader range of disciplines, better integration with campus infrastructure, and allocation of data resources and storage;
   - The development of a sustainable cyberinfrastructure integrating high- speed, end-to-end transmission with data curation, management, and storage to support communities doing data-intensive science (e.g.., genomics, phylogenomics, phenomics, biodiversity informatics, molecular modeling, economics, social systems, health-informatics, astronomy, astrophysics, Earth system modeling);
   - Alignment of data infrastructure plans with computational infrastructure plans.
4. Development of comprehensive education and workforce programs, from building deep expertise in computational, mathematical and statistical simulation, modeling, and CDS&E to developing a technical workforce and enabling career paths in science, academia, government, and industry. This strategy encompasses:
   - Education and workforce development is needed to support the next generation of computational and applied sciences as ACI and computational



This visualization shows instantaneous ground motions for a magnitude 8 earthquake simulation ‘SCEC M8’ along the San Andreas Fault. The image shows the rupture about 21 seconds after it started in central California, propagating south on the San Andreas Fault. The rupture will continue for nearly two minutes, passing Riverside, Palm Springs, and Indio before stopping south of Bombay Beach.

Credit: Amit Chourasia, San Diego Supercomputer Center, University of California, San Diego and data-intensive science goes mainstream. These efforts may include targeted Advanced Technological Education (ATE), Research Experiences for Undergraduates (REU), Graduate Research Fellowships (GRF), postdoctoral, and Faculty Early Career Development (CAREER) activities, curriculum development, and/or other programs aimed to serve these needs and that should include analysis of the range of new and emerging professional roles and the kinds of training and preparation needed. Such efforts should also be based on, and build evidence about, effective learning of new, complex domains;

- Development of new and diverse course curricula and career learning resources for parallel and distributed computer languages, data-intensive science, and data analytics, , again at levels ranging from the preparation of technicians to postdoctoral scientists to professionals in science, engineering, and education;
- Adaptation and expansion of current programs, e.g., ATE, Scholarship for Service (SFS), Transforming Undergraduate Education in Science, Technology, Engineering, and Mathematics (TUES), Integrative Graduate Education and Research Traineeship (IGERT), and GRF focused on developing technical workforce and career paths including community college and undergraduate education, and interdisciplinary and applied experiences at the graduate and postdoctoral levels;
- New research into how people learn concepts of concurrency parallelism; research and development of effective ways to teach parallelism and distributed computing and data-intensive science, simulation, and modeling; analysis of needed expert knowledge and capabilities in CDS&E, including computational science professional roles and investment in the development of learning progressions to inform curriculum and programs to build that knowledge;
- Investment in undergraduate and graduate education and curriculum development that will prepare the next generation of disciplinary scientists to engage in science with significant CDS&E/computational dimensions, including focus on the role of practica and apprenticeship experiences;
- Emphasis on broadening participation, new strategies for recruitment into undergraduate courses in this area, and development of the senior leadership talent pool.

5. Development and evaluation of transformational and grand challenge community programs that support contemporary complex problem solving by engaging a comprehensive and integrated approach to science, utilizing high-end computing, data, networking, facilities, software, and multidisciplinary expertise across communities, other government agencies, and international partnerships. This strategy encompasses:
   - New emphasis on transformational and grand challenge communities will build on enabling investments in infrastructure (e.g., Blue Waters, Stampede, NCAR, XSEDE, facilities and instruments, data services), core technologies (new methods and algorithms), software institutes, CDS&E (supporting individual investigators or small groups in fundamental approaches to CDS&E), learning and workforce (to develop a new generation of computational scientists), sharing interdisciplinary data across multiple

Advanced cyberinfrastructure provides secure, easy-to-use interfaces with instruments, data, computing systems, networks, applications, analysis and visualization tools and services, to support research and education.

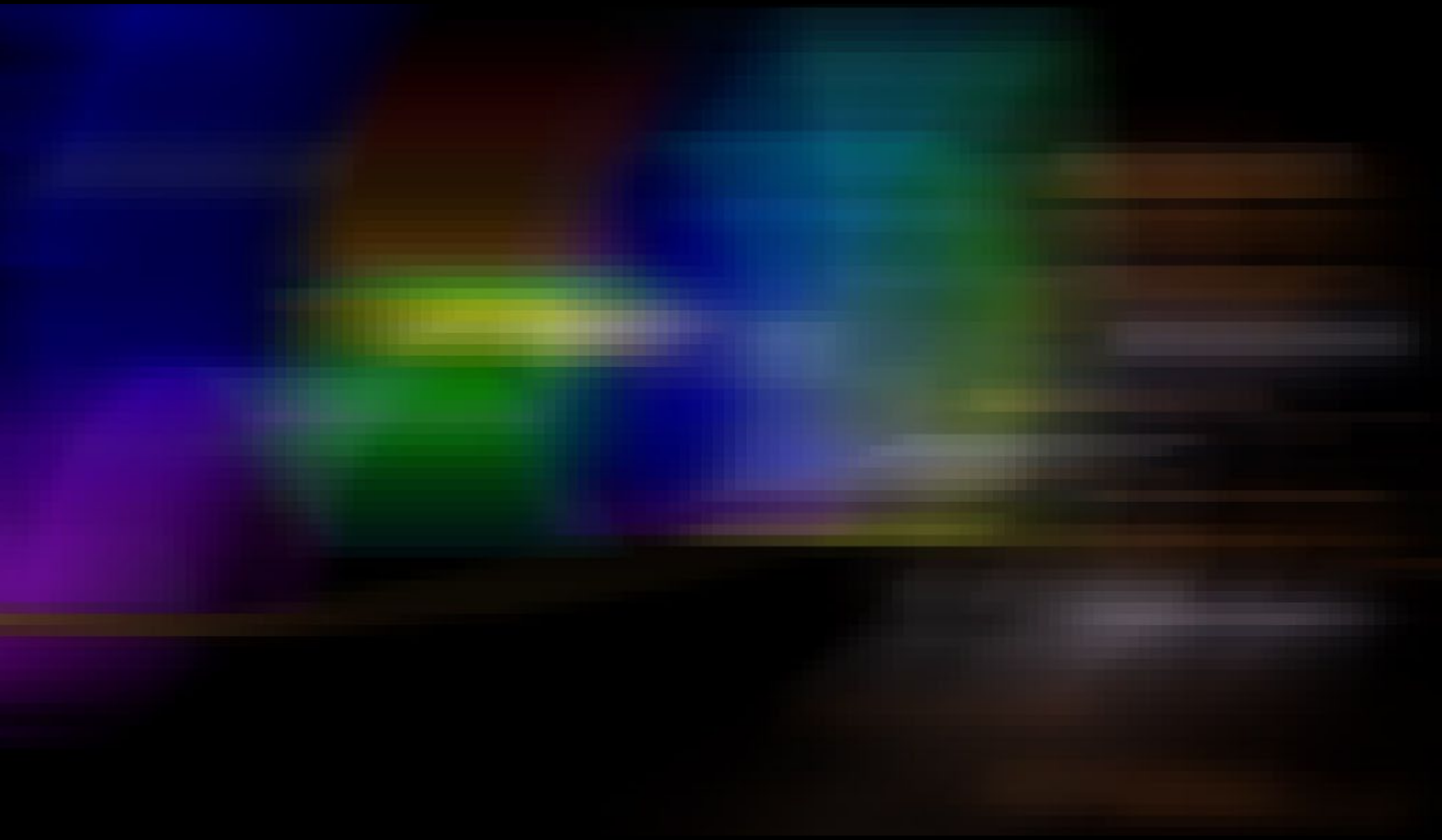
Credit: Blake Harvey, NCSA

institutions and agencies to enable teams and communities to directly address the next generation of major scientific challenges;

- On top of the integrated environment, these programs will entail the creation and development of comprehensive, multidisciplinary programs to support teams and communities in attacking complex transformational science and engineering problems, requiring integrative approaches to data, hypothesis testing, and computation, that cannot be adequately addressed by small groups; and requiring teams that include domain sciences and engineering, along with enabling sciences.
- Additional investments and focus on the long-term sustainability of research communities to address what are often decadal efforts for grand challenges;
- Investments to facilitate the creation of multidisciplinary expertise in partnership with campuses as a core competency for research and to develop CDS&E as an important career in the research enterprise.

Integrative approaches are required to solve complex problems and issues being addressed by science, engineering and education communities.

Credit: Thinkstock



NSF: 12-051

National Science Foundation