National Science Foundation
Advisory Committee for Cyberinfrastructure
Task Force on Grand Challenges

Final Report, March 2011

A Report of the National Science Foundation Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges

March 2011

THE NSF-ACCI TASK FORCE ON GRAND CHALLENGES

J. Tinsley Oden, Chair
The University of Texas at Austin

Omar Ghattas, Co-Chair,
Computational Methodologies Working Group Lead
The University of Texas at Austin

John Leslie King, Co-Chair
Virtual Organizations and Grand Challenge Communities Working Group Lead
University of Michigan

Barry I. Schneider, Liaison
The National Science Foundation

Klaus Bartschat
Drake University

Frederica Darema
Air Force Office of Scientific Research

John Drake
Oak Ridge National Laboratory

Thom Dunning
Software Working Group Lead
University of Illinois at Urbana-Champaign

Donald Estep
Computational Methodologies Working Group Co-Lead
Colorado State University

Sharon Glotzer
Education Working Group Lead
University of Michigan

Michael Gurnis
California Institute of Technology

Christopher Johnson
Data and Visualization Working Group Co-Lead
University of Utah

Daniel S. Katz
University of Chicago and Argonne National Laboratory

David Keyes
Columbia University and KAUST

Sara Kiesler
Carnegie Mellon University

Sangtae Kim
Morgridge Institute

James Kinter
Institute of Global Environment and Society Inc.

Gerhard Klimeck
Purdue University

C. William McCurdy
University of California Davis

Robert Moser
The University of Texas at Austin

Christian Ott
California Institute of Technology

Abani Patra
HPC Working Group Lead
University at Buffalo

Linda Petzold
Education Working Group Co-Lead
University of California, Santa Barbara

Tamar Schlick
New York University

Klaus Schulten
University of Illinois, Urbana-Champaign

Victoria Stodden
Virtual Organizations and Grand Challenge Communities
Working Group Co-Lead
Yale Law School

Jeroen Tromp
Princeton University

Mary Wheeler
The University of Texas at Austin

Susan J. Winter, Liaison
National Science Foundation

Cathy Wu
Data and Visualization Working Group Lead
University of Delaware

Katherine Yelick
Software Working Group Co-Lead
University of California, Berkeley

Task Force Administrative Coordinator:
Jon Bass, *The University of Texas at Austin*

Preface

This document contains the findings and recommendations of the NSF – Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges addressed by advances in Cyber Science and Engineering. The term Cyber Science and Engineering (CS&E) is introduced to describe the intellectual discipline that brings together core areas of science and engineering, computer science, and computational and applied mathematics in a concerted effort to use the cyberinfrastructure (CI) for scientific discovery and engineering innovations; CS&E is computational and data-based science and engineering enabled by CI. The report examines a host of broad issues faced in addressing the Grand Challenges of science and technology and explores how those can be met by advances in CI. Included in the report are recommendations for new programs and initiatives that will expand the portfolio of the Office of Cyberinfrastructure (OCI) and that will be critical to advances in all areas of science and engineering that rely on the CI.

The Task Force, one of six created by the ACCI during the summer of 2009, met many times since its inception, and held two workshops, one in August 2009, and another in April 2010. Over 100 scientists from the CS&E community participated in these meetings and contributed to the ideas that led to eight working drafts of this document before the present version was completed. A partial list of the Workshop attendees is given in Appendix A.

The Task Force consisted of six working groups, dedicated to six key components of the study: *Computational Methods and Algorithms*, led by Donald Estep and Omar Ghattas; *High Performance Computing*, led by Abani Patra; *Software*, led by Thom Dunning and Katherine Yelick; *Data and Visualization*, led by Cathy Wu and Christopher Johnson; *Education, Training, and Workforce Development,* led by Sharon Glotzer and Linda Petzold; and *Grand Challenge Communities*, led by John King and Victoria Stodden. Tinsley Oden chaired the Task Force, Omar Ghattas and John King acted as co-chairs, and Barry I. Schneider of NSF was the NSF liaison between OCI and the Task Force. Jon Bass served as the Task Force Administrative Coordinator.

Many others contributed to the writing of various sections, and the work of the following should be mentioned: Guy Almes (TAMU), Luc Anselin (Arizona State) George Biros (Georgia Tech), Robert Bonneau (AFOSR), James Brasseur (Penn State), Richard Brower (Boston U.), Peter Cummings (Vanderbilt/ORNL), Frederica Darema (AFOSR), Thomas Dietterich (Oregon State), Ron Elber (UT-Austin), Tom Evans (Indiana), Geoffrey Fox (Indiana), Gary King (Harvard), Alan Laub (UCLA), David Lazer (Northeastern), J. Scott Long (Indiana), Liz Lyon (U Bath), Dimitri Mavriplis (U. Wyoming), Thomas Maier (ORNL), Stephen McCormick (CU Boulder), Richard Moore (SDSC), Bernice Pescosolido (Indiana), Alex Pothen (Purdue), Mark Shephard (RPI), Renata Wentzcovitch (U. Minnesota), and John Ziebarth (Krell Inst.).

Although this report was prepared by a task force commissioned by the National Science Foundation, all opinions, findings, and recommendations expressed within it are those of the task force and do not necessarily reflect the views of the National Science Foundation.

Contents

Executive Summary

This document describes the major findings and recommendations of the NSF Task Force on Grand Challenges. It is one of six Task Forces created by the Advisory Committee for Cyberinfrastructure (ACCI) at NSF charged with the study of possible new programs and an expanded scope of CS&E within the foundation. The specific charges of this Task Force were:

1) Develop a thorough understanding of the requirements of science and engineering applications on the cyberinfrastructure that must be developed to make significant progress toward resolving Grand Challenge (GC) Problems.

2) Identify methods for enabling different communities to work together to solve complex problems. This will involve the study of virtual organizations and tools to support them.

3) Interact with other task forces to come forth with a set of well conceived recommendations on ideas for new programs that might be developed within OCI that will more tightly couple advanced problem solving in science and engineering with continuing investments.

4) Explore the role of Computational Science and Engineering enabled by Cyberinfrastructure in scientific discovery and engineering innovation and its place in the organizational structure and mission of NSF.

We provide definitions of a few key terms to make more precise the targets of this study and how we approach these goals.

1) Cyberinfrastructure (CI): the broad collection of computing systems, software, data acquisition and storage systems, and visualization environments, all generally linked by high-speed networks, often supported by expert professionals.

2) Cyber Science and Engineering (CS&E): computational science and engineering enabled by the cyberinfrastructure. Science is the enterprise dedicated to the acquisition of knowledge, and engineering is the innovative application of science for human needs. The classical pillars of science – the methods for acquiring knowledge – are theory (hypotheses put forth to explain physical realities) and experiments (knowledge gained through observation using human senses or instruments). In this document, computational science and engineering refers to science and engineering achieved through the use of computational methods and systems (generally, hardware, software, networks, etc.). Thus, computational science enables extensions of theory through computer modeling and simulation (but not exclusively), and enables extensions of experimental science through data-intensive computing (but not exclusively). CS&E is thus the intellectual and technological discipline lying at the intersection of applied mathematics, computer science, and all core science and engineering areas including data-based observational science and engineering and statistics, dedicated to the development and use of computational methods and systems in scientific

discovery and engineering innovation.

3) Grand Challenges (GC's): the "Grand Challenges" were U.S. policy terms set in the 1980's as goals for funding high-performance computing and communications research in response to foreign competition. They were described as "fundamental problems of science and engineering, with broad applications, whose solution would be enabled by high-performance computing resources…" (*cf.* http://www.nae.edu). Today, the Grand Challenges are interpreted in a much broader sense with the realization that they cannot be solved by advances in HPC alone: they also require extraordinary breakthroughs in computational models, algorithms, data and visualization technologies, software, and collaborative organizations uniting diverse disciplines.

Among the many Grand Challenges that can be listed are:

- Advanced New Materials*
- Prediction of Climate Change*
- Quantum Chromodynamics and Condensed Matter Theory
- Semiconductor Design and Manufacturing
- Assembling the Tree of Life*
- Drug Design and Development
- Energy through Fusion
- Water Sustainability
- Understanding Biological Systems*
- New Combustion Systems
- Astronomy and Cosmology*
- Hazard Analysis and Management*
- Human Sciences and Policy*
- Virtual Product Design*
- Cancer Detection and Therapy
- CO_2 Sequestration*

As representative examples, those marked with an asterisk are discussed in more detail in the body of this report. Common themes of all Grand Challenges include:

- All Grand Challenges face barriers due to challenges in software, in data management and visualization, and in coordinating the work of diverse communities that must work together to develop new models and algorithms and to evaluate outputs as a basis for critical decisions.
- Transformative discovery and innovation needed to address the Grand Challenges will often require capabilities approaching or exceeding exascale computing, and this will require dramatic changes in processor architecture and in power management.
- More faithful computational models and more stable and robust algorithms needed for large-scale Grand Challenge problems will have to adapt to emerging manycore and hybrid architectures, which appear to be a very promising path to energy-efficient increased computational power in the near future. Of critical importance are methods that are informed by observational data in a way that can cope with uncertainty in data and quantify uncertainties in predictions. New methods need to be developed to facilitate multiscale modeling, scalable solvers for multiphysics and stochastic problems, and large-scale data-intensive simulations.
- Of special significance is the need for acquiring relevant data for calibration and validation of large-scale computational models and the characterization and quantification of uncertainties. This will require the development of statistical representations of data on parameters and observations, statistical inverse methods and software that implement them, and methods to resolve the large stochastic systems that result from model and data uncertainties. The transition of conventional deterministic methods and models of complex physical events to those

accounting for uncertainties and stochasticity will increase by several orders of magnitude the size, complexity, and computational work needed for predictive simulations. Another challenge presented by data-intensive simulation is ensuring the ability of others to verify and reproduce the scientific results. This involves issues spanning software design, code building, and code and data dissemination.

- The combination of the development of computational models based on scientific and engineering principles, on principles and methods of computer science and computing technology, and on the use of core computational and applied mathematics that come into play to address effectively all Grand Challenge problems, represents the discipline referred to here as Cyber Science and Engineering. While NSF has supported many cross-directorate initiatives in basic CS&E over the years, there is no home for it within the NSF organizational structure. The result has generally been scattered, underfunded programs with low proposal success rates, and no sustainability for efforts requiring long-term investments in software and algorithm development and infrastructure. Effectively attacking pressing Grand Challenge problems under these conditions is extremely difficult.

*4) Grand Challenge Communities and Virtual Organization*s: these are organizational structures enabled by the effective use of modern CI to facilitate collaboration among geographically distributed and intellectually diverse multidisciplinary groups, necessary for addressing large-scale and critical Grand Challenges affecting many areas of society and areas of science and engineering. Grand Challenge Communities often include participants from intellectual disciplines that have different and conflicting conventions of collaboration, are not used to working with each other, and reside in distinct geographic locations. The organization of effective GC Communities and VO's is itself a formidable challenge requiring independent study in its own right. An introduction to the concepts and issues is given in Chapter 8 of this report, and a more complete study is to be the subject of a later report to the Advisory Committee on Cyberinfrastructure.

Findings and Recommendations

Overarching Recommendation

Throughout this study, the fundamental role of CS&E in scientific advancement and in addressing the Grand Challenges is repeatedly noted. This is a subject that has emerged since the advent of scientific computation and has grown to be of historic importance, affecting virtually every area of science and technology and revolutionizing the way science and engineering are done. It is now widely recognized as a third pillar of science and has become a subject indispensible to the nation's welfare, competitiveness, and standing in the international scientific community. Its importance has been noted in numerous studies sponsored by federal agencies including, in particular, the National Science Foundation. There is a wide consensus that it is truly a discipline, in the same spirit as applied mathematics or computer engineering or biochemistry, but its extraordinary value stems from its unique reliance on interdisciplinary collaborations, drawing adaptively from a core body of knowledge in mathematics, computer science, engineering and technology, and all scientific disciplines to address specific research challenges that invariably cross traditional boundaries.

CS&E differs from core computer science and applied math research in that it is more closely

intertwined with applications: it seeks to exploit the structure of particular scientific and engineering problems to design effective methods to overcome the challenges inherent in driving science and engineering problems. CS&E research seeks to advance mathematical methods to a greater extent than is done in core computer science research; also the methods it employs are more hardware-aware and software-oriented than is typical in applied math research. Finally, CS&E differs from core science and engineering disciplines in its greater focus on advanced computer science and applied math and its inherent reliance on interdisciplinary collaborations.

The fundamental importance of CS&E has frequently been recognized within the Foundation, which has attempted to fund cross-directorate programs in CS&E over the last two decades. Typically, these cross-cutting initiatives have been "ad-hoc", temporary programs with very low proposal success rates that are inadequate for creating the critical mass of knowledge and communities for systematically advancing research on the abiding and pervasive challenges in CS&E. Over the years NSF has supported a number of cross-cutting CS&E programs starting with the Grand, National, and CS Challenges Programs in the early-to-mid 90s, components of Knowledge and Distributed Intelligence (KDI) in the late 90s, ITR in the early-mid 00s, Dynamic Data Driven Application Systems (DDDAS) in 2005, the Collaborations in Mathematical Geosciences (CMG), Collaborative Research in Computational Neuroscience (CRCNS), Advances in Biological Informatics, and PetaApps in the late 00s, and Cyber-enabled Discovery and Innovation (CDI) today, but these programs are too few and far between to support research in an area so vital to the nation's competitiveness and future. Science agencies of foreign competitors have embraced CS&E and are investing heavily in this area, as is clearly spelled out in the NSF-supported Simulation-Based Engineering and Science (SBES) study [58]. All of these considerations lead to the following recommendations.

RECOMMENDATION:

It is recommended that permanent programmatic activities in CS&E be established within NSF. These activities should range from division- and directorate-level programs for discipline-specific aspects of CS&E, to permanent NSF-wide cross-cutting CS&E programs possibly managed by OCI. Interdisciplinary projects could be co-funded between cross-cutting and relevant disciplinary programs. The permanent NSF programmatic activities in CS&E would play a significant role in incentivizing universities to expedite the creation of CS&E research and educational programs, which in turn would go a long way in addressing the immense shortage of well-trained computational scientists and engineers in the workforce.[1]

[1] A resolution to create a new program in Computational and Data-Enabled Science and Engineering, coordinated by OCI, was unanimously endorsed by the Advisory Committee on Cyberinfrastructure and approved by the NSF Director on May 27, 2010. A copy of the letter to NSF Director Arden L. Bement recommending the creation of this program is included in Appendix B.

Owing to the breadth of research in CS&E across many federal agencies, a companion recommendation is provided as follows:

RECOMMENDATION:

NSF should work with the Department of Energy and other agencies in the creation of an Interagency Working Group on CS&E or generally on Computational Science and Engineering, including Data-Intensive Computing, in the spirit of other NSF-wide working groups. This broad-based Working Group could provide input leading to important interagency collaborations on new programs, particularly in HPC, and could lead to more focused and efficient use of resources to address the Grand Challenges facing our nation.

Findings and Recommendations Concerning Advances in CI Needed to Confront Grand Challenge Problems:

1) Computational Models, Methods, and Algorithms

Computational methods and algorithms have played a crucial role in the solution of complex scientific and engineering problems since the earliest days of computing. They form the key link between mathematical models of physical phenomena of interest and high performance software that can be used to carry out analysis and prediction of the behavior of complex physical systems. Synergistic advances in computing and computational methods have stimulated scientific and engineering breakthroughs, which have in turn motivated further advances in enabling technologies. Over the past half-century, advances in computational methods have led to speedups in the solution of important scientific problems that are as significant as those resulting from advances in the hardware alone. For example, Figures 7-10 in Chapter 3 illustrate breakthroughs on scientific problems that have been enabled by advances in algorithms. Computational methods, however, are often taken for granted, due to past successes and their largely hidden role in powering CS&E software. But while recent isolated successes have occurred, computational methods that can scale to petascale systems are in their infancy for *difficult* problems, such as those with strong heterogeneities and anisotropies, multiphysics couplings, multiscale/multirate behavior, stochastic forcing, uncertain parameters, dynamically evolving geometries, continuum-atomistic couplings, large-scale combinatorial structure, and so on. *But it is precisely these features that characterize next-generation Grand Challenge problems.* Absent a systematic research effort, continued progress on frontier CS&E problems is not assured, and Federal investments in hardware, networking, and software will be jeopardized. There is no question that building an exascale machine will be hard; but using it effectively to solve CS&E Grand Challenge problems will be even harder.

To address the difficulties in developing computational methods for scientific Grand Challenges such as those described in Chapter 2, a broad-based, comprehensive, long-term, and

vigorous research program in advanced computational methods must be established to overcome the challenges faced in devising, analyzing, replicating, scaling up, and applying new methods for critical CS&E problems on advanced computing systems. This program should support multidisciplinary and interdisciplinary teams that bring together applied mathematicians, computer scientists, and computational scientists and engineers. In turn, an additional CI challenge is to ensure that advances in computational methods and algorithms developed in one discipline are disseminated across all disciplines that face computational problems with similar structure.

Computational methods and algorithms play a key role at all stages of CS&E, including solution techniques for complex multiscale/multiphysics problems, advanced spatial and temporal discretization schemes for high fidelity simulations, scalable algorithms for solution of large linear and nonlinear algebraic systems and eigenvalue problems, methods for quantifying uncertainties in large-scale simulations, and algorithms for solution of large-scale optimization problems arising in design, control, and inversion.

RECOMMENDATION:

A broad-based, comprehensive, long-term, and vigorous research program in advanced computational methods should be established to overcome the challenges faced in devising, analyzing, replicating, and scaling up new computational methods for a new generation of critical CS&E problems on advanced computing systems. These should include advances in discretization methods, solvers, optimization, statistical methods for large datasets, and validation and uncertainty quantification methods including those in reproducible research, all targeted at enabling new frontiers in large-scale multiphysics, multiscale simulations on emerging architectures. This program should support multidisciplinary and interdisciplinary teams that bring together applied mathematicians, computer scientists, and computational scientists and engineers.

2) High Performance Computing

Transformative discovery and innovation in most disciplines important to addressing the Grand Challenges, such as climate, energy, environment, national security, disaster preparedness, and medicine, depend on the pervasive and seamless availability of computing at scale. According to many projections, general purpose exascale computing equipment is likely to be available in the next 10-15 years. However, this will likely be made possible only by dramatic changes in processor architectures, including very large scale of multi-core processing, power management, and packaging. New methodologies for power management at circuit, device, and system level, locality and concurrency of data and the computations that use/generate it, and resilience to system faults, are going to be crucial to the development of these systems.

NSF has taken on the challenge of providing and maintaining the computational infrastructure for advanced computing for over two decades. Providing the new infrastructure needed to address

the Grand Challenges in the future will be an especially daunting objective as the complexity and heterogeneity of the new systems and urgency of the research challenges require that a variety of innovative and "bleeding edge" systems be supported. HPC is an area where U.S. and NSF leadership has yielded great competitive advantage and sustained national security, but it is also an area in which that leadership is constantly challenged. Reliability and usability of modern HPC hardware is likely to be a Grand Challenge in research on par with the others listed above and will need a deliberate and long-term strategy.

RECOMMENDATION:

It is recommended that NSF, through OCI, continue to give high priority to funding a sustained and diverse set of HPC and innovative equipment resources to support the wide range of needs within the research community. These needs include support for the development of technologies to meet the foremost challenges in HPC, such as power-aware and application-sensitive architectures, new numerical algorithms to efficiently use petascale and exascale[2] architectures, and data flow and data analysis at the extreme scale.

3) Software

With the arrival of petascale computers and the expected progression toward multi-petascale and exascale computers in the next decade as well as the rapidly growing capabilities in data-driven discovery, opportunities for advancing science and engineering have never been higher. Also, with the expanding role of data-driven discovery and computational modeling and simulation in decision support as well as scientific discovery, the reproducibility of results places new demands on the robustness and documentation of software. As a result, the demands on innovative and sustainable software have never been higher. These considerations lead to the following recommendations.

RECOMMENDATIONS:

It is recommended that NSF:

1) Support the creation of reliable, robust science and engineering applications and data analysis and visualization applications for Grand Challenges as well as the software development environment needed to create these applications.

2) Provide support for the professional staff needed to create, maintain, evolve and disseminate the above applications as part of its grant funding.

[2] Petascale computing is the state-of-the-art in high performance computing. A one petaflop supercomputer performs 10^{15} floating point operations per second (FLOPS). A one exaflop supercomputer would perform 10^{18} FLOPS. (A typical PC performs on the order of 10^{9} FLOPS.)

3) Establish best practices for the release of science and engineering applications and data as well as the workflows involved in their creation to ensure the reproducibility of computational results.

4) Data and Visualization

Many areas of science and engineering are now becoming data-driven sciences, a shift that has led to a new era in computing identified by Jim Gray as the "fourth paradigm" of science. In this new paradigm, representing one of science's grand challenges, science follows a data-centric approach in which massive amounts of digital scientific data must be collected, integrated, and interpreted via visualization, mining, and modeling to generate new hypotheses and to accelerate discovery and innovation. Data-intensive science is characterized by the massive scale and complexity of data it relies on and by the interdisciplinary and multidisciplinary methods it requires for data generation, management, analysis, visualization, and re-using and re-purposing, including the reproducibility of results. Because data used in the data-centric approach to science are often heterogeneous, spanning multiple spatial and temporal scales, in distributed locations, and of varying levels of performance, reliability, security, and accessibility, the challenges to scientists are not only to find ways to physically manage and move the data, but also to develop new software tools for managing, migrating, and efficiently analyzing the data. These tools must employ an end-to-end approach that encompasses the entire data life cycle, from the initial data acquisition through data management and storage to data integration, analysis, visualization, and knowledge discovery.

However, we currently lack the robust data infrastructure, innovative research in data visualization and analysis, and interdisciplinary data scientists and data professionals needed to address the requirements of the new scientific paradigm. We must now embark on critical research and the development of cyberinfrastructure to address our shortcomings in data analysis and visualization, data integration and interoperability, data provenance and stewardship, scientific workflow and meta-tools, exascale computing, active storage and online analysis, data storage and management, and high-speed computer networks. As data-driven science continues to increase in its scope and impact, we need to better communicate the value digital scientific data and visualization bring to the broad scientific community, policy makers, and the public. To this end, the NSF must support research infrastructure, robust and persistent cyberinfrastructure, and the training of next-generation data scientists and professionals to empower data-driven science and data-intensive computing for discovery, innovation, and solution of society's pressing problems in health, energy, environment, and food.

RECOMMENDATIONS:

NSF, largely through and coordinated by OCI, should support research infrastructure and robust persistent cyberinfrastructure to empower data-driven science and data-intensive computing for discovery, innovation, and solution of society's pressing problems in health, energy, environment, and food.

1) Research: Funding for research on data management, network infrastructure, data analysis, and data visualization (i) to manage the pipeline from field

instruments to large-scale data analysis to end-user visualization and to public and policy makers, and (ii) to support data-intensive computing.

2) Data Infrastructure: Support for robust, persistent cyberinfrastructure to support the coordinated flow, storage, and management of data from instrument to (remote and local) computing resources to archiving and visualization.

3) Education: Support for building (i) the next-generation of data scientists who can work in a multi-disciplinary team of researchers in high performance computing, mathematics, statistics, domain-specific sciences, etc., (ii) data curation professionals who can support meta-data collection, indexing, and access, collaborating with scientists who collect and consume data.

5) Education, Training and Workforce Development

Universities are not adequately preparing today's students with the right background, skills, breadth and depth to become tomorrow's computational scientists and engineers, able to harness powerful new supercomputers for scientific discovery and engineering innovation. Our nation is losing its leadership position in CS&E among our principal competitors in the industrialized world, as other nations have embraced this challenge. New courses and curricula are urgently needed. Training in core CS&E skills needs to be widely available and easily accessible, to facilitate workforce development and accelerate research progress across the sciences and engineering. Much of the traditional compartmentalization of knowledge, both within our major universities, and to an extent within NSF itself, is not well suited for interdisciplinary research and education vital to CS&E. It is critical that actions be taken by NSF to address these issues.

RECOMMENDATIONS:

NSF should support education, training, and workforce development through the following grants and new programs:

1) Educational excellence grants at the undergraduate and graduate levels, which include funding for the development of new courses, curricula, and academic programs in CS&E that address the computational and analytical skills required in virtually all STEM disciplines.

2) Support for the formation of virtual communities engaged in CS&E education, including virtual entities leveraging expertise across colleges, universities, national and government laboratories, and supercomputing centers. Training, in the form of short courses, in core skills at all levels should be available online and supported 24/7, making the training broadly accessible.

3) Institution-based traineeship grants that train graduate students and postdoctoral fellows in the multidisciplinary, team-oriented iteration among

experiment, theory, and computation that is rapidly becoming a paradigm in critical STEM research areas and that has long been a standard in government laboratories and industry.

4) The creation of a pan-agency facility or program to coordinate training in CS&E education.

5) Grants that facilitate the transition of exceptionally talented graduate and postdoctoral students in computational science and engineering to permanent positions in academia as well as industry and government/national labs.

6) Sustainable, permanent programs in CS&E research and education at all funding agencies to demonstrate a long-term commitment to supporting CS&E as a discipline, thereby creating reliable partners for universities seeking institutional transformational change and for trained workers seeking careers in CS&E.

6) Grand Challenge Communities and Virtual Organizations

Collaboration is essential to meeting the Grand Challenges, and CI-enabled virtual organizations offer considerable promise for improving scientific and engineering productivity. However, there are many remaining obstacles to full exploitation of CI for collaboration. The scope of these obstacles goes beyond the purview of this report, and is addressed in a separate report to the ACCI. However, for the purposes of this report the following recommendations can be made.

RECOMMENDATIONS:

The NSF should initiate a thorough study outlining best practices, barriers, success stories, and failures, on how collaborative interdisciplinary research is done among diverse groups involved in Grand Challenge projects.

The NSF should invest in research on virtual organizations that includes:

1) Studying collaboration, including virtual organizations, as a science in its own right;

2) Connecting smaller virtual organizations to large-scale infrastructure by providing supplementary funds to such projects, supporting development of tools, applications, services, etc. with a mandate to disseminate those elements to other communities and users;

3) Investing in systematic, rigorous, project-level and program-level evaluations to determine the benefits from virtual organizations for scientific and engineering productivity and innovation;

4) Encouraging NSF program officers to share information and ideas related to virtual organizations with training and online management tools.

1 Introduction

No period in human history has witnessed the development of more technologies that affect scientific discovery than the years bridging the turn of the last century. These include major advances in high-performance computing (HPC), in broad areas of information technology, grid computing, advanced networking, the internet, data repositories, scientific visualization, and many more, all collectively called the cyber-infrastructure (CI). In recognition of the enormous importance of these developments to all areas of scientific and engineering research, the National Science Foundation created the Office of Cyberinfrastructure (OCI) in July 2005, to manage advances in CI across the Foundation.

The OCI is, by design, an overarching unit within the Foundation in that it provides support to all other NSF Directorates. While support of research in discipline-specific components of computational science and some of the development of related infrastructure is still the responsibility of individual directorates, OCI functions as both an agent enabling collaborations across disciplines and as a steward of research and new developments in CI itself that are critical to the success of interdisciplinary research.

The remarkable success of NSF-supported developments in CI during the short time period since OCI was created is an indication of the rising importance of interdisciplinary research and the critical role the CI plays in facilitating collaboration of diverse and widely separated communities of researchers. The success may also be an indication of the expanding role of computational science and engineering in all areas of scientific inquiry and technology, and of the advances in computational science and engineering made possible by CI. Underlying these advances are investments in the TeraGrid and the Open Science Grid, the Path to Petascale Computing, and in numerous services provided by CI in support of data intensive computing, software, HPC, networking, data storage, and education.

Looking forward, critical scientific and technological challenges loom ahead that will require major advances in science and engineering that cross the boundaries of many traditional disciplines. These are the Grand Challenges of science and technology that our country faces in the immediate future; at stake are our competitiveness, economy, security, general welfare, and leadership in scientific discovery. The challenges are daunting and they range from issues related to climate, energy, natural hazards, and defense to medicine, manufacturing, drug design, biology, and cosmology. To meet challenges of such importance and scale will require unusual coordination of and collaboration between the diverse communities of researchers referred to earlier, as well as corresponding advances in CI to facilitate these collaborations. These groups are the *Grand Challenge Communities*.

1.1 Cyber Science and Engineering (CS&E)

The expanding role of CI in providing the infrastructure for interdisciplinary research calls for the expansion of OCI's portfolio and the inclusion of new directions in CI-related research. It also calls for an expanded view of CI itself, now including many of the basic computational science activities – modeling, simulation, data-driven science – that should be developed in step with advances in the infrastructure itself. To develop a plan for executing these expanded programs, ACCI created several task forces, including the present Task Force, that has as its mission the study of the following broad issues: 1) what new developments in CI will be needed to impact new scientific research; 2) how can the work of Grand Challenge Communities and Virtual Organizations be facilitated by OCI; 3) what new programs within OCI are needed to carry out its expanded mission; and 4) explore the role of computational science and engineering enabled by cyberinfrastructure in scientific discovery and engineering innovation and its place in the organizational structure and mission of NSF. The key scientific target of this study will be referred to as *Cyber Science and Engineering (CS&E)*. This includes the traditional realm of computational science and engineering, a discipline at the intersection of applied and computational mathematics, computer science, and core science and engineering disciplines, but now dramatically enhanced by access to the full spectrum of CI-enabling-technologies: HPC, software, modern computational models and algorithms, data intensive computing, networking and storage, and visualization, as well as issues of education. But overlaid on such scientific goals are issues of new communities of domain science and computer science specialists that can employ CI to tackle Grand Challenges of great complexity and importance: the Grand Challenge Communities and Virtual Organizations.

1.2 Collaboration and the Cyberinfrastructure

It has been recognized since the 17th Century that much scientific research takes place in distributed communities involving multiple institutional venues, often separated by geographic distance. Successful communities organize production – their ways of doing things – to establish objectives, facilitate teamwork, and resolve disputes that might otherwise prevent them from meeting those objectives. Communities have developed various mechanisms to facilitate collaboration, including scientific societies, conferences, workshops, peer-reviewed publications, academic departments, institutes, and sabbaticals. Yet collaboration within and across communities remains difficult. Grand Challenge Communities often include participants from intellectual disciplines that have different and conflicting conventions of collaboration, are not used to working with each other, and are in various geographic locations. At the same time, Grand Challenge Communities have the potential to use modern cyberinfrastructure to enable more effective collaboration. CI offers a promising pathway to that high level of collaboration. Within the modern computing environment, multiple groups are able to form dynamic cooperative interrelationships – virtual organizations – that consolidate and share the computing, the data, and human resources as required to attack problems in advanced science and technology. OCI has taken an important lead on this, but collaboration in the large is a topic of such

broad importance that NSF as an organization should engage the matter. A separate report to the ACCI focuses on this.

1.3 Organization of this Report

This document describes the findings of the Task Force on Grand Challenges. Within this task force, two study areas and several working groups were created:

Area 1: CI Requirements for Next-Generation CS&E. Within this area are five working groups:

1) *High Performance Computing (HPC)*

   Focuses on the new opportunities for scientific discovery that could be achieved through advances in HPC.

2) *Software*

   Focuses on the critical software developments needed to address the Grand Challenges amid a changing landscape of computer architectures; also focuses on approaches to the maintenance and support of relevant software.

3) *Data and Visualization*

   Addresses how the OCI can prepare for and capitalize on the enormous increases in data relevant to scientific discovery, as well as methods of data acquisition, storage, and analysis.

4) *Advanced Computational Methodologies*

   Addresses a central area of CS&E, namely, the development of new and effective algorithms and computational methods that optimize the use of CI, so that computational science does indeed become the third pillar of the scientific method.

5) *Education and Workforce Preparation*

   Educates future generations of scientists and engineers in the foundations of CS&E and prepares them for using CI and contributing to its development.

Area 2: Collaboration, Including Grand Challenge Communities and Virtual Organizations. The goal of this area is to develop technologies, and organizational strategies that enable CI to facilitate effective collaboration of distributed multidisciplinary groups. Such collaboration is essential if science and engineering are to be effective in overcoming the overarching societal Grand Challenges.

In what follows, examples of several Grand Challenge problems are described, the solutions of which will require extraordinary advances in each of the components of CS&E and, correspondingly, significant advances in CI. To address these challenges, one must create advanced computational models to provide a basis for representing our knowledge of the physical realities involved in each Grand Challenge, and extensive data to inform the models or to represent information from which new knowledge can be obtained. Ultimately, to resolve these models or process these data, advances in High Performance Computing and computational methods and algorithms and, correspondingly, scientific software are needed. A great challenge is also the organization of the work itself in a way that the GC Communities and VO's can function effectively and efficiently to meet the challenges. Finally, the advances toward resolving the Grand Challenges will have no lasting value if they are lost to our own generation: we must find ways to equip the next generation of scientists and engineers with the tools, concepts, and principles of CS&E. These component issues are dealt with in the chapters following this introduction.

2 Grand Challenges in CS&E

This chapter contains brief expositions on several examples selected as representative of principal technological and scientific problems requiring new developments in CI to enable advances in scientific discovery. They embody critical issues in the ever-expanding vistas of high performance computing; in the ubiquitous area of software; in data and visualization; in the fundamentally important area of advanced computational methods; and on the critical area of education in CS&E and CI.

2.1 Addressing the Grand Challenges

The *Grand Challenges* were U.S. policy terms set in the 1980's as goals for funding high-performance computing and communications research in response to foreign competition. They were described as "fundamental problems of science and engineering, with broad applications, whose solution would be enabled by high-performance computing resources…" (*cf.* http://www.nae.edu) Today, the Grand Challenges are interpreted in a much broader sense with the realization that they cannot be solved by advances in HPC alone: they also require, as noted earlier, extraordinary breakthroughs in computational models, algorithms, data and visualization technologies, software, and collaborative organizations uniting diverse disciplines. Many communities have come forth over the past two decades with reports that indentify specific Grand Challenges in their respective fields. These GC's virtually all require breakthroughs in CS&E enabled by advances in CI. An incomplete list of examples is:

- Advanced New Materials (electronic structure properties, chemical catalysts, …)
- Prediction of Climate Change
- Quantum Chromodynamics and Condensed Matter Theory
- Semiconductor Design and Manufacturing
- Drug Design
- Energy through Fusion
- New Combustion Systems
- Astronomy and Cosmology
- Cardiovascular Engineering
- Water Sustainability
- Cancer Detection and Therapy
- CO_2 Sequestration

In the subsections below, we give brief accounts of several representative problems that attempt to not only identify the open problems that complicate the challenges themselves, but also the advances in CS&E and CI needed to confront them. We emphasize that these are merely examples of Grand Challenges, and many other problems could have been chosen.

2.2 Climate Change Prediction to Advise Regional Adaptation Strategies and Global Mitigation Policies

Decades of careful evaluation of weather and climate measurements, paleoclimate proxy records, and the output of global climate models have produced convincing evidence that the earth's climate is undergoing change at a rate more rapid than that of any previous period in human history. More over, human activities may be at least partially responsible for that change. To address the threat posed by such change, the global society rightly demands accurate projections of climate change, with ever-decreasing levels of uncertainty. A complementary demand exists for means to anticipate climate changes with greater spatial discrimination over the next 30 years, especially as those changes may affect extreme weather and climate events.

The history of capability computing is coincident with the history of the development and standard use of weather and climate models, ensembles, and earth system models. Since the early experiments with numerical weather prediction on the first general-purpose computer, ENIAC, the scope of the modeling system has expanded along with computer capability. Phillips' development of a global circulation model of the atmosphere in 1956 introduced the modern age of numerical weather prediction. In 1967, Manabe and Weatherald projected climate change based on doubling of atmospheric CO_2 concentrations, which

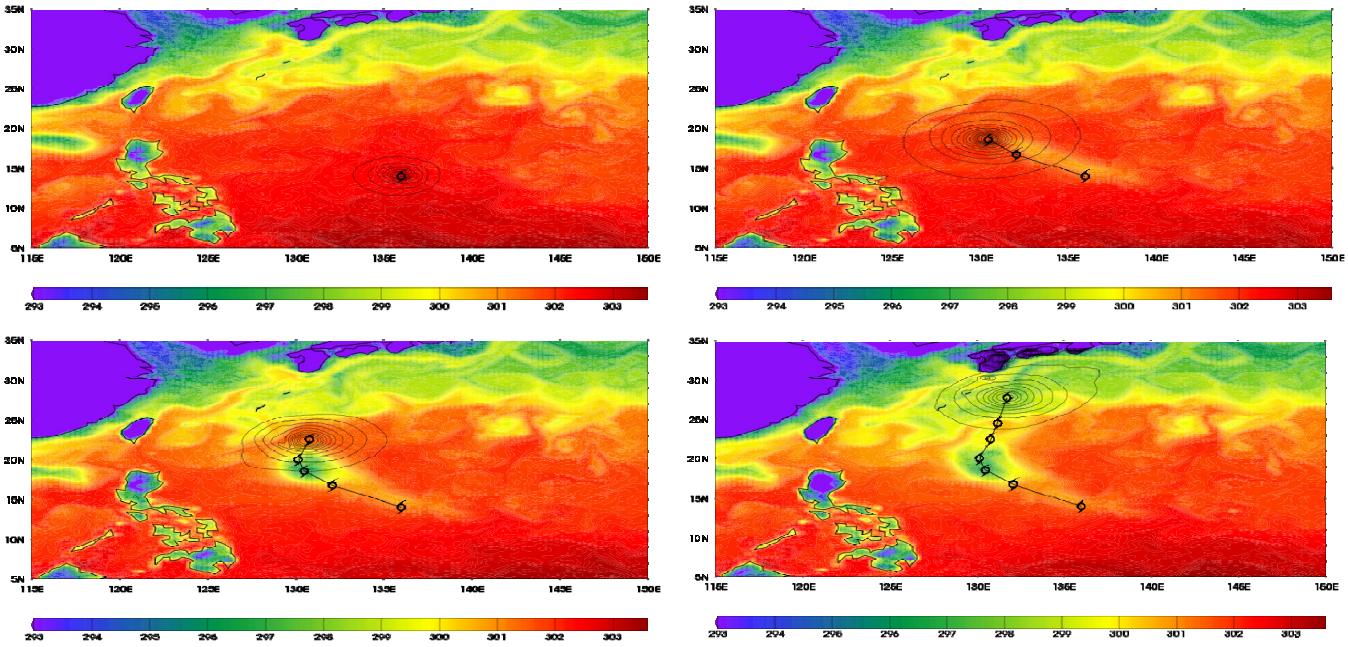

Figure 1: A simulation of a self-generated Category 4 tropical cyclone at Day 0 (A), Day 2 (B), Day 4 (C), and Day 6 (D) from the modeling experiment described in McClean, et.al. (2010). The model uses 0.25-degree grid spacing for the atmosphere and 0.1-degree spacing for the ocean. The colors show sea-surface temperatures and the contour lines display surface pressure. At this resolution, the phenomenon of cold water upwelling produced by the storms winds can be realistically simulated, and it appears as a cold water "wake" behind the storm track. (Source: Charles Doutriaux, LLNL, 2008. This work was performed under the Contract DE-AC52-07NA27344.)

required radiation balance calculations with long wave absorption in the atmosphere. In 1976, NCAR became the first recipient of a Cray 1
supercomputer performing at a rate of 4 Mflops. By 1981, Hansen projected a cooling effect of aerosols in the atmosphere. The Cray X-MP, introduced in 1982, benchmarked at 21 MFlops. In 1991, the global cooling effects of the Mt. Pinatubo eruption were predicted correctly. The largest

The climate models used in the most recent IPCC assessment showed unequivocally that human activities are responsible for the change in the global mean climate, but they are unable to provide regional information suitable for adaptation to climate change.

computer in the world, as described in the Top500 list, crossed the 100 gigaflop line in 1993, and by 1997 the first teraflop machines had arrived. This additional power allowed coupled three-dimensional ocean and atmosphere models to be explored, and in 2001 the observed warming of the ocean basins was explained using simulation. In 2002, the fastest computer in the world was the Japanese Earth Simulator with a peak speed of 40 Tflops. The Kyoto treaty went into effect in the same year (2005) that Hurricane Katrina raised new questions about regional effects of global warming. The Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report of 2007 utilized massive computing resources in an international effort to bound the possible future consequences of climate change, but the questions about implications for the environment only intensified. In 2009, the first Petaflop computer became available and, in 2010, the first generation Earth System Model (CCSM4) was released, coupling ocean, atmosphere, land, ice and the carbon cycle with terrestrial and oceanic ecosystems. The research focus of the NSF expanded to include shorter-term decadal climate predictions.

Despite progress in observing, understanding, and modeling the climate, the current generation of climate models have reached a plateau in their ability to simulate salient features of Earth's climate. The models cannot discriminate climate change signals observed in different parts of continents, nor can they provide the detailed regional information that is critically needed for developing regional adaptation strategies. Worse, the current models have large systematic errors in critical parts of their framework for the global climate system, and they severely underestimate the variability of weather and climate. As a result, the models may fail to predict the extremes that have the largest impact on human society and natural ecosystems.

In 2008, the international weather and climate modeling community came together at the World Modeling Summit (WMS) and reached a consensus: the time is ripe to revolutionize the application of numerical models to the prediction of climate through the development of seamless prediction methodologies, that unify the weather and

The challenge for high-performance computing is formidable and feeds into challenges in software, data management, analysis, and visualization, as well as the necessarily virtual global organization that must work across national boundaries to develop the models and evaluate their output.

climate forecast problems [68]. At the heart of the WMS findings was the hypothesis that the ability to resolve important processes in the atmosphere and ocean and at the land surface, as well as the interactions among them (already the case in weather prediction models), can dramatically improve the fidelity of the climate models.

A report from the WMS called for a revolution in climate modeling that would begin by establishing multiple international high performance computing facilities, for example, one each in the Americas, Asia, and Europe. These facilities would be virtually interconnected and dedicated to the development and application of high-resolution climate models. The global climate models would be capable of resolving clouds, ocean eddies, the variations of the landscape, and the cracks and seams in sea and land ice. Such models would require a spatial resolution of a kilometer or less and be able to run for simulated centuries or longer within a few days of wall-clock time. The challenge for high performance computing would be formidable, and it would also feed into modeling challenges related to software and data management, analysis, and visualization. Equally important is the challenge of creating a global virtual organization in which institutions can work together to develop models and evaluate their output.

As will be explained in Chapter 3, predictive models of climate require extensive amounts of accurate data. The data are necessary to quantify uncertainty and to enable meaningful validation and verification of processes. Not only do the data provide a detailed record of relevant physics and chemistry of the environment, but they must also adequately inform the complex computational models.

In summary, the Grand Challenge here is to improve our understanding of weather and climate by building the next generation of models. Those models must accurately reflect and predict climate conditions at the regional decision-making scale, and they must include the full distribution of weather events that compose the delivery system of climate. In addition, we must organize our efforts, via CI, at regional, national, and global levels to address the pressing problem of global climate change.

2.3 Human Sciences and Policy

New sources of data and new means for analysis are transforming the human sciences in ways that advance knowledge, solve grand challenges, and inform policy. Archives of text -- historical and contemporary – can be examined using automated information extraction from digitized libraries, blogs, email messages, speeches, government reports, and other web sources. Data from individual-level registration, primary participation, campaign contributions, ballot images and automated precinct-level result reporting can inform electoral studies. Credit card and real estate transactions, RFID product tracking and geographic location information from cell phones or toll booths using transponders (*e.g.,* Fastlane or EZPass) can be used to study commercial behavior. Digital medical records, hospital admittance data, and location-based data might transform our understanding of health care.

New sources of data and new means for analysis are transforming the human sciences in ways that advance knowledge, solve grand challenges, and inform policy.

Developments in genomics, proteomics, metabolomics, and brain imaging allow study of person-level variables never before possible. Increasingly powerful models allow the study of phenomena from person to globe, and in reverse, pushing beneath the person to the organ, cell, or something even smaller. Cyberinfrastructure can bring to researchers data on single individuals as well as networks of individuals. Satellite pictures of human-generated light at night and networks of roads and other infrastructure by day can provide opportunities to study phenomena not previously observable. New techniques allow the exploitation of such resources without infringing on personal privacy or causing similar social problems.

These changes bring both opportunities and challenges. As more powerful and more widely applicable knowledge arises, new scientific challenges appear that require advances in cyber science and engineering to resolve them. The human sciences are part of this remarkable change, as illustrated by three examples: societally-informed climate models, global-scale epidemiological models, and understanding human networks.

2.3.1 Societally-Informed Climate Models:

A new generation of computationally intensive models is needed to represent processes such as cloud formation at finer scales, "well enough to provide the sorts of prediction that policy-makers and other stakeholders need" [62]. Human activities such as land clearing, urban expansion, and agriculture create complex mosaics of highly fragmented land cover that become increasingly important as modeling is refined. Equally important, such refinement makes it possible to study how human activity is altered by climate change. Global warming might affect zones of agriculture, sea level rise might alter affect coastal urbanization, and reductions in rainfall might make some areas uninhabitable. Human-climate feedback is essential to climate dynamics, but current models cannot meet the need Meso-scale climate models often rely on 'scenarios' of land cover based on assumed conditions rather than actual data, and produce compromised short- and long-term predictions of climate change. A new generation of models must explicitly incorporate social processes and critical feedback between human activity and the climate system. This requires:

A new generation of computationally intensive models is needed to represent processes such as cloud formation at finer scales, "well enough to provide the sorts of prediction that policy-makers and other stakeholders need".

- Coherent databases of social activities (*e.g.,* demography, transportation, and other factors) for use in climate-change modeling to provide guidance for land managers, policy makers, commuters, agriculturalists and others who make decisions based on environmental conditions. This requires the development of ontologies to facilitate data integration and exchange.
- High performance computing capable of handling state-of-the-art climate models and dynamic models of land cover change, emissions, urban growth and other effects of human activity.

- Calibration and validation of packaged modeling products for decision-makers to address uncertainties in model results. Land use modeling must be incorporated effectively into the climate models now serving as benchmarks.
- Data visualization for understanding and explaining complex human and climate dynamics arising from this new generation of models to provide tangible, accessible and comprehensible explanations of model results. Abstract representations do not provide policy makers and stakeholders with realistic understanding of their own geographic 'backyards.'
- Reconciliation of the three-dimensional structure of climate models with the two-dimensional structure of land use models.

2.3.2 Global Scale Epidemiological Models:

A global pandemic could kill millions, enabled by rapid spread of pathogens in the jet age. Computational power linked to data streams about human movement has spawned the field of computational epidemiology, but it lags behind weather and climate modeling. The granularity of data necessary to predict disease spread is not yet known, nor are the means to model the micro-movements of individuals and macro flows of groups. Effective modeling of global disease spread would probably surpass current cyberinfrastructure capability, and work must be done to enable global-scale epidemiological models that allow researchers, medical practitioners, and public officials to implement mitigation strategies. Human movement and how human behavior might change given exposure to particular pathogens under differential social and behavioral conditions must be incorporated into models, using a biological framework of transmission probabilities for particular pathogens. Such models would require integration of heterogeneous data regarding human movement and behavior of individuals within communities, ranging from commuting patterns in India to school attendance in the US. Computational power and effective algorithms for modeling the movements of billions of human and nonhuman actors (*e.g.,* animal disease vectors) must be developed. All of this must be done with careful attention to the sensitivity of individual location and movement in order to prevent infringement of privacy, including measures to keep individual identities hidden during data collection and analysis.

Effective modeling of global disease spread would probably surpass current cyberinfrastructure capability, and work must be done to enable global-scale epidemiological models that allow researchers, medical practitioners, and public officials to implement mitigation strategies.

2.3.3 Understanding Human Networks:

Much human behavior involves networks of individuals, groups, communities, and societies. The challenges discussed above involve human network behavior, and other challenges depend on this as well. Recent research has demonstrated the ability to analyze small to moderate-sized networks and understand why people gain weight, express political views or communicate as they do with colleagues or friends. Network research offers the opportunity to understand collective intelligence in knowledge accumulation (*e.g.,* Wikipedia), prediction of event outcomes (*e.g.,* the Iowa Electronic

Markets), or the sourcing of engineering solutions (*e.g.,* InnoCentive). Cyberinfrastructure enables the phenomena and the means to study them, but creates challenges as well. One challenge is the analysis of very large social networks involving network ties of variable strength and duration, as well as greater information about individuals who are connected in such ways.

Improved understanding of human networks is key to increasing the value of investments in science, along the path leading from knowledge to innovation to economic welfare. Current scientometric analyses focus on authors, institutional affiliation, topic, publications and patents or other simple variables. Future analyses will include complete individual biographies with educational and employment history, histories of scientific activity, and connections between scientists and those within and outside their professional worlds. Human connections are sometimes contained within boundaries that can be drawn easily (*e.g.,* organizational networks), but human connections are often complex, starting from an individual and moving outward to ties that increasingly exist in "virtual" worlds such as cyberspace. Cyberinfrastructure provides the potential to link existing information sources (databases, published literature) with data from social networks, distributed sensor webs, and other sources in ways that could revolutionize the human sciences. This development would also place huge demands on cyberinfrastructure and present fundamentally new challenges such as reworking the science of sampling (*e.g.,* studying part of a population with confidence that the sample represents the whole) and understanding the multi-faceted nature of social network ties and their effects on human behavior. These challenges are welcomed by the human sciences.

Improved understanding of human networks is key to increasing the value of investments in science, along the path leading from knowledge to innovation to economic welfare.

2.4 Macromolecular Structure and Complexes

Biology can anticipate unprecedented opportunities in the 21st century, because it stands to benefit enormously from the confluence of three trends in scientific methodologies: advances in experimental techniques in biomolecular structure determination, progress in theoretical modeling and simulation for large biological systems, and breakthroughs in computer technology. Experimental data can now be analyzed and interpreted further by modeling,

To study such biomolecular systems successfully and reliably, new methods and models need to be systematically developed: force fields, hybrid quantum/molecular mechanics models, enhanced sampling techniques, rigorous coarse graining of multiscale models and integration of all these tools to allow "telescoping" from one level of resolution to another.

and predictions for different approaches can be tested and advanced through computational methodologies and technologies.

Markedly enhanced computational resources will allow systematic solutions of various important biomolecular problems. In the increasing complexity of temporal and spatial dimensions, such problems include macromolecular folding, biochemical binding and reaction mechanisms, macromolecular pathways, and supramolecular cellular processes. Prominent examples of macromolecular folding are protein folding and RNA folding. Examples of reaction mechanisms include enzyme catalysis and protein/ligand interactions. Macromolecular pathways include DNA replication and repair fidelity, protein synthesis, chromatin organization, and RNA editing. Supramolecular cellular processes include protein signaling networks, plant cell wall formation, and endocytosis.

If the study of such systems is to be successful and reliable, new methods and models need to be systematically developed, including the use of: force fields, hybrid quantum/ molecular mechanics models, enhanced sampling techniques, and rigorous coarse graining of multiscale models. All of these tools must be integrated to allow "telescoping" from one level of resolution to another to focus on specific details. In concert with these developments, infrastructural support for generating and analyzing voluminous molecular data requires development of simulation management tools for clustering, archiving, comparisons, debugging, visualization, communication, and replication. Such new capabilities must be developed in a focused manner to avoid computational bottlenecks (*e.g.,* the microsecond timescale for protein folding due to long-range intermolecular interaction computations, or the lack of rigorous coarse-graining models to allow scaling up to macromolecular pathways and supramolecular cellular processes).

2.5 Hazard Analysis and Management

Hurricanes, earthquakes, tornadoes, contaminant releases, wildfires, or incendiaries – all of these catastrophic events have disruptive implications for society and must be properly managed. Keys to hazard management are, first, the ability to predict a

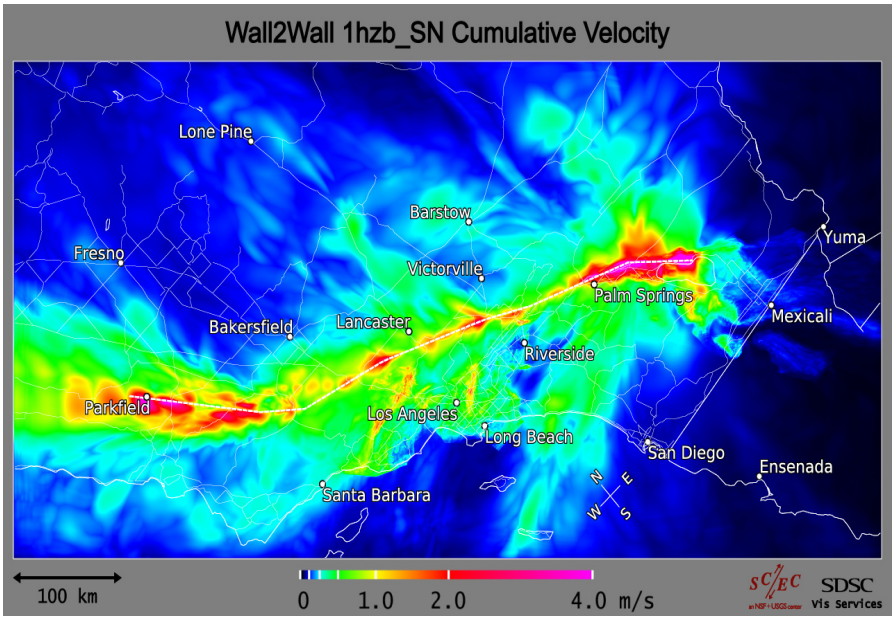

Figure 2: Peak ground velocities for a southeast-to-northwest Mw8.0 scenario on the San Andreas fault from the Salton Sea to Parkfield ('Wall-to-Wall'). The simulation computed 350 s of wave propagation in a 800 km x 400 km x 100 km subset of the SCEC Community Velocity Model (CVM) V4 (32 billion grid points with a spacing of 100 m everywhere) and a minimum shear-wave velocity of 500 m/s up to a maximum frequency of 1 Hz. The source description was generated by combining several Mw7.8 dynamic source descriptions. 'ShakeOut-D'The simulation used 96,000 NICS Kraken cores, took 2.6 hours wall clock time. (Source: SCEC, Nov 2009)

wide range of possibilities for *a priori* planning and, second, the means to perform simulation and near-term prediction to support decision-making strategies to manage specific issues. High resolution models of the physics that are calibrated by sparse observation data and laboratory scale experiments need to integrate methods of uncertainty quantification to predict the effects of the extreme scale. Multiscale and multiphysics methodologies are essential to our ability to represent the complex physics inherent in most of these phenomena. Such models of physics must be systematically coupled to social and behavioral models of public actions that affect populations.

In many of these hazard applications (*e.g.,* storm surge computations using the ADCIRC code) the basic physics model evaluations for a single computation often consumes O(1) hour on a petaflop class computer, i.e., simulations require 10^{15} FLOPS and commensurate memory. Predictive simulations will require ensembles of hundreds if not more of these in hours with appropriate analytics on the outcomes of these computations. Among the computational needs here are thus the ability to do vast ensembles of simulations in a timely fashion and the ability to integrate the high data volume outputs of these simulations into usable predictions using appropriate analytics. A second major issue of hazard analysis is the lack of observational data at extreme scale (*e.g.,* data on Category 5 hurricanes, 9.0-magnitude earthquakes, or 10^9 m^3 volcanic eruptions are not readily available). Consequently, predictions have to rely on very large ensembles of models and high resolution simulations with quantified error and uncertainty.

2.6 Managing Greenhouse Gases

There is consensus in the scientific community that increased levels of greenhouse gases – particularly carbon dioxide – can adversely affect the global climate. A main contributor to the increasing atmospheric concentration of CO_2 is fossil fuel combustion for power generation. The demand for energy is expected to grow in developed and, in particular, developing countries. Alternative fuels are unlikely to replace fossil fuels in the short term, and fossil fuels will be in demand for the foreseeable future.

One of the most promising approaches for reducing atmospheric CO_2 is geological sequestration, that is the injection of CO_2 into deep brine aquifers and oil and gas reservoirs. In geological sequestration, CO_2 from power plant emissions, natural gas fields, and other sources is captured, compressed, and injected as a supercritical fluid into deep brine aquifers and depleted oil reservoirs.

Predictive computational simulation may be the only means to account for the lack of complete characterization of the subsurface environment.

While geological sequestration is a proven means of permanent CO_2 storage, it is difficult to design and manage such efforts. Predictive computational simulation may be the only means to overcome problems from the lack of complete characterization of the subsurface environment, the multiple scales of the various interacting processes, the large areal extent of saline aquifers, and the need for long-term predictions. Key issues for modeling CO_2 injection in saline formations

are the large uncertainty in predicting subsurface CO_2 flow rates which is the direct result of uncertainty in characterizing formation permeability and porosity and multiphase fluid behavior as a function of pressure and temperature. The flow of CO_2 is dominated by gravity and viscous forces during the injection period, whereas gravity and capillary forces dominate any movement of CO_2 after the injection has ceased. Computational capabilities at the peta- and exascale will be necessary for the type of predictive simulations needed.

2.7 Assembling the Tree of Life

Knowledge of evolutionary relationships is fundamental to biology. Those relationships are captured in the form of phylogenetic trees. A grand challenge for biology is to reconstruct the detailed shape of the "tree of life" – the phylogeny of all known organisms. Such phylogenetic trees help us understand and predict

- Functions of and interactions between genes,
- Relationships between genotype and phenotype,
- The co-evolution of hosts and parasites,
- The origins and spread of disease,
- Drug and vaccine development, and
- The origin and migrations of human populations [36].

Figure 3 shows small fragments of the tree of life, those concerning (a) the relationships among herpes viruses that affect humans, (b) the evolution of the West Nile Virus, and (c) the relationships among antivenins for various poisonous snakes [36].

The process of reconstruction begins with descriptions of species (behavior, morphology, metabolism, and DNA) and

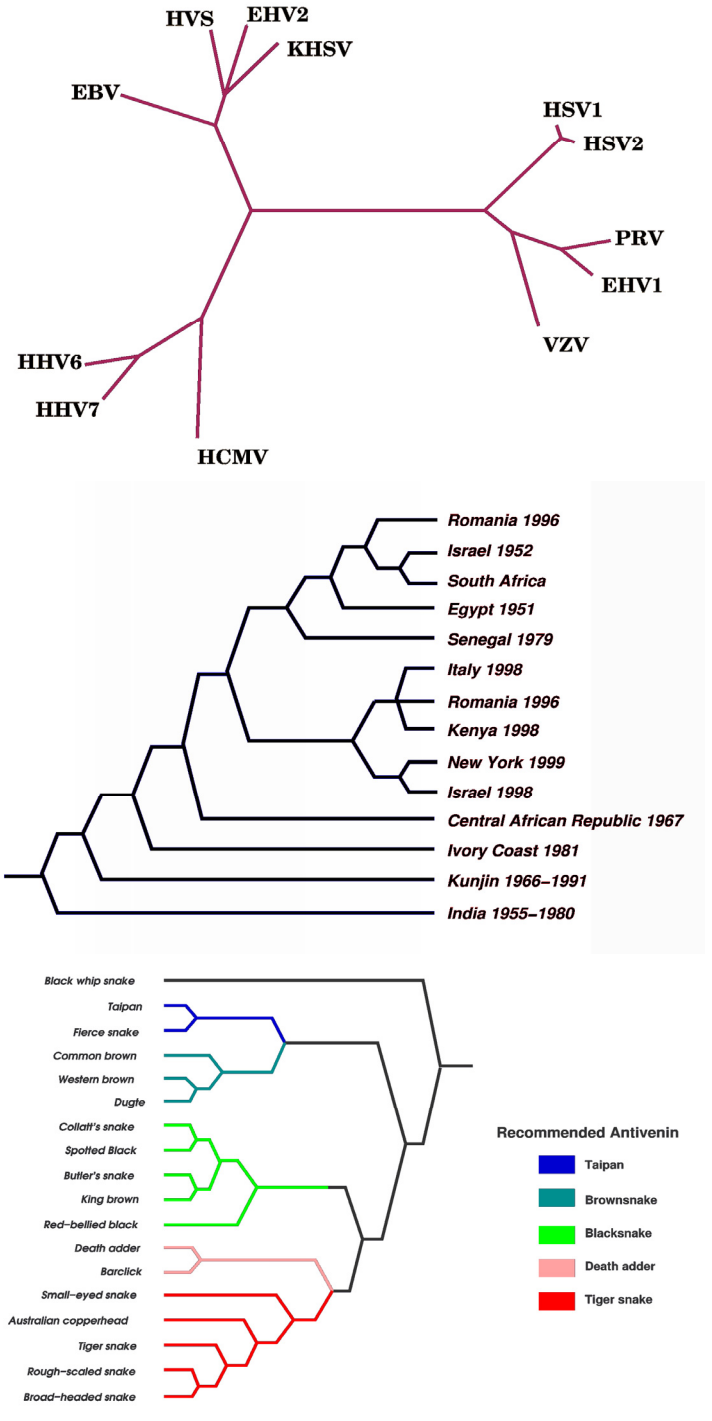

Figure 3: Phylogenetic trees for (a) Herpes viruses (b) West Nile Virus, and (c) snake antivenins [36]. Images courtesy of Bernard Moret and Joel Cracraft.

models of evolutionary processes (speciation, population evolution, molecular character evolution, etc.) and then proceeds to search the space of all possible trees to find the tree that conforms best to various criteria such as maximum parsimony, maximum likelihood, and minimal evolution (including distance-based methods). Because of the vast number of possible trees, most algorithmic formulations of this problem are NP-Complete, and immense computing resources are required to construct even relatively small trees (*e.g.,* involving 100-500 taxa) [3]. Furthermore, many algorithms construct an unrooted tree that provides only a partial constraint on the evolutionary processes that produced the observed variations across taxa. Moving from our current capability to handle 10s-100s of taxa to the ultimate requirement to handle the estimated 10-200 million species on the Earth will require major advances in both HPC and algorithm development. In addition, assembling the data to describe these species is a major undertaking that will involve the development of methodologies and strategies for prioritizing which species should be included and in what order. Tools and methodologies (simulation, visualization, etc.) are also needed to validate the algorithms and the resulting phylogenetic trees.

To achieve those goals, we must build a Grand Challenge Community that includes scientists in phylogenetic biology and computer science and engineering. Initial efforts in this direction include the iPlant community [26] and CIPRES (Cyberinfrastructure for Phylogenetic Research) [10].

2.8 Gamma Ray Bursts

Ninety years after Einstein first proposed his General Theory of Relativity (the GR), astrophysicists are probing deeper into regions of the universe where gravity is very strong and where, according to GR's geometric description, the curvature of spacetime is large.

Regions of strong curvature are notoriously difficult to investigate with conventional observational astronomy, and some phenomena might bear no observable electromagnetic signature at all and may only be visible by neutrinos (if sufficiently close to Earth) or by gravitational waves - ripples of spacetime itself that are predicted by Einstein's GR. To date, gravitational waves have not been observed directly, but gravitational-wave detectors (*e.g.,* LIGO [30], GEO [19], and VIRGO [64]) are in the process of reaching sensitivities sufficiently

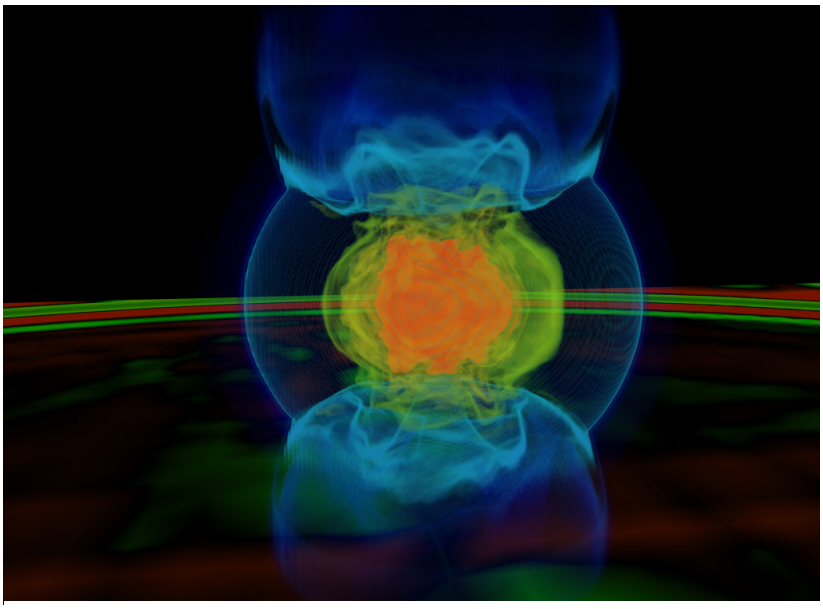

Figure 4: A rapidly spinning deformed newborn neutron star at the center of a dying massive star that may produce a Gamma-Ray Burst. Simulation by C. D. Ott (Caltech), rendering by R. Kaehler (ZIB/KIPAC).

high to observe interesting astrophysical phenomena.

Until gravitational-wave astronomy becomes a reality, astrophysicists must rely on computationally and conceptually challenging large-scale numerical simulations. Simulations allow us to grasp the details of energetic processes occurring in regions of strong spacetime curvature that are shrouded from direct observation in the electromagnetic spectrum by intervening matter or that have little or no electromagnetic signature at all. Such astrophysical systems and phenomena include the birth of neutron stars (NSs) or the collapse of evolved massive stars into black holes (BHs), the coalescence of compact[3] binary systems, gamma-ray bursts (GRBs, [41]), active galactic nuclei harboring supermassive black holes, pulsars, and quasi-periodically oscillating NSs (QPOs).

Of those phenomena, GRBs, intense narrowly-beamed flashes of gamma rays of cosmological origin, are among the most scientifically interesting, and the riddle concerning their central engines and emission mechanisms is one of the most complex and challenging problems of astrophysics today. GRBs last between 0.5 to 1000 s, with a bimodal distribution of durations [34], indicating two distinct classes of mechanisms and central engines, one known as short-hard (duration less than 2 s) and the other known as long-soft (duration 2 to 1000 s).

Hypotheses regarding these classes exist, and while observations are aiding our theoretical understanding, much that is said about the GRB central engine will remain speculation until it is possible to generate self-consistent models of the following: (a) the processes that lead to the formation of the GRB central engine and (b) the way the central engine utilizes gravitational (accretion) and rotational energy to launch the GRB jet via magnetic stresses and/or polar neutrino pair-annihilation processes. The physics necessary in such a model includes general relativity, relativistic magneto-hydrodynamics, nuclear physics (describing nuclear reactions and the equation of state for dense matter), neutrino physics (weak interactions), and neutrino and photon radiation transport. In addition, it is necessary to adequately resolve physical processes with characteristic scales from about 100 meters near the central engine to about 5 to 10 million kilometers, the approximate radius of the collapsar progenitor star.

It is necessary to adequately resolve physical processes with characteristic scales from about 100 meters near the central engine to about 5 to 10 million kilometers, the approximate radius of the collapsar progenitor star.

Any comprehensive approach to GRBs must naturally draw upon techniques and tools both from numerical relativity and from the theory of core-collapse supernovae and neutron stars. Furthermore, both areas have had dramatic progress in the past decade. In numerical relativity, immense improvements in the long-term stability of 3D GR vacuum and hydrodynamic evolutions (*e.g.,* [1, 38]) allow, for the first time, calculations for long-term stable binary black hole merger, binary neutron star merger, and neutron star and evolved massive star collapse. For its part, Supernova theory has made giant leaps from

[3] The term "compact" refers to the compact-stellar nature of the binary members in such systems: white dwarfs, neutron stars, black holes.

spherically symmetric (1D) models with approximate neutrino radiation transport of the early 1990s to Newtonian or approximate-GR to 2D and the first 3D [18] calculations. Those calculations address detailed neutrino and nuclear physics and energy-dependent multi-species Boltzmann neutrino transport [6], neutrino flux-limited diffusion [7], and magneto-hydrodynamics [8].

This modeling cannot be fully realized on present-day computers. By computing at multiple sustained petaflops of performance, however, we would be able to tackle the full GRB problem and build complete numerical models whose output could be compared with observations. Current terascale codes, such as the spacetime evolution code Ccatie and the GR hydrodynamics code Whisky, can be and have been applied to the realistic modeling of the inspiral and merger phase of NS–NS and NS-BH binaries, the collapse of polytropic (cold) supermassive NSs, and the collapse and early post-bounce phase of a core-collapse supernova or a collapsar. As the codes are upgraded and readied for petascale applications, the remaining physics modules will be developed and integrated. At that time, energy-dependent neutrino transport and magneto-hydrodynamics, both likely to be crucial to the GRB central engine, will be given high priority.

This modeling cannot yet be fully realized on present-day computers. Computing at multiple sustained petaflops of performance will allow us to tackle the full GBR problem and provide complete numerical models whose output can be compared with observations.

To estimate roughly the requirements for a full collapsar-type GRB calculation, we assume a Berger-Oliger-type [4] adaptive-mesh refinement setup with 16 refinement levels, resolving features with a resolution from 10,000 km down to 100 m across a domain of 5 million cubic km. To simplify, we assume that each level of refinement has twice the resolution as the previous level and covers approximately half the domain. Taking a base grid size of 1024 and 512 3D grid functions, and storing the curvature and radiation-hydrodynamics data on each level, we estimate a total memory consumption of about 0.0625 PB (64 TB). We compute the number of time steps that are necessary to evolve for 100 s in physical time by assuming a time step that is half the light-crossing time of each grid cell on each individual level. Therefore, the base grid has to be evolved for about 6000 time steps, while the finest grid will have to be evolved for 2^{15} steps, which is a total of $(2^{16}-1) \times 6000$ updates of the 1024^3 points. Current best practice codes require approximately 10K FLOPs per grid point per time step. When we assume that additional physics (neutrino and photon radiation transport and magnetic fields, some of which may be evolved with different and varying time-step sizes) requires, on average, an additional 22K FLOPs, one time step of one refinement level requires 50 TFLOPs.

Summing up over all levels and time steps, we arrive at a total of about 18 million

Summing up over all levels and time steps, we arrive at a total of about 18 million PFLOPs needed to run a single simulation. On a machine with 2 PFLOPS sustained, this will take about 100 days, using the full machine.

PFLOPs needed to run a single simulation. On a machine with 2 PFLOPS sustained, this will take about 100 days, using the full machine, and assuming that no faults occur and no other jobs need to use the system. For this reason, GRBs pose a true petascale problem.

2.9 Virtual Product Design for Manufacturing Industries

Engineering innovation in almost every discipline has been revolutionized through the use of virtual models to replace the construction and testing of expensive prototypes, leading to dramatic cost reductions and reduced design cycle times, and resulting in more competitive designs. Historically, engineering product development in areas as diverse as aircraft aerodynamics, automotive crash simulation, nuclear reactor core analysis, and semiconductor design has been an important driver of CS&E, as well as HPC technology. However, recent studies have revealed that, apart from a select group of industries and/or organizations, the adoption of advanced CS&E technology has essentially stagnated in most engineering disciplines [24, 66]. For example, in aerospace engineering, computational fluid dynamics has progressed over the last 30 years from simple panel methods in the 1970's to Reynolds averaged Navier-Stokes models in the 1990's, but it has not embraced more complex and expensive large-eddy simulations or other multi-physics simulations. Instead, the discipline has chosen to reduce the cost of a fixed-simulation capability rather than to explore the potential of higher fidelity simulations on leading-edge hardware [33]. In most cases across diverse application areas, component-level analysis involving single-physics simulations on commodity hardware represents the state of the practice. Since foreign industrial competitors are investing aggressively in advanced CS&E methodologies, these findings carry important implications for national competitiveness [66].

An aggressive insertion/adoption of CS&E methods into the product development cycle, including approaches such as comprehensive high-fidelity multiphysics simulations, numerical optimization for non-intuitive and better designs, and uncertainty quantification for reliable and certifiable product design, constitutes a Grand Challenge that offers the potential for large gains in efficiency, cost reduction, and overall competitiveness. Common barriers to increased industrial adoption of high performance CS&E include the lack of effective simulation software, overall software and hardware costs, lack of suitable manpower and demonstration of provably

An aggressive insertion/adoption of CS&E methods into the product development cycle, including approaches such as comprehensive high-fidelity multiphysics simulations, numerical optimization for non-intuitive and better designs, and uncertainty quantification for reliable and certifiable product design, constitutes a Grand Challenge that offers the potential for revolutionary gains in efficiency, cost reduction, and overall competitiveness.

beneficial return on investment in the short term. A Grand Challenge in virtual physics-based product development can serve to illustrate the potential of leading-edge CS&E in the product development cycle, while at the same time serving to advance the development of new enabling techniques and software targeted at emerging exascale hardware.

2.9.1 Turbomachinery Engine Design:

As an example, in the aerospace industry, current aircraft turbofan engine design relies heavily on zero dimensional cycle models with maps that represent the different engine components, such as compressor, turbine or combustor. These components themselves are traditionally designed with low dimensional models, although more recently three-dimensional steady-state Reynolds-averaged Navier-Stokes computational fluid dynamics simulations have been used at the component level. High fidelity unsteady component simulations are currently pushing the state of the art largely due to the geometrical and physical complexity present even at the component level. For example, a full compressor or turbine simulation may contain 10 to 30 rows of fixed and rotating blades with up to 100 blades or more per row. Current day simulations of these types of configurations using on the order of 100 million grid points, with sliding grid interfaces, and fully implicit time-stepping strategies, can be run on several thousand processors requiring on the order of 10^{15} flops and 10^{12} bytes of memory. However, the incorporation of higher resolution and additional physics, made possible with the advent of petaflops and exaflops capabilities, will result in dramatic advances in simulation predictive capability. For example, because turbine blade operating temperatures are directly linked to failure rates (i.e., a 20°C rise in blade temperature corresponds to a 50% reduction in blade life), the simulation of cooling flows and associated conjugate heat transfer from first principles will have a dramatic effect on engine component performance predictions.

The combination of multiphysics turbine simulations with a full unsteady compressor simulation and a large turbulence eddy resolving simulation of the combustor, including fuel spray and complex combustion chemistry, will clearly require exaflop level resources.

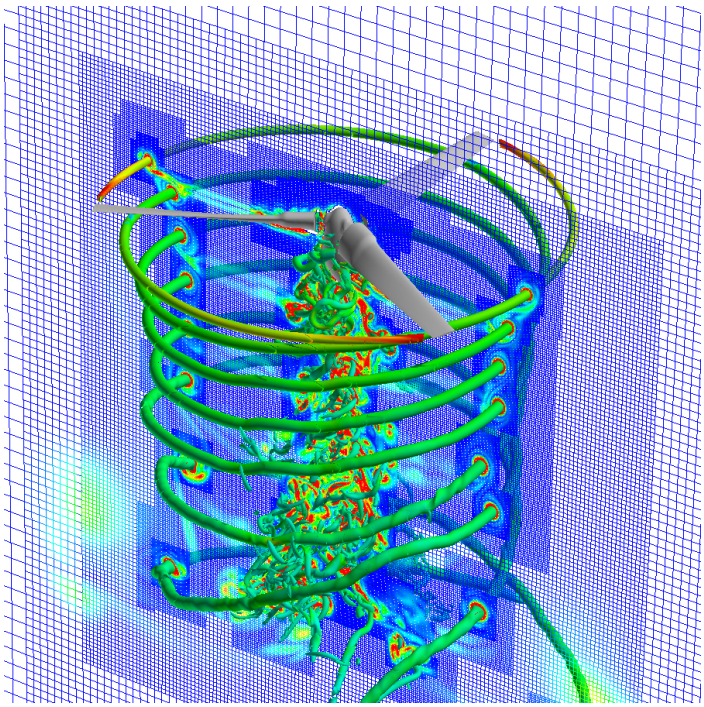

Figure 5: High fidelity simulation of helicopter rotor dynamics. Simulation includes a rotating unstructured mesh fixed to the blades and hub and a fixed Cartesian mesh in the off-body region where a high (6^{th}) order accurate discretization adaptive mesh refinement strategy is used for accurately capturing the wake vortices. The overlap and interpolation patterns between fixed and rotating meshes are recomputed in parallel at each time step. Reproduced from [67].

A typical high pressure turbine blade can contain up to 400 cooling holes and it has been estimated that 1 million grid points are

required to simulate a single cooling hole flow to sufficient accuracy for conjugate heat transfer predictions. Simply based on the number of rows and blades per row, this would translate into over 10^{12} grid points or more than a factor of 1000 increase in required resolution, putting such a simulation clearly in the petaflops range, requiring a total of 10^{18} flops and 10^{15} bytes of memory. The combination of multiphysics turbine simulations with a full unsteady compressor simulation and a large turbulence eddy resolving simulation of the combustor, including fuel spray and complex combustion chemistry, will clearly require exaflop level resources. However, new frontiers in product design and reliability will be enabled through the availability of such simulations especially when used in design optimization loops, and for managing manufacturing uncertainties to provide reliable estimates of fleet engine performance or life cycle wear and predictive performance degradation.

2.9.2 Wind Engineering: Wind energy represents an area that has seen aggressive use of CS&E since its inception. Current leading-edge high-fidelity wind turbine aerodynamics simulations can be achieved using on the order of 100 million grid points, with overlapping or sliding mesh interfaces, and implicit time-stepping procedures, usually limited to time steps corresponding to less than 1 degree of revolution, due to temporal accuracy limitations. Total computational requirements for such simulations, assuming the simulation of ten complete revolutions, can be estimated to be of the order of 10^{15} flops and 10^{12} bytes of memory. However, large eddy turbulence resolving simulations including blade transition effects, geographic terrain effects and atmospheric turbulent boundary layer interactions can be expected to require at least one to two orders of magnitude more resolution in space and time, putting such simulations clearly in the petaflops range. Simulating interference effects between turbines will require the inclusion of adaptive mesh refinement techniques and/or higher-order methods to capture and preserve wake and vortex effects over long distances. The simulation of arrays of turbines (for example a dozen closely spaced turbines) with terrain and atmospheric turbulence effects can thus be expected to require of the order of 10^{21} flops and 10^{15} bytes of memory. Furthermore, design optimization using these types of simulations and uncertainty quantification, due both to extreme atmospheric events and manufacturing variability, can be expected to add another one to two orders of magnitude in computational requirements.

2.9.3 Virtual Flight Testing: In aerospace engineering, complete aircraft steady-state aerodynamic analyses are now commonplace, as well as linear structural analysis of complex structures. The next logical step involves the adoption of time-dependent large-eddy simulations for aerodynamics, time-dependent non-linear structural analysis, and the coupling of these two disciplines for dynamic aeroelasticity. Furthermore, aeroacoustics and propulsion disciplines (and aerothermal in the case of hypersonic vehicles) need to be integrated into the simulation process, as well as simulation of the flight control system, in order to enable controlled virtual flight simulations. Numerical optimization techniques can then be devised to explore optimal configurations and to design flight system control laws with specified handling characteristics. Finally, the design process will require the simulation of the complete flight envelope, including cruise conditions, extreme conditions, and unanticipated emergency conditions. Building the complete

flight-envelope data-base involves hundreds of thousands of individual conditions and is currently achieved through a combination of expensive wind-tunnel testing and flight testing. The ultimate long term goal should be digital airworthiness certification.

All of these virtual product design Grand Challenges share many of the same requirements and obstacles as the other Grand Challenges described in this report. However, there are some particular issues that are specific to product design Grand Challenges. For example, the use of ever increasing spatial resolution is often not the best path forward for increased simulation outcomes in many virtual design problems. Often, the extension of steady-state simulations to time-dependent problems, and/or the incorporation of additional tightly coupled physics represent the critical elements required for increased simulation effectiveness. Additionally, the natural progression from conceptual to detail design must rely on a hierarchy of low-to-high fidelity models, all of which must work together to provide the most optimal and reliable final design. Finally, because product design is a time critical exercise, there is a limit on acceptable simulation turnaround time (often taken as 24 hours) to be useful in the design cycle. These aspects make the development of enabling analysis and optimization algorithms that scale effectively to the exascale particularly challenging. However, these problems are so complex that exascale resources will be required in order to realize the full potential of CS&E in the product design cycle.

Because product design is a time critical exercise, there is a limit on acceptable simulation turnaround time, making the development of enabling analysis and optimization algorithms that scale effectively to the exascale particularly challenging. These problems are so complex that exascale resources will be required in order to realize the full potential of CS&E in the product design cycle.

A physics-based virtual product Grand Challenge will provide a catalyst for focusing resources on the development of computational methods, implementations, and software that enable higher fidelity simulations, tight coupling of disparate physics, effective optimization strategies, and novel uncertainty quantification methods targeting risk reduction, reliability, virtual certification, and complete life-cycle assessment. In addition, the Grand Challenge will demonstrate the potential for accelerating engineering innovation and will provide the basis for reliable software that can be deployed cost-effectively across a range of hardware scales.

2.10 High-Temperature Superconductor Material Design

Superconductivity—the ability of some materials to conduct electricity without resistance—was discovered nearly a century ago in materials such as mercury and niobium-titanium alloys. The potential applications of superconductivity are innumerable—with revolutionary advances possible in such areas as power generation and transmission, grid technology, and high-speed levitating trains. However, these materials must be cooled to well below 20 K

(or -400°F) before they make the transition to the superconducting state; for this reason, they are known as low-temperature superconductors (LTSCs). So-called high-temperature superconductors (HTSCs), discovered a little more than two decades ago, require far less cooling; some copper-oxide materials, known as cuprates (an example is shown in Figure 6), are superconducting at temperatures above 77 K (or -320°F) – which is very significant, since 77K is the boiling point of the relatively cheap coolant, nitrogen. HTSCs are much more complex than LTSCs: examples of cuprate and the recently discovered (2008) iron-based HTSCs are $YBa_2Cu_3O_7$ and CeFeAsO respectively.

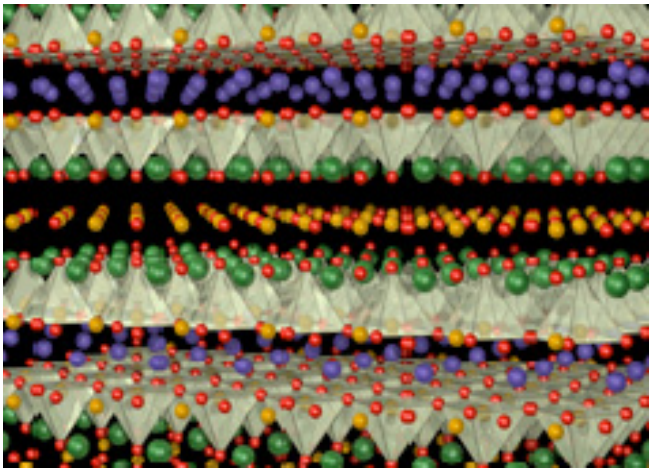

Figure 6: The high-Tc superconductor YBa2Cu3O7. Atoms are rendered as follows: O – red; Cu – copper; Ba – green; Y – pruple. Image courtesy of Jeremy Meredith, Oak Ridge National Laboratory.

Understanding the fundamental origins of HTSC behavior has been a theoretical challenge since the discovery of HTSCs. Recently, a relatively simple model that provides a coarse-grained description of the electrons in a cuprate's copper-oxide layers – the Hubbard model – has provided new insights into HTSCs. Despite its coarse-grained nature, the simulation of the Hubbard model, in common with the simulation of correlated electron systems in general, is computationally intensive. For example, the landmark simulations of the Hubbard model by a group of Oak Ridge National Laboratory researchers that showed that the model could predict HTSC successfully and delineate its fundamental mechanisms won the 2008 Gordon Bell prize for highest-performing sustained scientific computation (1.352 petaflops). However, by virtue of its coarse-grained nature, the Hubbard model does not reflect atomic composition or structure, and thus cannot predict the superconducting transition temperature of a specific material. In order to do this – i.e., material-specific modeling of HTSC candidates – the single-orbital Hubbard model needs to be turned into a multi-orbital model, significantly increasing the computational complexity, since for these calculations problem size grows exponentially with the size of the system. In addition, the robust extraction of parameter-free materials-specific multi-orbital models from first principles electronic structure calculations – a process called down-folding – requires peta-scale simulations in itself. The merger of these two programs and its embedding into an overall design methodology will result in simulations in need of exascale infrastructure.

Hence, materials-specific HTSC simulations are exascale-level computational grand challenges, which must be addressed if we are to reach the point of designing new HTSC materials. Imagine the impact on society of new HTSCs incorporated into HTSC cables that would enable resistance-less transmission of electricity around the U.S. and around the world: areas of the U.S. (or the world) that are rich in sunlight (*e.g.,* deserts) could be home to massive solar energy conversion farms that powered the rest of the U.S. (or other parts of the world). HTSCs would make possible the widespread

and cost-effective use of magnetically levitated (maglev) personal vehicles.

2.11 Common Themes to the Grand Challenges

A review of the problems typifying the Grand Challenges reveals a number of common themes. The following identifies those with clear impact on computational science and engineering.

- All Grand Challenges face barriers in the areas of software, data management, and visualization, and the coordination of the work of diverse communities that combine efforts and resources to develop models and algorithms and to evaluate the outputs.
- All Grand Challenges require transformative discovery and innovation, which in turn demand capabilities approaching or exceeding exascale computing. Computing at this scale requires dramatic changes in processor architecture and power management.
- All Grand Challenges need advanced computational models and algorithms, including methods that are informed by observational data in a way that can cope with uncertainty in the data and that can quantify uncertainties in predictions. New methods are necessary to facilitate multiscale modeling, enhanced sampling, and vast simulations while integrating high data volume outputs of the simulations along with new methods to encourage the publication of code and data to facilitate verification of computational results.
- Significantly, all Grand Challenges must have the ability to acquire relevant data for calibration and validation of large-scale computational models and to characterize and quantify uncertainties. This ability depends on the development of statistical representations of data on parameters and observations, statistical inverse methods, and software that implements them. It depends also on methods to resolve the large stochastic systems that result from model and data uncertainties. The transition of conventional deterministic methods and models of complex physical events to those accounting for uncertainties and stochasticity will increase by several orders of magnitude the size, complexity, and computational work needed for predictive simulations.
- All Grand Challenge problems call for the development – in some combination - of computational models based on scientific and engineering principles, on the principles and methods of computer science, and on computing technology and the use of core computational and applied mathematics. The advance of that combination of disciplines defines the purpose of Cyber Science and Engineering (CS&E): the discipline bringing together computational science and engineering as they can be exploited via the cyberinfrastructure.

The transition of conventional deterministic methods and models of complex physical events to those accounting for uncertainties and stochasticity will increase by several orders of magnitude the size, complexity, and computational work needed for predictive simulations.

Although NSF has supported many cross-directorate initiatives in basic CS&E over the

years, there has been no home for it within the NSF organizational structure. As a result, efforts in CS&E have been fitful: underfunded programs, low-proposal success rates, and no sustainability for efforts requiring long-term investments in software and algorithm development and infrastructure. Under those conditions, an effective attack on Grand Challenges is extremely difficult.

It is clear that important discipline-specific programs in computational science and engineering are vital to advancements in every discipline, and such problems must be encouraged and supported at NSF. But mechanisms should also be created for sustained support of CS&E across multiple disciplines (and directorates), for interdisciplinary work is an essential attribute of all Grand Challenge efforts. Also, the best work in CS&E will be built on a foundation of solid applied mathematics and computer science not always in the scope of discipline-specific approaches, while, conversely, core mathematical and computer science, by themselves, do not generally fit the needs of Grand Challenge projects. The distinction is often that new mathematics and computer science must be developed to resolve specific barriers to progress on Grand Challenge problems, and these developments are rarely anticipated as relevant topics for mathematical or computer research.

These considerations suggest that the Foundation would be best served in the broad area of CS&E if it developed policy and structures that support directorate-specific activities in CS&E, on the one hand, and, on the other hand, that support Foundation-wide initiatives that involve multiple disciplines. Those latter initiatives will always be needed for addressing legitimate Grand Challenge problems.

These considerations suggest that the Foundation would be best served in the broad area of CS& E if it developed policy and structures that support directorate-specific activities in CS&E, on the one hand, and, on the other hand, that support Foundation-wide initiatives that involve multiple disciplines. These latter initiatives will always be needed for addressing legitimate Grand Challenge problems.

3 Advanced Computational Methods & Algorithms

3.1 Introduction

Computational methods and algorithms have played a crucial role in the solution of complex scientific and engineering problems since the earliest days of computing. They form the key link between mathematical models of physical phenomena of interest and high performance software that can be used to carry out analysis and prediction of the behavior of complex physical systems. Synergistic advances in computing and computational methods have stimulated scientific and engineering breakthroughs, which have in turn motivated further advances in enabling technologies. Over the past half-century, advances in computational methods have led to speedups in the solution of important scientific problems that are as significant as those resulting from advances in the hardware alone. For example, Figures 7-10 illustrate breakthroughs on scientific problems that have been enabled by advances in algorithms.

There is no question that building an exascale machine is hard; but using it effectively to solve CS&E Grand Challenge problems is an even harder goal.

Computational methods, however, are often taken for granted due to past successes and their largely hidden role in powering CS&E software. But while recent isolated successes have occurred, computational methods that can scale to petascale systems are still in their infancy for difficult problems, such as those with strong heterogeneities and anisotropies, multiphysics couplings, multiscale/multirate behaviors, stochastic forcing, uncertain parameters, dynamically evolving

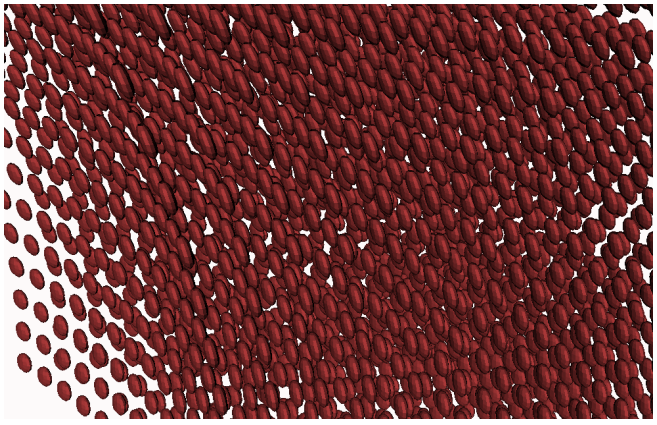

Figure 7: Direct numerical simulation of blood flow, using a complex fluid model that resolves dynamical interactions between deformable cells and surrounding fluid plasma are instrumental to gaining a better understanding of hemodynamic phenomena. The computational challenges associated with such microstructural simulations of blood flow are immense: modeling just one microliter of blood, with over four million cells, results in more than a trillion space-time variables. Work at Georgia Tech and New York University (led by George Biros and Denis Zorin and supported by an NSF PetaApps project) aims to overcome these challenges using new parallel kernel-independent fast multipole methods. The project has developed new parallel algorithms and hybrid OpenMP/MPI implementations that have enabled scalability to 200,000 cores on a Cray XT5 while achieving 0.7 PFlops/s of sustained performance.

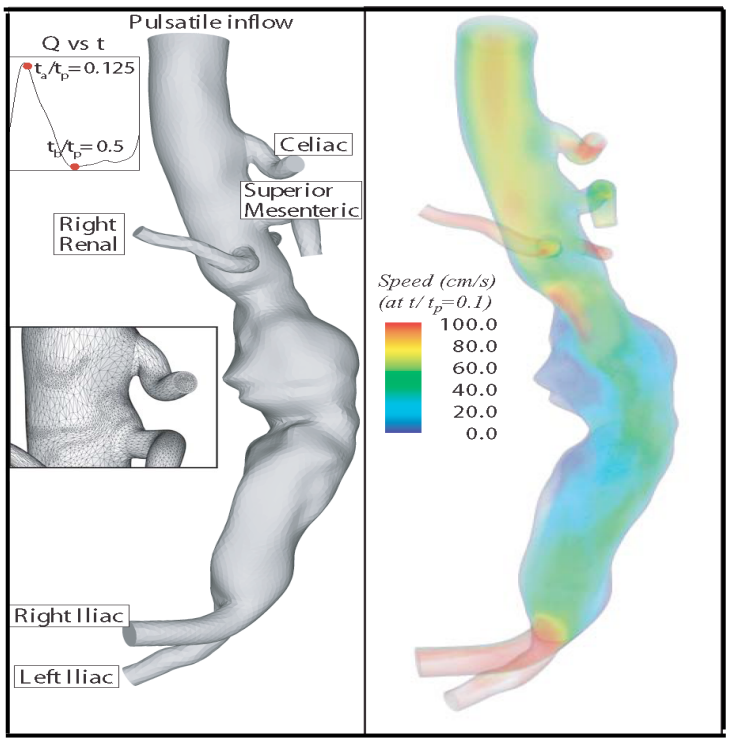

Figure 8: Simulation of blood flow in a patient-specific abdominal aortic aneurysm (AAA) model. The complexity of the geometry and physics dictates the use of adapted anisotropic unstructured AAA meshes. This presents the challenge of developing implicit methods for solving the Navier Stokes equations that scale to petascale systems. Research at RPI led by Kenneth Jansen and Mark Shephard and supported by an NSF PetaApps project has resulted in strong scaling of the PHASTA finite element flow code on a 5 billion element mesh to nearly 300,000 IBM Blue Gene/P cores with 95% efficiency. Such simulations have the potential to revolutionize planning of surgical procedures.

geometries, continuum-atomistic couplings, large-scale combinatorial structure, and so on. *But it is precisely these features that characterize next-generation Grand Challenge problems.*

Absent a systematic research effort, continued progress on frontier CS&E problems is not assured, and federal investments in hardware, networking, and software will be jeopardized. Let there be not doubt: building an exascale machine will be hard; but using it effectively to solve CS&E Grand Challenge problems will be even harder.

To address the difficulties in developing computational methods for scientific Grand Challenges such as those described in Chapter 2, a broad-based, comprehensive, long-term, and vigorous research program in advanced computational methods must be established to overcome the challenges faced in devising, analyzing, scaling up, and applying new methods for critical CS&E problems on advanced computing systems. As noted earlier, this program should support multidisciplinary and interdisciplinary teams that bring together applied mathematicians, computer scientists, and computational scientists and engineers. In turn, an additional CI challenge is to ensure that advances in computational methods and algorithms developed in one discipline are disseminated across all disciplines that face computational problems with similar structure.

Computational methods and algorithms play a key role at all stages of CS&E,

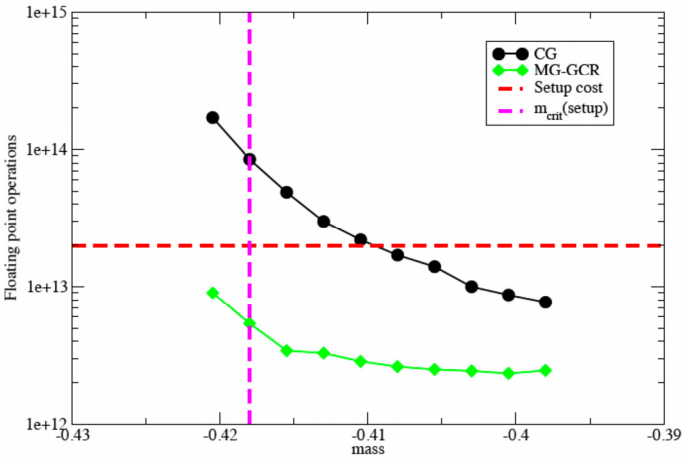

Figure 9: In another NSF PetaApps project, a team of applied mathematicians and computational physicists led by Thomas Manteuffel and Stephen McCormick (CU Boulder), James Brannick (Penn State), and Richard Brower and Claudio Rebbi (Boston University) is developing advanced multigrid algorithms for the Dirac inversion problem of lattice quantum chromodynamics (QCD). The log plot above compares a new adaptive multigrid preconditioned Generalized Conjugate Residual algorithm with a conventional QCD solver (red/black preconditioned CG) in terms of the floating point operations needed to solve the Wilson-Dirac system on a 32 X 32 X 32 X 96 lattice for various quark masses. Production parallel multigrid codes are now showing an order of magnitude speed up, nearly eliminating the problem of critical slowing down at small quark mass, which plagued all previous solvers in lattice QCD.

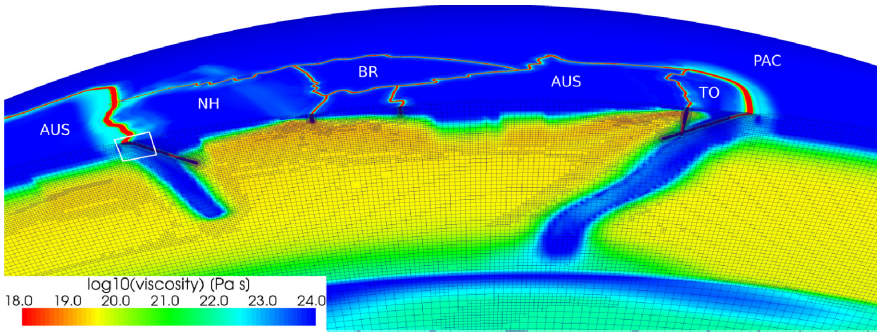

Figure 10: Portion of an adaptively refined mesh from a global mantle convection simulation with refinement both around plate boundaries and dynamically in response to the nonlinear viscosity, with plastic failure in the region from the New Hebrides to Tonga in the SW Pacific. The mesh contains elements on 7 different refinement levels globally with a finest resolution of about 1 km. The key algorithmic challenge is to overcome the difficulty of adapting meshes in parallel on the petascale supercomputers necessary for these simulations. A team led by Omar Ghattas (UT-Austin) and Michael Gurnis (Caltech) has developed parallel AMR algorithms that scale to over 200,000 cores, adapt to complex geometries, and deliver high order accuracy. These new algorithms have resulted in a factor of 5000 reduction in problem size, making tractable the global mantle convection simulations on TACC's Ranger illustrated above, and leading to new insights into the dynamics of plate boundaries.

including solution techniques for complex multiscale/multiphysics problems, advanced spatial and temporal discretization schemes for high fidelity simulations, scalable algorithms for solution of large linear and nonlinear algebraic systems and eigenvalue problems, methods for quantifying uncertainties in large-scale simulations, and algorithms for solution of large-scale optimization problems arising in design, control, and inversion. In this section, we summarize research issues in advanced computational methods that must be addressed to enable solution of frontier science and engineering Grand Challenge problems using next generation computing systems.

3.2 Simulation of Complex Multiscale, Multiphysics, Multi-model Systems

Science and engineering are increasingly concerned with the study of multiscale, multiphysics systems that intimately couple different phenomena occurring at different spatial and temporal scales and governed by different physical laws. Such systems can arise in a variety of ways. For example, systems may involve a single physical process that must be modeled using a multiscale approach. The approach couples several descriptions of the process valid at different scales, for example, deterministic and stochastic. Other systems involve the coupling of multiple physical processes described by different models. Examples include: modeling the transport of pollutants in ground water, which couples the simulation of multiple fluid phases, geomechanics, and a complex set of biogeochemical reactions; simulating a fusion reactor, which involves fluid dynamics, deformation of solid materials, thermal effects, ablation, fracture, corrosion and aging of materials, and radiation; and simulation of climate systems, which couples atmosphere, ocean, land surface, and sea/land ice models. Complex engineered systems constitute yet another class of examples. For instance, models of a regional power grid involve a mixture of a large number of continuous and integer

Science and engineering are increasingly concerned with the study of multiscale, multiphysics systems that intimately couple different events occurring at different scales and governed by different physical laws.

variables encompassing a wide range of scales coupled through descriptions that include various nonlinear dependencies and constraints.

Generically, multiscale, multiphysics systems present significant challenges for numerical simulation. It is rarely possible to simulate a complex system to such a degree that behavior is resolved uniformly at the finest scale; such systems exhibit complex stability properties resulting from a fusion of the stability properties of component physics, and the linkage between physical components has a strong impact on the model behavior. At the same time, accurate simulation of multiscale, multiphysics systems presents a challenge for high performance computing because existing paradigms for efficient use of high performance platforms for single physics models are inadequate for treating multiphysics systems. Thus, the challenge of faithfully simulating a system that encompasses a wide range of scales and physical processes requires the development of new computational algorithms that provide robust accuracy in a multiphysics, multiscale context yet scale to the millions of processor cores that characterize future multi-petaflop and exaflop systems.

3.3 Advanced Discretization Methods for Partial Differential Equations

In many areas of computational science and engineering, increasing attention is being devoted to advanced or special discretization methods. These methods yield much higher fidelity to the detailed physical description than is possible with traditional discretization methods. In these applications, standard discretization methods either miss important physical properties or achieve them at great inefficiency, even at the highest possible resolutions.

Examples of applications requiring advanced discretization methods abound. Discontinuous Galerkin (dG) finite element methods, smoothed particle hydrodynamics (SPH) methods, and hp-adaptive finite element methods that allow for locally high order discretization yet provide great flexibility in local element and discretization geometry are increasingly used for applications where the geometry of the physical domain is complicated or multiscale. Likewise, specialized methods for treating problems with dynamic interfaces and free boundaries are undergoing rapid development. Integral equation-based discretizations are increasingly deployed, motivated by advances in fast multipole methods for rapid evaluation of the relevant kernels. Many problems in science and engineering, for example, ranging from the modeling of black holes to the modeling of DNA and protein molecules to the study of

Increasing attention is devoted to advanced or special discretization methods that yield much higher fidelity to the detailed physical description than is possible with traditional discretization methods.

A central challenge in the development of these advanced discretization methods is to ensure that they map well onto forthcoming multi-petaflops and exaflops systems.

the propagation of nerve impulses, involve the evolution of physical phenomena on complex domains and manifolds. In these situations, the geometry of the domain is a critical consideration in the construction of good numerical methods. Motivated originally by the solution of Maxwell's equations, interest has intensified recently in the systematic study and use of compatible spatial discretization methods that inherit or mimic fundamental properties of the model, such as topology, conservation, symmetries, and positivity structures and maximum principles. These issues are also important in time discretization for evolution problems. So-called multirate integration methods that allow for different time steps for different components or over different regions of space are very important, for example, in reacting flow simulations, solid state circuit simulation, and biochemical network simulation. Geometric integrators that preserve properties such as a Hamiltonian structure are extremely important for simulations involving long times, such as the construction of trajectories for space vehicles.

In addition to discretization of the governing continuous equations, discretization of the geometry is a critical issue. In cases where the domain is simple, it is easy to generate uniform meshes of well-shaped elements. The generation of such meshes for geometrically complex 3-D domains, combined with anisotropic physics, can ultimately dominate the overall run time of the simulation. In addition, such meshes may yield much larger systems of algebraic equations than more optimal mesh configurations. An alternative is fully automatic unstructured mesh generation that can interact with CAD (solid model) representations to generate and adapt the more optimally configured meshes over general domains. However, these meshes require more complex data representations and can yield more poorly conditioned algebraic systems when not carefully controlled and/or combined with appropriate equation discretization methods (*e.g.,* stabilized methods). The application of high order equation discretization techniques requires the use of curved meshes for problems with curved boundaries, adding substantial additional complexity to the mesh generation process.

Although substantial progress has been made in the areas of structured and unstructured mesh generation, there are a number of critical areas requiring further development for parallel simulations. These include: generation and adaptive control of meshes that are matched to the equation discretization methods used, parallel generation of meshes of many billions of elements on massively parallel computers, effective dynamic partitioning of adaptively defined unstructured meshes, and methods for the representation and generation of properly controlled curved meshes for use with higher order methods, including consideration of the interactions of the mesh generator with the geometric model representation.

Linear solvers constitute a critical component of modern implicit scientific simulation codes and are often the barrier to scalability on massively parallel systems.

It is essential, however, as the Grand Challenge problems become increasingly complex, that the continued development of advanced discretization methods honors the underlying physics. Simultaneously, we must

ensure that they map well onto forthcoming multi-petaflop and exaflop systems.

3.4 Scalable Solvers

Solvers constitute a critical component of modern implicit scientific simulation codes and are often the barrier to scalability on massively parallel systems. Large, structured, linear and nonlinear algebraic systems and algebraic eigenvalue problems arise after discretization of complex engineering and scientific models. Overall scalability of a solver is the product of algorithmic scalability (work required as a function of problem size) and implementation scalability (which depends on having a large computation-to-communication ratio). It is often the case that large-scale scientific simulation codes spend the majority of their time in the linear solver phase, because other components usually scale linearly with problem size and require nearest-neighbor communication, while the solver typically scales superlinearly and involves global communication. Naive solvers can scale quadratically (or worse), rendering them unsuitable for the weak scaling required to capitalize on increasing numbers of processors.

In principle, linear solvers are capable of scaling well on parallel systems: for elliptic-dominated problems (and for parabolic, which resemble elliptic after time discretization), the Green's functions decay exponentially and hence effective preconditioners that coarse-grain the global communication can be designed. For hyperbolic-dominated problems, the dependencies are local. Unfortunately, a number of features of emerging computational science problems provide serious impediments to the scalability of modern solvers. These include the presence of severe anisotropies and heterogeneities, multiphysics couplings, strong nonlinearities, dynamic mesh adaptivity, interface dynamics, mixed-order discretizations, and multiscale models. For such problems, algorithmic scalability is not at all ensured, and implementation scalability is questionable due to the dynamic load balancing and significant communication required. It is absolutely critical that these challenges be overcome in order to ensure continued progress on the next-generation Grand Challenge problems as exemplified by those described in Chapter 2.

3.5 Algorithms for First Principles Models

Models that represent First Principles descriptions of physical phenomena also present numerous computational challenges. Such models generally do not involve partial differential equations. A primary example is provided by molecular dynamics (MD). MD models involve solving equations of motion of particles in order to compute statistical information such as temporal and spatial ensemble averages. This general simulation technique allows for a statistical mechanics description of matter at finite temperatures and of open systems. For example, MD simulations have proven to be a useful bridge between microscopic modeling of (bio)molecules and properties at larger scales such as elasticity, conduction, and mechanical properties of biological assemblies. Significant computational challenges include:

- Current computational capabilities limit the sampling to insufficient resolutions. The complexity and ruggedness of the energy landscape of molecular systems, with thousands to millions degrees of freedom, suggest that achieving complete sampling will remain difficult for the foreseeable future. Besides new

approaches to computing important statistics, rigorous tests of convergence of computed statistics and assessment of phase space coverage are desired for further progress in the field.

- Current descriptions of force fields are not sufficiently accurate. As the scope of MD simulations increases to longer times and larger systems, we observe significant flaws in the current energy function descriptions, which are largely empirical. Even qualitative features are wrong, *e.g.*, some proteins, that are accessible to straightforward simulations do not fold while others fold too quickly, RNA molecules are computed to be unstable, and quantitative experimental data is hard to reproduce.
- Hierarchical temporal and spatial coarse graining are necessary to overcome the gap between molecules and biological cells. We need algorithms that will help us choose the next set of variables in the (coarser) hierarchy, compute their effective interactions and assess the reliability of these models.

In the biophysics field, these advances are necessary to expand our knowledge of large protein assembly and their cell functionalities. For example, studying microtubules—the most functional cytoskeletal filaments of the cell, requires modeling of the basic building block, the protein tubulin. Rigorous coarsening based on atomistic models is a promising direction to consistent computational models for cellular behavior with limited external parameters.

3.6 Combinatorial and Discrete Problems

Combinatorial scientific computing (CSC) is the field in which researchers design graph and hypergraph solutions to solve combinatorial problems that arise for example in computational science and engineering information science, social networks, and bioinformatics, as well as create high performance software implementing these algorithms. CSC plays a critical enabling role in applications requiring parallelization, differential equations, optimization, eigenvalue computations, analysis of massive data sets, and so on. When large-scale problems demand increasing accuracy and fidelity, effective algorithms for solving combinatorial problems on emerging computing systems are needed.

Combinatorial problems have a number of features that present challenges in designing scalable parallel algorithms for their solution. The runtime of graph algorithms is dominated by communication costs and memory latency rather than processor speed. There is little work to do when processing the data at a vertex or an edge, and so computation cannot hide memory access costs. Since access patterns are determined by the structure of the input graph, prefetching techniques cannot be applied. Graph algorithms possess poor data locality, making it difficult to obtain good memory system performance. While concurrency is abundant, dependencies between computations at nodes or edges have to be satisfied, and these costs can limit the performance.

Several innovative ideas have been used to design scalable combinatorial algorithms. These include: approximation, when algorithms with higher concurrency are available if the problem can be solved approximately rather than optimally; speculation, when dependent computations are performed concurrently on multiple processors, with roll-backs if conflicts are detected; randomization to reduce the necessity of synchronization of tasks; and

partitioning, mapping, and scheduling tasks to reduce communications and synchronization costs. However, a broad long-term research effort in the design of innovative algorithms in cooperation with research in exascale architectures and applications is vital for the solution of these exascale problems.

An example of a challenging combinatorial problem in exascale computing is dynamic load balancing for multi-scale, multi-physics problems. Here computations at multiple phases of the computation need to be mapped to processors in a way that balances the sum of the computations in the phases while reducing communication and synchronization costs. In adaptive computations, the collection of computational tasks changes from iteration to iteration, and the costs of data migration have to be included among the multiple objectives of balancing the load. At the exascale, the imbalances in each phase of the computation that could be tolerated at the tera-scale would impact performance adversely, and hence dynamic load balancing would be vital for good performance. Combinatorial problems also arise in enabling solvers for linear and nonlinear systems of equations. Domain decomposition requires graph partitioning, Algebraic Multigrid solvers use combinatorial methods in coarsening grids, and preconditioners based on incomplete factorizations rely on graph models of the factorization. Another key kernel here is an algorithm for computing sparse matrix-vector multiplications, and combinatorial analysis is needed to make the computation efficient for the memory system. Automatic Differentiation (AD) is a software methodology that can compute analytic derivatives of functions, represented by programs, both accurately and efficiently. AD relies on a computational graph representation to apply the chain rule to compute the needed derivatives, and this graph is transformed to reduce the operations and storage needed for the derivative computations. The computation of Jacobians and Hessians is feasible for large-scale problems only when the scarcity and symmetry available in the computations is exploited by graph coloring models to reduce the number of passes by computing several columns or rows of the matrix simultaneously. AD has applications in nonlinear differential equations when sensitivities of the solutions are required and in Uncertainty Quantification.

3.7 Uncertainty Quantification

Because complex systems are often inaccessible to experiment and direct observation, and building and testing prototypes are extremely expensive, there is often only a small set of observational data available for analysis and predictions of behavior. Hence, a fusion of observational and experimental data with computational modeling provides the only means to gain the required understanding of complex systems. Error and uncertainty in such cases arise in many ways, for example, in data and parameters measured by experiment and observation, from discretization, and from a lack of knowledge about the physical processes in the system. Moreover, they are represented in different ways, for example, statistically, probabilistically, and deterministically. As computational modeling has become a fundamental tool in the analysis and prediction of the behavior of complex systems in science and engineering, the need to quantify the effects of error and uncertainty has become critical. This is true on scientific grounds, but in addition, computational science is increasingly used to inform policy-making or mitigation solutions where significant resources are at stake. For example, an understanding of predictive uncertainty plays an essential role in the political acceptance of the need to design

policies to address global warming when the cost of different policies varies by trillions of dollars. Policy and decision makers need analyses of complex systems that are supported by quantitative characterizations of error and uncertainty.

In terms of computational costs, quantification of uncertainty and error estimates and control are tremendously expensive undertakings that raise entirely new sets of challenges for both mathematical algorithm development and high performance computing. The underpinnings are the problems of forward and inverse sensitivity analysis. Forward sensitivity analysis is concerned with how errors in data, parameters, and discretization propagate through a model to affect output. Inverse sensitivity analysis reverses the point of view to determine the allowable uncertainty in inputs to a model given a desired degree of uncertainty in the model output. This is an ill-posed inverse problem that provides a powerful link between model results and experimental observation. Both types of problems involve determining how model output changes with changes in input and discretization. Whereas in the past one-time solutions of simple models might have sufficed for scientific investigation, both forward and inverse sensitivity analysis involves the simulation and analysis of model behavior for *many* sets of data/parameter values and discretizations. What are needed are entirely new classes of efficient and robust algorithms for sensitivity analysis and uncertainty quantification that can scale to very large numbers of parameters and expensive simulation models that can efficiently utilize the millions of processor cores that characterize future exaflop systems.

In terms of computational costs, quantification of uncertainty and error estimates and control are tremendously expensive undertakings that raise entirely new sets of challenges for both mathematical algorithm development and high performance computing.

3.8 Large-Scale Simulation-Based Optimization

Advanced CI has the potential to enable a transformation from simulation to simulation-based decision-making, which gives rise to complex optimization problems that include large-scale forward problems as constraints. Those optimization problems arise in design (in which the decision variables represent the configuration and constitution of the system) and in manufacturing and operations (in which the decision variables represent control parameters). Moreover, decision-making informed by predictive simulation requires estimation of uncertain parameters that characterize the simulation. The resulting inverse problems seek to estimate these parameters by minimizing discrepancy with observations.

Advanced cyberinfrastructure has the potential to enable a transformation from simulation to simulation-based decision-making.

Unfortunately, the solution of simulation-based optimization problems, whether in the

form of optimal design, optimal control, or inverse problems, is notoriously more challenging than the corresponding forward problem. First, the optimization problem is often ill-posed and requires careful regularization, despite the usual well-posedness of the forward problem. Second, it usually results in a 4D space-time boundary value problem, despite the evolutionary nature of the forward problem. Third, the optimization problem often includes inequality constraints, which create difficulties not encountered in the forward problem. Fourth, the optimization objective and/or constraints are often formulated in probabilistic terms. And fifth, the forward problem is merely a subproblem associated with optimization, which can be orders of magnitude more computationally challenging. Indeed, when the forward problem requires petascale resources, the optimization problem will usually be in the realm of the exascale.

Because of those difficulties, contemporary optimization methods are inadequate for the solution of frontier optimization problems that are governed by large-scale complex simulations. We need entirely new classes of efficient and robust optimization algorithms that address the difficulties listed above and can scale to the millions of processor cores that characterize future exaflop systems. The challenges in creating those algorithms are of the highest order, but they must be overcome to elevate decision-making for complex multiscale, multiphysics simulations from a practice relying on simple interpolative models to a more rigorous science based on high-fidelity predictive simulation.

3.9 Integrated Sensor-Simulation Systems

Many of the algorithms and methods discussed in this chapter must be merged together to address the challenges of creating integrated online sensor-simulation systems. In such systems, the goal is to assimilate data from sensors, often dynamically, to infer unknown parameters and states of large-scale simulations of physical systems (for example, an evolving hurricane), using methods from inverse theory. The updated simulation models are then advanced forward in time to yield predictions (such as storm path), which can then be employed for simulation-based decision-making (such as how to deploy emergency responders) or used as a basis for simulation-based optimal control. Optimal experimental design theory can then be used to determine optimal locations of sensors (for example to reduce uncertainties in the estimation of current atmospheric state); these locations are then fed back to steer the sensors to new locations. This entire cycle of sensing-assimilation-simulation-prediction-control-steering is then invoked repeatedly, often in real time, over the life cycle of the evolving event. Such systems are becoming known as *Dynamic Data-Driven Application Systems* (DDDAS) [14].

Algorithms and computational methods underlying DDDAS face enormous challenges. While such systems have been realized in practice, the models that are at the core of DDDAS simulations tend to be simple (such as lumped parameter models or ODEs), or else if they are high-fidelity models, the underlying data assimilation/control/steering algorithms tend to be simple and often heuristic. The challenge is to create online dynamic data driven application systems that employ high fidelity (multiphysics, multiscale, multi-model PDE) simulations of the evolving event in conjunction with provably optimal algorithms for the assimilation, control, and steering components. A further serious

challenge is to carry out the DDDAS framework while rigorously accounting for uncertainties—in data, in models, in predictions, and in the design and control phases. Success in developing high-fidelity DDDAS systems that operate with quantified uncertainties will lead to significant beneficial impacts on many societal problems in such areas as manufacturing, commerce, transportation, hazard prediction and management, and medicine, to name a few.

3.10 Verification, Validation, and Reproducibility

The viability of predictive simulation in CS&E is founded on the ability to carry out verification and validation (V&V) of complex CS&E models and codes. We distinguish among four different entities: the physical system of interest, the mathematical model of that physical system, the numerical approximation to the mathematical model required to render it solvable on a computer (that is, the computational model), and finally the software implementation of the numerical approximation.

Verification is the process of determining if a computational model of the physical system is an acceptable approximation of the mathematical model of the system. Verification comprises both code verification, and solution verification. Code verification is the process of confirming that the computer code implementing the computational model correctly employs the algorithms developed for the implementation. Solution verification is the process that determines that the equations and mathematical constructions governing the model are numerically solved with sufficient accuracy for specific quantities of interest (QoI) and for the specific simulation at hand. Code verification employs analytical and manufactured solutions to assess expected convergence rates. The adoption of systematic software engineering practices gives further credibility in the development of complex scientific codes.

An additional mechanism for verification lies in the emerging initiative for reproducible computational research, which advocates that all details of computations, code, and data, be made conveniently available to other researchers [59]. Often the steps taken to generate computational results are embodied in software scripts or code. The predictions of large-scale complex simulation codes involve large numbers of small decisions, from data collation and filters to parameter settings in algorithms and software invocation sequences. Those decisions are often impossible to capture completely in the final published papers, simply because of their large number. In those cases a convenient way to communicate research methodology is to release the underlying code for inspection. Release of the accompanying data is the second necessary step for reproducibility of published computational findings.

Solution verification is the province of the field of *a posteriori* error estimation. Several decades of advances in this area provide the capability of yielding rigorous and, in many cases, guaranteed bounds on errors for specific applications. New error estimation techniques have been advanced that enable the estimation and control of modeling error, error due to uncertainty, and approximation error for multiphysics and multiscale models. The challenge for CS&E is the development of rigorous *a posteriori* error estimates and adaptive control of all components of error for the sort of complex, multiphysics, multiscale models characterizing Grand Challenge problems (as exemplified by those in Section 2.1) remains a challenging problem in CS&E that must be overcome to build confidence in the

predictions of such codes.

Validation of a model, in contrast, involves comparisons with reality, that is, experimental observations. We understand that validation is a *process* designed to give confidence in a model or to reject it: no model can actually be "validated"; it can be "not-invalidated" by physical observations, and common terminology is to declare such models as valid, a subjective decision. But before we can validate, by a criterion, any uncertain model parameters and other input data must be inferred from available experimental observations via a calibration process. However, the experimental data are themselves uncertain (due to measurement errors), and there may be uncertainties associated with the mathematical form of the model (structural uncertainty). The calibration problem thus seeks to determine probability distributions of model parameters that are consistent with the probability distributions of the observed quantities in the calibration experiment, the model uncertainty, and any prior information on the parameters. Bayesian inference provides a systematic framework for accounting for all of these sources of uncertainty in the calibration process. However, for problems with high-dimensional uncertain inputs and expensive forward models, the Monte Carlo sampling techniques that underlie standard approaches used in Bayesian inference quickly become untenable.

A calibrated model can then be subjected to a validation test using additional validation experiments not used for calibration. The uncertainties of the calibrated parameters, which are now random variables, are propagated through the calibrated model to produce probability distributions of outputs; these outputs are compared with the probability distributions of measurements from validation experiments using appropriate metrics and associated rejection criteria to assess model validity. If the model is rejected, then either it must be recalibrated with more or better data, or the model form itself needs to be refined, for example, to include a neglected phenomenon or to remove a simplifying assumption. In this way, the validation process drives model development and experimental measurements. The major computational challenges here are propagating uncertainties from model parameters to code outputs. As is the case with inverse uncertainty propagation, contemporary methods for forward propagation of uncertainty break down for expensive models and high-dimensional parameter spaces. *Overcoming the curse of dimensionality* for forward and inverse uncertainty propagation is critically needed for the development of rigorous and scalable methods for validation of large-scale complex models.

Overcoming the curse of dimensionality for forward and inverse uncertainty propagation is critically needed for the development of rigorous and scalable methods for validation of large-scale complex models.

A vital component of reproducible research in computational science is openly accessible code and data. With the expanding role of data-driven discovery and computational modeling and simulation in scientific discovery, the reproducibility of results places new demands on the robustness and documentation of software.

3.11 Recommendations

The life's blood of CS&E is computational methods. Often undervalued and taken for granted, excellence in research in this area is key to international leadership in broad areas of CS&E and CI. This chapter has reviewed the challenges associated with computational algorithms and methods as they face a new generation of complex problems in CS&E that must be solved on new generations of computing systems. The main conclusions drawn are:

- An algorithmic Moore's Law has held over the past four decades, with simulation capability progressing as much from developments in more sophisticated computational algorithms as from advances in hardware capability.
- Advanced computational methods become even more critical as problem size and complexity (multiscale, multiphysics, multimodel) increase.
- Advanced computational methods that map to emerging (manycore, hybrid) architectures are generally not available and will need to be developed and supported.
- A lack of investment in computational methods will result in our inability to make effective use of new HPC systems, thus jeopardizing NSF HPC and other CI investments.

The overall recommendation from this chapter is given below.

RECOMMENDATION:

A broad-based, comprehensive, long-term, and vigorous research program in advanced computational methods should be established to overcome the challenges faced in devising, analyzing, and scaling up new computational methods for a new generation of critical CS&E problems on advanced computing systems. These should include advances in discretization methods, solvers, optimization, and validation and uncertainty quantification methods, including the facilitation of reproducible research through affirmative steps such as the creation of repositories for code and data, all targeted at enabling new frontiers in large-scale multiphysics simulations on emerging architectures. This program should support multidisciplinary and interdisciplinary teams that bring together applied mathematicians, computer scientists, and computational scientists and engineers.

4 High Performance Computing for Grand Challenge Problems

4.1 Challenges of Exascale Computing

Transformative discovery and innovation in most disciplines important to meeting the Grand Challenges, such as climate, energy, environment, national security, disaster preparedness, and medicine, depend on the pervasive and seamless availability of computing at scale. According to many projections, general purpose exascale computing equipment is likely to be available in the next 10-15 years [35]. However, this will likely be made possible only by dramatic changes in processor architectures, including very large scale of multi-core processing (perhaps in the range of 1000 cores per chip or beyond), power management, and packaging. New methodologies for power management at circuit, device, and system level, locality and concurrency of data and the computations that use/generate it, and resilience to system faults, are going to be crucial to development of these systems. Adoption and efficient use is thus likely to require many advances in programming models, tools, and techniques, training, and workforce development, and will also require significant investment in upgrading the applications and software stack.

As CS&E moves more towards adopting rigorous standards for validation, verification, documentation, and reproducibility, routine access to HPC is crucial.

Concurrent to this revolution is another paradigm shift in the quantity, quality, and availability of digital data and its use in driving modeling and simulation. Computing is increasingly data driven and conversely HPC is often constrained by the high volume of data it generates and consumes. HPC needs to respond to these needs with architectures that are more flexible in terms of the balance between data handling capabilities and processing power. While we keep our focus on catering to the evolving science needs, it is also clear that advancement in the grand challenge science areas is greatly dependent on low barrier and adequate access to the

Computing is increasingly data driven and conversely HPC is often constrained by the high volume of data it generates and consumes.

existing models of computing at scale. Ubiquitous and seamless availability of high end computing from the scientists' workbench are necessary to deliver on the promises of computational science. As CS&E moves more towards adopting rigorous standards for validation, verification, documentation, and reproducibility, routine access to HPC is crucial.

Resource provision modalities like those advocated by "clouds" and grids will need to be integrated with HPC provisioning.[4] Such provisioning, which can demystify the usage of HPC when coupled with appropriate CI for collaboration, can transform scientific investigation for entire disciplines, leading to an age of unmatched innovation and discovery. However, much needs to be done to ensure that these cloud environments that are currently driven by the needs of commercial processors are also addressing the needs of the NSF community. Recent partnerships between NSF and commercial vendors on making these environments available to NSF funded scientists through a peer-reviewed competition are commendable [54]. A balanced provision of extreme and moderate scale computing is clearly needed as advocated in the three tier structure of Track I, II and III in the cyberinfrastructure vision [13]. The interpretation of the Tracks will need to be done in the context of available hardware options. The intermediate and lower end tracks must clearly include these mechanisms for modest scale HPC provisioning. NSF investments in these areas are crucial to ensure that these promising technologies are developed in manners consistent with the needs of NSF users.

[4] The present NSF sponsored investment in the TeraGrid gateway is an early and successful example of such provisioning of HPC for well-defined communities.

This is also an area where there exist significant opportunities for cooperation nationally (across multiple federal agencies NSF, DOE/OASCR, NASA, DOD, etc.) and internationally (*e.g.,* with the PRACE Project) for synergistic investments. Progress in use and development of these resources will require much joint investment – see for example [25] for details of such an initiative.

4.2 Core HPC Advances Needed for GC Communities

A quantum increase in machine, software, and human resources must continue to be provisioned for widespread use among NSF researchers. The great quality and quantity of science enabled by existing resources (see, for example, http://www.teragrid.org) clearly shows that these resources are now an essential part of the research methodology for the NSF research community. HPC architectures, methodologies, and software to exploit these architectures are in a constant state of flux. Grand challenge applications need to take the fullest advantage of those systems at the earliest possible opportunity. Thus, it is necessary for the NSF research community to *have early and low barrier access to the best equipment, methodologies, and software*.

Apparent from the grand challenges listed earlier is the diversity of the applications addressed by NSF researchers. That diversity requires support for very different computational methodologies and computer architectures in the NSF HPC arsenal. The constant evolution of these architectures and development of new methodologies dictate *a consistent policy of forward looking investment in innovative hardware*. The equipment available in the NSF portfolio of HPC resources should not only be able to meet the existing needs of the user community but also act as a driver for the

development of effective strategies and tools for best use of forthcoming machines by the considerable NSF computational science research community.

A list of developments and special capabilities needed for exascale platforms for HPC systems in the future (O(10) years) was recently proposed by a group of international experts as part of a roadmap for exascale software [25]. Significant among these are a need to reduce power costs to 25 pico-joules per floating point operation, 10 billion-way concurrency for simultaneous operation and latency hiding, O(100) petabytes capacity mix of DRAM and nonvolatile memory, and bandwidths of the order of 1 terabit enabled by optical technologies. Each one of these features is going to be necessary to attain the target of a usable exascale computer that satisfies the "politico-economic" constraint of 25MW for the maximum power consumed. Ensuring that the software and hardware developments targeted here do actually happen is challenging and will need the detailed and balanced program of investment we outline below.

4.3 Software Stack - Programming Models, Compilers, Debuggers, and Development Environments for Extreme Scales

The Message Passing Interface (MPI) based programming model, which requires the inherently flat architecture of most of the high end computers used today, will need to be reinvented to meet the application challenges outlined in the previous section. New programming models that allow researchers to exploit the heterogeneity of processing elements and the hierarchy of memory and data storage need to be developed. Current attempts at using, for example, Partitioned Global Address Space (PGAS) languages need to be investigated, but there also needs to be fresh thinking on alternate approaches that enable Grand Challenge applications to take advantage of locality of data and memory usage where they exist. New generations of architectures will need innovative combinations of different types of processors and accelerators for processing, memory and storage structures, and application-driven asynchronous usage of machine elements. These architectures will need new programming modalities and tools.

Computer architecture development is guided by the power budget as much as by the computing needs of the Grand Challenge applications. That constraint will be crucial to the move to the next stage of development. Recent analysis [16] indicates that, to deliver the necessary computing at power budgets that are realistic, significant reductions will have to be made in the power consumed at all scales of computing equipment. Usage of graphics processing units and accelerators to provide customized computing capability has made multiteraflop and petaflop scale computing affordable. However, this approach is unlikely to be adequate for the next generation. A more holistic approach, one integrating application-aware approaches to power management, is likely to be needed.

4.4 New Numerical Algorithms to Efficiently Use Petascale and Exascale Architectures

The scaling to a million chips (with a thousand cores per chip) and the use of special-purpose accelerators, graphics cards, and data appliances will require the development of new algorithms and methodologies to deal with the fundamental shift in the computing and data storage and

access architectures. The parallel algorithms currently used in most applications presume a relatively flat processor layout, static partitioning, and scheduling based on *a priori* knowledge of the architecture and problem. The next generation of codes will need to exploit the fine-grain parallelism at the intranode/socket level (O(1000) cores with shared memory and coherent caches) and the coarse-grain inter-node parallelism (O(10000) nodes with relatively low bandwidth) simultaneously and respond to changes in machine and code needs at runtime. Algorithms that contain even limited fractions with O(P) schemes (P is the number of nodes) will have insurmountable bottlenecks in effectively scaling to future architectures. Locality of data and avoidance of nonlocal memory references will be needed. New algorithms must be created that enable such methodologies.

4.5 Data Flow and Data Analysis at Extreme Scale

Many of the next generation Grand Challenge applications will be "data intensive." The ratio of data movement to computing required by such applications is quite different from that for compute-intensive applications. Very large volumes of data need to be moved among processing elements and from secondary and tertiary storage. To enable such applications, great attention has to be paid to the data flow among the interacting components of the applications and the computing devices. Data required for a particular computation may reside immediately next to the processing element and available instantly, or the data may be as far away as a central repository a continent away. The systematic development of input and output systems to enable such data access dynamically will be a great challenge and will require much investment. Novel I/O architectures taking advantage of either localized preprocessing and/or solid state disks are also being developed. These architectures will fundamentally change the programming model that is appropriate for most applications. These core needs naturally lend themselves to a preliminary set of recommendations. Those recommendations must be coordinated with the companion groups working on a plan to address HPC needs of the NSF community before they are submitted to the community of stakeholders.

Coordinated investment in developing these critical technologies is needed. Ad hoc and diffuse funding models are unlikely to succeed. A piecemeal and unfocused strategy governing investment in these technologies in conjunction with other broad research goals is unlikely to succeed. Thus, success also requires that a single organization take the lead role in coordinating this investment. The recommendations summarized in the Executive Summary are to be interpreted in the context of the recommendation towards evolution of the Office of Cyberinfrastructure into an agent for the support of CS&E at large.

4.6 Recommendations

NSF has taken on the challenge of providing and maintaining the computational infrastructure for advanced computing for two decades. Providing the new infrastructure needed to meet the Grand Challenges in the future will be an especially daunting objective.

RECOMMENDATION:

It is recommended that NSF, through OCI, continue to give high priority to funding a sustained and diverse set of HPC and innovative equipment resources to support the wide range of needs within the research community. These needs include support for the development of technologies to meet the foremost challenges in HPC, such as power-aware and application-sensitive architectures, new numerical algorithms to efficiently use petascale and exascale architectures, and data flow and data analysis at the extreme scale.

5 Software Infrastructure for Grand Challenge Communities

5.1 Introduction

Computational modeling and simulation as well as data analysis and visualization are critical to advancement in many areas of science—from astronomy and astrophysics; to climate change prediction and hazard analysis, mitigation and response; to nanoscale science and technology and the biological sciences. Engineering innovation has also been revolutionized with high performance computing, especially replacement of expensive physical prototypes with virtual ones that lead to more optimal designs at much lower cost in less time. Although high performance computers are the enabling technology, such advances are driven by scientific and engineering applications – software – that capture the physics, chemistry, and biology in the description of the natural or engineered system. *The software infrastructure for science and engineering is a critical part of the national cyberinfrastructure*.

The software infrastructure for science and engineering is a critical part of the national cyberinfrastructure.

The software used in the solution of science and engineering problems, including ones that are Grand Challenges, is a complex hierarchy of software components, often referred to as the software stack. As discussed in the Task Force Report on Software, this software hierarchy includes:

- Computing systems software for operating and managing computer systems, such as operating systems and file systems for individual computers and middleware for distributed computing systems;
- Tools for developing computational science and engineering applications and data analysis and visualization tools, such as compilers, debuggers, and numerical libraries; and
- Science and engineering applications, including the tools needed to analyze and visualize the data produced by these applications.

Advances in computing systems software are needed to operate and manage the increasingly complex computing systems

Advances in software are required in all areas: computing systems software, middleware, and science and engineering applications to solve Grand Challenge problems.

required to solve Grand Challenge problems; middleware is needed to provide the required functionality in distributed computing environments for addressing a growing class of data-intensive Grand Challenges; advanced tools are needed to facilitate the development of sophisticated applications and analysis tools; and, finally, a new generation of science and engineering applications is needed to take full advantage of the extraordinary capabilities provided by the advanced computing technologies needed to solve Grand Challenge problems. In this chapter we will focus on the science and engineering applications needed to address the types of Grand Challenges outlined in Chapter 2.

5.2 Key Issues in Software Development

As the scientific and engineering communities solve ever more complex problems in an environment of ever advancing computing technologies, a number of key issues must be taken into account, including:

Evolution of Computing Technologies. Computing technologies continue to evolve with major changes arising nearly every decade. In the past two decades, we have seen the decline of vector computers and the emergence of parallel computers powered by microprocessors with performance approaching that of the earlier generation of supercomputers. These single core microprocessors are now being replaced by multi-core and many-core microprocessors. Single processor performance increased by a factor of 1000 from the mid 1970s to the mid 2000s, but in 2004 performance leveled off as thermal effects prevented further increases in processor frequency. Now, increases in cores per chip provide the engine for advances in performance and software must adapt to exploit this new level of parallelism. Many other factors are also in flux, for example, the performance of memory subsystems and the speed of the processor interconnect relative to that of the microprocessor; these changes also have significant implications for the design of scientific software.

Evolution of Scientific Software. Major science and engineering applications have lifetimes measured in decades. However, they are constantly changing as the science questions advance and understanding of phenomena improves while, in parallel, the underlying computational algorithms and numerical methods continue to improve.

Major science and engineering applications have lifetimes that are often measured in decades and, thus, must evolve as the underlying computing technologies change.

Science and engineering applications, which may contain tens of thousands to millions of lines of code, must also adapt as the computing technology changes—from minor revisions as the computer hardware and systems software evolve to major revisions as disruptive changes in computing hardware and algorithms occur. As an example, consider GAUSSIAN, which was developed by John Pople, who won the Nobel Prize in Chemistry in 1998. Pople began the development of GAUSSIAN, which is still the most widely used computational chemistry application today, in the late 1960s [23] and it has continued to evolve over the past forty years [42]. However, the performance of GAUSSIAN has now

reached a plateau and will likely not increase further unless it is rewritten to exploit multi- and many-core processors. This is a major undertaking for such a complex molecular modeling application.

Reproducibility of Computational Results. Another example of a disruptive change in the evolution of scientific software comes with the increasing use of workflows to coordinate the numerous steps in a research task and their impact on the reproducibility of results. Use of computational tools is appearing in an increasing number of scientific research settings in which the tools are used in complex, highly differentiated, and granular workflows. As a result, reproducibility is poised to become a key issue in computational science and engineering and in data management. Tools for provenance tracking are emerging but need to be developed at a faster rate and for a much wider number of research problems [57]. Version control systems for code development exist but are not routinely used by all computational scientists. Version control and provenance are not just important for software but for data as well. With the diverse background of the many researchers using computation, provenance tools must be easy to use and be applicable to new problems.

5.3 Multiple Activities of Software Development

In the past, development of applications for science and engineering was driven by a patchwork of individually funded projects with little attempt to coordinate and integrate the best approaches into a single package. Some research communities addressed this problem by developing applications that integrated the advances being made into a "community code," although there was usually no direct support for this effort (GAUSSIAN is an excellent example of this practice). With the increasing complexity of computing systems, it is no longer possible to take full advantage of the advances in computing technology using this ad hoc approach. Now is the time to re-examine how best to support the development, maintenance, and upgrading of the software needed by the nation's researchers to advance science and engineering, However, it must be recognized that the development of computational science and engineering software requires the support of multiple activities:

Now is the time to re-examine how best to support the development, maintenance, and upgrading of the software needed by the nation's researchers to advance science and engineering.

1. *Development of software to test new concepts, mathematical models, and/or algorithms.* These activities are usually targeted at the development of software to demonstrate a concept, model or algorithm and are undertaken by individual scientists or small groups.

2. *Development of community codes.* These applications are intended to support the research of a well-defined community of users, many of whom may participate in the development of the software. This software may employ the new concepts and algorithms developed either by small groups, those developed by the community of researchers involved, or by the teams of researchers tacking Grand Challenge problems (see below).

Facilitating collaborative code development has the corollary effect of supporting reproducibility of published computational results by making the code widely available, independently tested, and useful well beyond the originating research group.

3. *Development of software targeted for use on the nation's most powerful computing systems.*[5] These are often large collaborative efforts that include distributed teams of computer scientists and applied mathematicians as well as disciplinary scientists and/or engineers. The development of this software often poses unique challenges since it is targeted at computers at the leading edge of computing technologies of a scale and complexity previously nonexistent.

The current funding mechanisms at NSF work well for activity 1, where principle investigators are funded for a limited number of years to develop an innovative new concept, model, or algorithm. However, they are far from optimum for activities 2 and 3, where long-term funding is needed to support the production of reliable, robust software for a broad audience of scientists and engineers.

Long-term funding is needed to support the production of reliable, robust software for a broad audience of scientists and engineers.

[5] Such as the Track 1 and 2 systems funded by the NSF and the leadership computing facilities funded by the U.S. Department of Energy.

The technical issues encountered in developing software for advanced computing systems rival those encountered in other research activities and require an understanding of a broad range of disciplines, from the scientific problems that are being targeted to the underlying computing technology and algorithms. The computer science community has, at times, provided high-level languages and programming aids to ease the programming effort. Yet, science and engineering software often lags behind, as illustrated by the GAUSSIAN example, due to lack of sufficient funding to rewrite or revise the software, uncertainties in support for some of the basic software infrastructure provided by computer scientists, and the detailed knowledge and expertise required to implement the most sophisticated applications on modern highly parallel computers.

In the current funding environment, many software advances are not fully exploited and may even be lost because there is no mechanism for the identification of the most valuable software and its long-term maintenance and evolution. Although, as noted above, some communities have dealt with this issue by establishing loose coalitions devoted to the development and maintenance of selected software packages, the sustainability of these efforts is questionable. In addition, because of the lack of direct funding, the level of effort is often

NSF must recognize the critical role of software in the nation's cyberinfrastructure and ensure that widely used software is maintained and continues to evolve.

minimal, which can lead to suboptimal software. *NSF must recognize the critical role of software in the nation's cyberinfrastructure and ensure that widely used software is maintained and continues to evolve in response to community needs*. Professional staff will play a key role in achieving this goal. These individuals have a unique combination of knowledge and expertise—disciplinary science or engineering, computational science or engineering, software engineering, and high performance computing—that is essential for creating, maintaining, and evolving these usually complex software packages. Funding agencies should explicitly include support for professional staff in their grants and help create a satisfactory career path for them.

A major development within the past decade has been the use of an open source license as a means of making software freely available to the computer and computational science community. Although Linux is the best-known example of open source software [43], many science and engineering software developers have long made their software freely available to the community. This practice allows researchers to quickly build on the innovations of others to advance their work. Because of the clear benefits of this approach, federal agencies have begun to require that software developed under their support be made available via an open source license. This requirement is critical for the development of a computing software stack for science and engineering, where the software developed by several groups, who may be funded by more than one federal agency, must work together. In fact, discussions are currently under way on the development of the software stack for exascale computers that will likely require the integration of software developed by researchers funded by agencies in several countries [44]. Also, recent work on code and data release accompanying published results indicate the use of the "Reproducible Research Standard," or similar open licensing structure [59].

Making software developed under federal support available via an open source license is critical for the development of a computing software stack for science and engineering.

5.4 Exemplary Programs and Projects in Software Development

Although the importance of software for advancing science and engineering has long been recognized at federal agencies that support research, few explicitly support software development as an end in itself. As two examples to the contrary, the National Institutes of Health (NIH) has supported the development and continuing evolution of NAMD, a molecular dynamics application targeted at computational modeling of large biomolecular systems [45], since the late 1980s through its National Center for Research Resources program [46] and, in the 1990s, as part of the Environmental Molecular Sciences Laboratory Project, the U.S. Department of Energy supported the development and continuing evolution of NWChem, one of the few molecular science applications that scales to 100,000 cores [2]. These applications are used by tens of thousands of researchers worldwide and are exemplary of major software development projects that have had an extraordinary

impact.[6] However, even in these cases, the full cost of the software development effort are not being fully supported.

More recently, programs have arisen that target the development of software for science and engineering. Principal among these are the Biomedical Information Science and Technology Initiative (BISTI) [47] at the National Institutes of Health, which has focused on the development of scientific applications important to biomedical research, and the Scientific Discovery through Advanced Computing (SciDAC) program [48] at the U.S. Department of Energy, which is focusing on the development of the software stack needed for scientific and engineering research relevant to DOE's mission. Both of these initiatives have seen significant investment and are serving their target community's needs, although, again, few would argue that the level of support is adequate to meet the challenges posed by petascale computing (and beyond) technologies needed by the Grand Challenges.

At NSF, funding for the National Center for Atmospheric Research has long included support for the development of software for the atmospheric sciences research community. In particular, NSF supported development of the NCAR Graphics package [49], the netCDF library and interface [50], and the Community Climate System Model [51] and the Weather Research Forecasting [52] applications. All of this software has found application outside NCAR and the community that it directly supports. The NSF's Supercomputing Centers Program also funded the development of a substantial suite of software that facilitated the use of the supercomputers installed in those facilities. More recently, these supercomputing center's efforts have been replaced by several independent programs, *e.g.,* SDCI (Software Development for Cyberinfrastructure) and STCI (Strategic Technologies for Cyberinfrastructure), which are funding a number of important efforts; *e.g.,* the Alpaca project [53], which is developing high-level tools to help scientists develop and maintain large, complex applications. It has yet to be seen if the SDCI program will be able to create and maintain the integrated software stack needed by the nation's science and engineering research community.

It is yet to be seen if the SDCI program will be able to create and maintain the integrated software stack needed by the nation's science and engineering research community.

5.5 Recommendations

Software to support the Grand Challenges in science and engineering are a critical part of the national cyberinfrastructure. Software investments often have a broad impact because, more often than not, the software created in one project is widely disseminated and incorporated in other research projects, enabling them to achieve their goals. Computational science and engineering will advance most rapidly if the National Science Foundation develops a cyberinfrastructure

[6] It should be noted that some of the technology developed in the NWChem project, *e.g.,* Global Arrays, has been used in many other molecular science applications, *e.g.,* GAMESS-US and MolPro, and Charm++, which lies at the heart of NAMD, is now being used in a number of other scientific applications.

program that carefully balances its investments in computer hardware and computer software.

It is recommended that NSF establish a program to support the development of scientific software for Grand Challenge communities that complements the single investigator programs that are currently so successful. It must not only support the creation of new concepts, models, and algorithms, it must also support the creation, maintenance, and evolution of major science and engineering applications. While it is clearly premature to prescribe new mechanisms for supporting the development of the software needed for computational science and engineering research, a few core ideas are beginning to emerge.

- Groups that can integrate expertise in computing technology, software engineering, and computational methodology with science and engineering and are embedded in application domains, can provide a much needed locus for the sustained development and maintenance of essential science and engineering software.
- Professional staff, individuals who have expertise in science or engineering, computer science and/or engineering, software engineering, and high performance computing, are critical to creating science and engineering applications that can evolve to meet the needs of the communities they serve as well as evolve as computing technology changes.

With the arrival of petascale computers and the expected progression toward multi-petascale and exascale computers in the next decade as well as the rapidly growing capabilities in data-driven discovery, opportunities for advancing science and engineering have never been higher. Also, with the expanding role of data-driven discovery and computational modeling and simulation in decision support as well as scientific discovery, the reproducibility of results places new demands on the robustness and documentation of software. As a result, the demands on innovative and sustainable software have never been higher. These considerations lead to the following recommendations.

RECOMMENDATIONS:

It is recommended that NSF:

1) Support the creation of reliable, robust science and engineering applications and data analysis and visualization applications for Grand Challenges as well as the software development environment needed to create these applications.

2) Provide support for the professional staff needed to create, maintain, evolve and disseminate the above applications as part of its grant funding.

3) Establish best practices for the release of science and engineering applications and data as well as the workflows involved in their creation to ensure the reproducibility of computational results.

6 Data and Visualization

6.1 The Data Challenge

Digital technologies have transformed every facet of research, from the questions asked and the methods used to the ways in which researchers interact. Since 2003, digital information makes up 90 percent of all information production, vastly exceeding the amount of information on paper and film. As simulations and experiments generate many petabytes and even exabytes of data, science is becoming increasingly data intensive. For example, climate models are expected to generate hundreds of exabytes by 2020 [9], and the Large Hadron Collider (LHC) will produce roughly 15 petabytes of data annually over its estimated 15-year lifespan [86]. Thus, one of the greatest scientific and engineering challenges of the 21st century is the endeavor to understand and make effective use of this growing body of information. Scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate, explore, and model massive datasets. Indeed, a deluge of data has shaped a new era in computing, a shift called by Jim Gray the "fourth paradigm" of science, which will focus on the power of data-intensive computing [27], following the first three paradigms of science—experiments, theoretical hypothesis, and computer simulation. In this paradigm, science follows a data-first approach in which massive amounts of data are collected by automated instruments and then processed via visualization, data mining, and statistical modeling to discover regularities and generate and test hypotheses.

Since 2003, digital information makes up 90 percent of all information production. One of the greatest scientific and engineering challenges of the twenty-first century is to understand and make effective use of the growing body of information.

Several reports and books have been published in recent years that discuss the challenges created by the "tsunami" of scientific data and the potential transformative opportunities for science and society [*e.g.,* 22, 61]. Many studies have begun to address the major issues in the management, policy, research challenges, and use of digital data. These issues include the integrity, accessibility, and stewardship of data [15], the long-term preservation of digital data [31], and the economical sustainability of data [60], as well as the challenges in scientific visualization [37, 28, 65], modeling and simulation [58, 35], data analysis [32], and software development [17, 25]. To address those issues [5], major committees have been established, including the Interagency Working Group on Digital

Data (IWGDD) and the National Research Council Board of Research Data and Information (BRDI).

This chapter focuses on data and visualization issues relevant to Grand Challenge Communities and provides scientific case studies of each. Those aspects are (i) value proposition of digital scientific data and visualization, (ii) data science and data infrastructure as a major component of research in cyberinfrastructure, and (iii) the scientific and user communities of data-intensive science.

6.2 Broad Impact of Digital Scientific Data

The continuous cycle of generation, access, and use of an ever-increasing range and volume of digital data is transforming all elements of science. To harness the accelerating data explosion, our most important tools now include data management and visualization. Indeed, the re-use and re-purposing of digital scientific data and visualization capabilities will have a dramatic impact on scientific, biomedical, and engineering research; defense and national security; and industrial innovations.

Success Stories. Access to high performance computing resources to collate, interpret, model, and visualize scientific data in real-time has led to significant advances in our understanding of weather, fundamental physics, chemistry and structural biology, and earthquakes among other fields of scientific discovery. A relevant and representative success story for data and visualization is the real-time prediction of tornados as exemplified by the NSF Center for Analysis and Prediction of Storms (http://www.caps.ou.edu/) where weather data is sampled in real-time and sent over dedicated links to high performance computing resources for modeling and simulation followed by visualization to map out high probability regions and tracks for tornados (http://www.psc.edu/science/2007/storms.html/).

Several examples of projects under way illustrate the demands on data management, and challenges and opportunities of data-driven science:

1) *Global Earth Observation System of Systems (GEOSS)*. Earth observations are the data collected about the earth's land, atmosphere, oceans, biosphere, and near-space environment. These data are collected by means of instruments that sense or measure the physical, chemical, and/or biological properties of the earth. These data provide critical information to assess climate change and its impacts; ensure healthy air quality; manage ocean, water, mineral and other natural resources; monitor land cover and land use change; measure agricultural productivity and trends; and reduce disaster losses. The Strategic Plan for the U.S. Integrated Earth Observation System directly supports the efforts of more than 70 countries, who are working together to achieve a GEOSS – which will interconnect a diverse and growing array of instruments and systems for monitoring and forecasting changes in the global environment (http://usgeo.gov/docs/EOCStrategic_Plan.pdf).

2) *Reverse Engineering the Brain.* The brain is the most complex biological system we know, and understanding its functionality is the compelling biological challenge of the century. How do thought, action, and emotion arise from the building blocks of life? The National Academy of Engineering has selected *reverse engineering the brain* as one of its grand challenges. The cerebral cortex of the human brain contains more than 160 trillion synaptic connections that originate from billions of neurons. Given the

complexity of the nervous system, it is not surprising that the neurosciences are rich in the use of and need for data. The neurosciences now rely heavily on *in vivo* imaging methods and computational models, both of which depend on computing power and mathematical techniques. Large-scale, high-resolution images of small sections of the brain are already measured in tens of petabytes and increasing. In addition, neuroscientists must work across multiple scales of resolution and must integrate such diverse data sets as cellular neuroimaging, gene expression data, genotype data, neuronal morphology, and clinical data. With new technologies, there are hopes to ultimately create a "connectome"—a complete circuit diagram of the brain. This goal will require intensive and large-scale collaborations among biologists, engineers, and computer scientists.

3) *Integrated Public Use Microdata Series (IPUMS).* The study of powerful large-scale trends such as economic development, urbanization, expanding migration, population aging, and mass education by social, behavioral, and other scientists requires access to global-scale micro-data – data about individuals, households, and families collected by census offices around the world. IPUMS provides researchers and educators with interoperable access to data from more than 130 censuses in 35 countries, representing more than 279 million person records. This powerful digital collection meets critical research needs while successfully preserving appropriate privacy and confidentiality rights. In addition, IPUMS allows researchers to construct frameworks for analyzing and visualizing the world's population in time and space. This broader view allows researchers to identify agents of change, to assess their implications for society and the environment, and to develop policies and plans to meet or prevent future challenges at local, regional, national, and global scales (https://international.ipums.org/international/).

4) *Next-Generation Sequencing and Genomics.* Genome sequencers are making rapid advances, with those newly available in 2010 having up to 1000 times more throughput, thus enabling results up to 1000 times less costly than the previous generation DNA sequencing tools; these advances are transforming life sciences research. It is now feasible to study metagenomics—the collection of genomes recovered directly from environmental samples – to characterize unculturable organisms and complex microbial communities in their natural environment. This also includes the study of the human microbiome—the collection of genomes of microbes in the human body, and their impact on human health and human disease. The large-scale sequencing of individual human genomes can uncover genetic variants associated with disease, leading to the potential use of the personal genome in medicine. Where high-throughput sequencing was previously limited to large genome centers, next-generation sequencing has brought the field of genomics back into the laboratories of single investigators or small academic consortia. With a throughput of one terabyte per day and increasing, it requires cost-effective, compact solutions that provide high performance computational power combined with vast storage capabilities in order to fully exploit the success of modern sequencers.

5) *Sensor Networks.* Many areas of science and engineering, such as terrestrial ecology, oceanography, and geosciences, are rapidly becoming data-driven sciences. Large, spatially-distributed and heterogeneous networks of sensors are being deployed including fixed instruments, mobile sensors, and citizen scientists. The resulting data are

voluminous, spatially distributed, and collected on different time scales and with different sampling plans and sampling biases. Furthermore, the miniaturization of sensor technology is driving a tradeoff from deploying a small number of expensive, highly-reliable sensors to deploying many thousands of cheap, unreliable sensors. This gives rise to many challenges: (i) data cleaning: quality assurance and quality control of raw data must be fully automated, because the data are too voluminous for manual inspect and cleaning; (ii) data fusion: data at multiple spatial-temporal scales and collected under different protocols must be fused to a single scale or by employing novel multi-scale modeling methodologies; (iii) data filtering/disposal/transport: typically data are collected at distributed remote locations; analyzed at a second site, usually a computing facility; visualized and interpreted at a third site such as the scientist's desktop; and finally must be assimilated into predictive models and decision-making tools such as for forest management, public health management, etc. Necessary cyberinfrastructure includes greatly improved networking, algorithms, and low power hardware for in situ data reduction/compression/sampling, multi-resolution, multi-tiered distributed data storage, algorithms for petascale data mining, and distributed algorithms for visualization and decision support.

6.3 The Need for a Data Infrastructure

Data-centric science is characterized by the massive scale and complexity of data and the interdisciplinary and multidisciplinary nature of data generation, management, analysis, and use. The heterogeneous methods and devices for data generation and capture and the inherently multi-scale, multi-physics nature of many sciences have resulted in a mass of data with hundreds or thousands of attributes or dimensions and spanning multiple spatial and temporal scales. Thus, not only are centralized storage repositories "data-intensive", but so are the far greater volumes of data that are network-accessible in offices, labs, and homes and by sensors and portable devices. Thus, data-intensive computing should be considered more than just the ability to store and move larger amounts of data. The complexity of the new datasets, as well as the increasing diversity of the data flows, is rendering the traditional compute/datacenter model inadequate for modern scientific research.

Indeed, as the International Exascale Project Roadmap [25] emphasizes, "*The potential impact of Exascale computing will be measured not just in the power it can provide for simulations but also in the capabilities it provides for managing and making sense of the data produced.*" New infrastructures are needed to enhance capabilities for finding, using, and integrating data to accelerate its use in discovery and innovations.

Deficiency of Current Data Infrastructure and Lessons Learned. There is currently a lack of robust cyberinfrastructure for data science. The recent DataNet program, for example, is of an exploratory nature and not structured to provide general-purpose data infrastructure. Similarly, the storage and archive resources of the TeraGrid are designed only to support infrastructure for TeraGrid compute systems. As the world becomes more instrumented and our processes for dealing with significant data in-flow from sensors or experiment and data out-flow from simulation models becomes limiting, data and visualization become an ever-increasing challenge. This challenge is not only the physical management of the data, but also in the software tools to

manage, migrate, cache, and efficiently analyze the data. Lessons from TeraGrid show that the standard model of running on high performance computing resources remotely followed by transferring the data back locally for analysis and visualization is breaking down. Large scale data easily generated in a few days may presently takes weeks to transfer back to home institutions, and even this assumes high speed interconnect and the availability of local resources to store the data. Researchers are spending more time on data management and simulation workflow than actually doing the science. Previously, obtaining data was rate-limiting; now the aggregation, analysis, interpretation, and visualization have become the limiting step. What is needed are ways to facilitate both remote, local, and on-the-fly data analysis and management. A step in the right direction is the InCommon federation of resources (http://www.incommonfederation.org/) that allows an authenticated user to get direct access to TeraGrid. However, this needs to be broadened to facilitate access to the data, for example to mount remote disk resources from the TeraGrid locally.

Data characteristics and infrastructure requirements. Scientific data must be thought of hierarchically or as tiered with different levels of performance, reliability, security, and accessibility. Scratch disk space, high performance parallel, global file systems, archives, and real-time data streams necessitate different requirements; specifically, different policies, economies, and expectations related to lifetime, costs, value, and reliability. In the ideal scenario, real-time data flow would come into and out of simulations running on large parallel resources with on-the-fly analysis and visualization capturing data as fast as possible as a means to understand and potentially steer the simulations. Yet, not all of the data can likely be saved, so data subsets will be either aggregated, reduced, or less frequently time sampled and subsequently saved on high performance parallel disks. Data will then be analyzed by other loosely coupled resources, which in turn may be reduced for longer term storage on more cost effective, ideally globally accessible, file systems for further analysis, and processing, ideally remotely and locally.

When data has to move between resources, middleware should facilitate migration, data integrity, caching, and also provide metrics for the expected timescale for the data availability. Gold standard data should be archived, and policies set to understand worth, cost, lifetime, need for annotation, and dissemination. Another complication is the notion of distributed data where collaborations independently work on pieces of the puzzle but later need to manage, integrate, and analyze the data.

The characteristics of scientific data further necessitate robust high-speed computer networks for several distinct reasons: (i) as instruments grow in diversity of location and richness of data, increasingly the data needed for the workflow will be large and not co-located with the processing resources; (ii) workflows are becoming more complex, and include data acquisition and processing, simulation and modeling, data analysis, and visualization by the discipline scientists; and (iii) the resources needed by the various workflow steps are not likely to all be in the same location. All these observations emphasize two key applications of computer networks: (a) moving large data sets and (b) supporting effective remote visualization (*e.g.,* by streaming visualization flows at high speed over wide area.).

Components of a Robust Data Infrastructure. Major advances in computer science and engineering will be key to

addressing the cyberinfrastructure needed to empower data-intensive science. An end-to-end approach is required that encompasses the entire data life cycle from the initial data acquisition, to data management and storage, and to data integration, analysis, visualization and knowledge discovery. An important rationale underlying the end-to-end approach is the central role reproducibility plays in our scientific efforts. Without the communication of the entire data life-cycle and data processing it is difficult – if not impossible – for fellow scientists to verify and replicate data-driven findings. While there are various domain-specific scientific applications, data-intensive science shares major common cyberinfrastructure needs [17, 25, 32, 35, 65]. A robust persistent data infrastructure will consist of several major components:

1) *Data Analysis and Visualization.* Innovative research is needed in the areas of data analysis, mining, and visualization to promote enhanced capabilities for finding, understanding, visualizing, and interacting with data, and to gain novel insights from extreme scale, complex scientific data. Visual analysis systems that enable interaction between the scientist users, the data analysis system, and the data are critical for supporting scientific discovery and for enhancing communication about science outcomes. Visual data analysis is needed for extreme scale, heterogeneous, and high-dimensional scientific data. New mathematical and statistical approaches and algorithms are needed to scale with the size of the data, along with related parallel implementations able to scale with the exascale computing. New models and tools are needed for indexing, querying, and searching massive datasets. New algorithms should make effective use of new computer architectures being developed and the associated development of scalable algorithms, libraries and tools.

New mathematical and statistical approaches and algorithms are needed to scale with the size of the data, along with related parallel implementations able to scale with the exascale computing.

2) *Data Integration and Interoperability*. To promote the effective integration and interoperability of data and data tools, systems, services, and resources will require the use and development of common standards. Also needed are ontologies for semantic data integration and analysis, support for collaborative data analysis, as well as knowledge representation and machine reasoning research to support automated analysis of large data sets and integration of data from multiple sources.

3) *Data Provenance and Stewardship*. Concepts, strategies, tools, and automated protocols should be developed for data quality assessment and control, validation, authentication, provenance, and attribution. Data should be documented adequately enough to find it, interpret it, and understand its provenance – the processes that gave rise to it. This requires a robust infrastructure for uniquely naming data sets that will persistently resolve to the underlying data. In addition, high quality metadata will be necessary to properly interpret data for subsequent visualization and analysis. This will facilitate repeatability. By encapsulating context with the data, associated metadata can be properly interpreted in light of new discoveries. Support for automatic tracking of data usage as well as attribution of origination are needed to assess a data contribution. Indeed reproducibility can be

used as a framing mechanism for the drive toward open data and shared analysis. On data disposition, deciding what data to keep or discard can be guided by the application. Best practices need to be developed for disposition decision-making, including strategies and practices for understanding the relationship between cost and benefits of archiving data.

Not only is the data volume rapidly heading towards exabytes, but there are significant scientific and engineering challenges in both simulation and data analysis that are already exceeding petaflops and rapidly approaching the exaflops.

4) *Scientific Workflow and Metatools*. Scientific workflow allows a scientist to specify end-to-end control and data flow as a series of structured activities, computation, data analysis, and knowledge discovery. Meta tools are needed to aid process navigation, hypothesis tracking, workflows, provenance tracking, advanced collaboration, and sharing, as well as to support a proper balance between batch mode and interactive data exploration. Such tools are crucial to facilitating reproducibility and the ability of computational scientists to effective store and communicate the analysis underlying their results. Efforts to develop such tools are vital to the integrity and verifiability of data-driven computational findings.

5) Exascale Computing. Increasingly, experiments and observational systems are finding that not only is the data volume rapidly heading towards exabytes, but there are significant scientific and engineering challenges in both simulation and data analysis that are already exceeding petaflops and rapidly approaching the exaflops range. Hardware architectures, programming models, and algorithms for such data- and compute-intensive scientific applications must be explored. In particular, using exascale performance to rapidly do model simulations will allow the integration of data analysis and visualization into simulations to avoid storing vast amounts of data.

6) *Active Storage and Online Analysis*. Extreme scale data sets are too large to easily move and often infeasible to analyze in their raw form. Modern storage architectures can be exploited for performing various important analysis tasks. Online analytics can potentially reduce the need to store certain types of data. The needs include active storage processing studies, software libraries to embed functions within storage, and data analysis techniques. Also needed are data reduction methods and hierarchical representations for data reduction prior to post-analysis.

7) Data Storage and Management. Data storage needs include new scalable storage devices, caching algorithms to move data in/out from dynamic storage providing high level of performance, as well as scalable file systems with improvements in parallel I/O libraries. New database system approaches are needed to scale in performance, usability, query, data modeling, and an ability to incorporate complex data types in scientific applications. Scalable data format and high-level libraries for data access need to be extended and redesigned. New storage formats that emphasize scalability and parallel I/O along with the capabilities to incorporate analytics and workflow mechanisms need to be developed.

8) High-Speed Computer Networks. As the data increasingly flow physically from instrument and archive to various computing facilities to visualizations that involve the optic nerve of the science user, high-speed wide-area networks will be essential to the success of data-intensive science. Note specifically that the needed computer networks extend to the campuses of research universities. Thus, the architectures and designs of these networks will require a coordinated national effort that will include network leaders that operate at the national backbone, regional, and campus levels.

6.4 Communities for Data-Intensive Science

Digital access can multiply the value of information through repeated use. While the ready availability of diverse data is shifting scientific approaches from the traditional, hypothesis-driven scientific method to science based on exploration, current analysis and visualization methods lag far behind our ability to create data. Multi- and interdisciplinary skills are needed to handle diverse issues such as automatic data interpretation, summary visualizations, and data integration from multiple disciplines and domains. High performance computers will be needed to analyze the massive scale and complex data on a time scale that is practical in human terms. A systematic effort is needed to train the next generation of data scientists who can work in a multi-disciplinary team of researchers in high performance computing, mathematics, statistics, domain-specific sciences, etc.

Current analysis and visualization methods lag far behind our ability to create data. Multi-and interdisciplinary skills are needed to handle diverse issues such as automatic data interpretation, summary visualizations, and data integration from multiple disciplines and domains.

The value of scientific data is realized only when the data are effectively analyzed and the results are presented to the science community, policymakers, and public in an understandable way. This means that computational scientists must have the tools to track, save, and communicate their data analysis so that others are able to reproduce the findings. There are numerous examples of data re-use and re-purposing beyond the communities that generate the data. Because scientific data are often used in different ways according to their contexts and have varying life cycle requirements, solutions should support communities of practice and leverage their capabilities while promoting data integration and interoperability. Because those communities of practice are changing the way data are used and re-used and the way science in those communities is done, the community processes present an opportunity for research in the social, behavioral, and other sciences.

The challenge is to take data sets that were collected for a variety of other purposes and synthesize them to address important scientific and policy questions. Thus cyberinfrastructure requirements include support for data discovery (finding these existing data sets), schema mapping and data transformation (to convert the data into common frames of reference), and new computational statistics, machine learning, data mining, and visualization algorithms

that can support the modeling and visualization needed for synthesis studies.

An example of data repurposing is bird migration modeling. Bird migration is poorly observed and poorly understood, because birds are generally too small to carry instruments. One promising approach to obtaining data is to re-analyze existing data collected by the network of NEXRAD Doppler radar stations operated by the National Weather Service. Fortunately, this data, in relatively unfiltered form, has been archived for the past 15 years. The BirdCast project is analyzing this data and combining it with citizen science bird checklists (http://www.ebird.org), a network of microphone arrays that capture species-diagnostic flight calls (http://www.xbat.org), and several other data sources (weather forecasts, MODIS land cover data products, etc.) in order to develop statistical models of migration. Migration modeling is a grand challenge for ecology and conservation. It will only be possible by fusing heterogeneous data from many sources, including data collected for a wide variety of other purposes.

Another example, GEOSS, is a "system of systems" that supports policymakers, resource managers, science researchers, and many other experts and decision makers. Built on existing observational systems and incorporating new systems for earth observation and modeling, this emerging public infrastructure links a diverse and growing array of instruments and systems for monitoring and forecasting changes in the global environment. GEOSS further highlights the need for coherence among data-sharing principles adopted by international science collaborations and the policy and legal frameworks in place in the national jurisdictions where researchers operate (http://earthobservations.org/geoss_dsp.shtml).

Data-intensive computing promises breakthroughs across a broad spectrum of sciences and engineering and presents significant opportunities in the areas of energy, climate, socioeconomics, biology, and medicine. It will require the close collaboration of stakeholders in all sectors to fully realize the value of scientific data for science and society.

6.5 Recommendations

New opportunities are on the horizon for the development of creative uses of digital scientific data in innovative combinations for purposes of discovery, innovation, and progress. At the same time, the increasing demand for data processing, storage, and transfer in large-scale simulations based on data-informed models, stochastic systems, and requirements for model validation and uncertainty quantification represent new challenges in data-intensive computing.

With the increasingly significant impact of data-driven science, we need to better communicate the value proposition of digital scientific data and visualization to the broad scientific community, policy makers, and the public—how science will be enabled with open access data, including novel visualization and interpretation; how discovery will be enabled by integration, transcending fields; and how new data types will motivate new applications and discoveries. To fully realize the value of research data, NSF must support research infrastructure, robust and persistent cyberinfrastructure, and training of data scientists and professionals to empower data-driven science and data-intensive computing for discovery, innovation, and solution of society's pressing problems in health, energy,

environment, and food.

RECOMMENDATIONS:

NSF, largely through and coordinated by OCI, should support research infrastructure and robust persistent cyberinfrastructure to empower data-driven science and data-intensive computing for discovery, innovation, and solution of society's pressing problems in health, energy, environment, and food.

1) Research: Funding for research on data management, network infrastructure, data analysis, and data visualization (i) to manage the pipeline from field instruments to large-scale data analysis to end-user visualization and to public and policy makers, and (ii) to support data-intensive computing.

2) Data Infrastructure: Support for robust, persistent cyberinfrastructure to support the coordinated flow, storage, and management of data from instrument to (remote and local) computing resources to archiving and visualization.

3) Education: Support for building (i) the next-generation of data scientists who can work in a multi-disciplinary team of researchers in high performance computing, mathematics, statistics, domain-specific sciences, etc., (ii) data curation professionals who can support meta-data collection, indexing, and access, collaborating with scientists who collect and consume data.

7 Education, Training, and Workforce Development in Computational Science and Engineering

7.1 The Status of Education in CS&E in the U.S.

As earlier chapters have made clear, at the heart of all Grand Challenge projects is CS&E, a discipline built on interdisciplinary collaborations and deep knowledge of computational and applied mathematics and the scientific and engineering disciplines, as well as sophistication in computational skills. However, students often lack a fundamental understanding of the mathematical basis for scientific computation. In the U.S., this deficiency of understanding begins even before high school, and it continues to affect students throughout their undergraduate and graduate careers. Rather than teach these fundamentals as we did in earlier decades, the Science, Technology, Engineering, and Math (STEM) undergraduate curricula at many colleges and universities increasingly rely on the "black box" use of commercial software packages. As a result, students fail to learn the underlying concepts of modeling, programming, and "algorithmic thinking" that are critical to using computers in a scientific context [63].

Universities are not adequately preparing today's students with the right background, skills, breadth, and depth to become tomorrow's computational scientists and engineers, able to harness powerful new supercomputers for scientific discovery and engineering innovation.

The problem worsens as computers increase in complexity. Universities are falling behind in providing students with the knowledge needed to design algorithms and write software for modern architectures, such as those comprising multicore and hybrid processors that will take us to exascale computing and beyond [21,24]. The problem is exacerbated by the fact that computational science and engineering is considered neither the responsibility of the computer science departments, nor of domain sciences or engineering. As a result, teaching the core competencies of CS&E falls between the

cracks. Many students, therefore, fail to learn what is required to apply computing to the pressing and multifaceted technological challenges we face as a nation. For example, predicting and mitigating the impacts of climate change and designing clean energy technologies will require a robust understanding of CS&E. Yet our universities are not delivering the formal education required to address those global challenges. The core competencies of CS&E – including high performance computing (HPC) – are rapidly evolving, and most universities are not keeping pace. There are over 100 graduate programs in CS&E at U.S. universities, yet few schools have the necessary expertise or curricula in "bleeding-edge" HPC to prepare students to use next-generation architectures. The gap is widening between what is currently taught at most institutions and the skills needed for 21st century R&D [20, 24, 40].

The current generation of students favors the culture of open-source software in which individuals and teams can contribute in a shared community to public-domain codes. Those students, however, receive no formal training in creating *sustainable* codes that are essential for robust and effective software engineering. The skills essential for applying CS&E in modern scientific and technological enterprises are not broadly taught in the sciences or engineering, and the lack of those skills is significantly hampering the innovative potential of U.S. industry [66]. Topics missing from the curricula include 1) uncertainty quantification, 2) verification and validation, 3) risk assessment, and 4) decision making [24].

Programming the next generation of petascale and exascale computers for discovery and innovation requires new skills and knowledge that are rare among today's computational scientists and engineers. New curricula, and new approaches to teaching and learning CS&E, are urgently needed.

Of course, there are some graduate and postdoctoral students who do receive the proper education and training in CS&E and who want to pursue careers in academia, but those students often fail to blend in with existing faculty and departmental cultures. Computer science departments may be loathe to hire *computational* scientists. Similarly, application fields typically retain only a small fraction of computational scientists on the faculty, preferring to hire experimentalists over "theorists." As a result, universities are evolving slowly in their ability to transform research and education by fully leveraging CS&E. The lack of a long-term commitment to funding of CS&E is a major contributing factor. Programs such as NSF's CDI program (and, before it, ITR and KDI) focused on or embraced elements of CS&E, but they were short-term initiatives, not permanent programs. Thus, ostensibly, NSF *demonstrates no long-term commitment to the support of CS&E researchers,* and that support is essential to attract new cyber science talent to the field of CS&E. U.S. universities follow the lead of the NSF: there will be no sustainable infrastructure for CS&E in most universities until there is one at NSF.

7.2 Global Considerations

Many countries embrace CS&E and recognize the role it will play – indeed, is already playing – in driving R&D. France, Germany, China, and Japan have made major, long-term commitments to HPC. Europe

already leads the U.S. in critical elements of CS&E. China is dedicating the equivalent of $1 billion/year to a new university program that requires research projects for graduate students to have integrated simulation and modeling components. Germany is restructuring universities to leverage university-industry partnerships. Singapore and Saudi Arabia are investing enormous sums into CS&E. In comparison, the U.S. is at risk of losing its leadership position in a field it invented. That decline in our research status would have a major impact on the nation's ability to compete and innovate in the 21st century [11, 39, 58, 66].

We believe these issues necessitate concrete action and leadership from policymakers. While there are some federal programs aimed at addressing inadequacies in CS&E training and education, we believe more resources and a coordinated approach are needed to address and overcome these pressing challenges.

7.3 Existing Programs

The challenges and needs outlined above demonstrate the need for US policymakers to rethink their approach to funding CS&E education. The NSF and other scientific agencies have programs in place supporting individual CS&E initiatives, but it is clear that additional resources and programs are required for the US to remain competitive in the CS&E field and in the global knowledge economy. The following examples illustrate a few of the existing programs, along with their potential for impacting CS&E education.

The NSF CISE directorate houses the Pathways to Revitalized Undergraduate Computing Education (CPATH) [12]. CPATH focuses on providing K-12 and undergraduate students with fundamental computing concepts and methodologies necessary for building more advanced computational skills. Of the 26 awards made in FY 2009, only four address aspects directly related to CS&E generally by enabling curricula that incorporate concepts of simulation, modeling, and/or parallel computing. While this program is an important mechanism for enhancing CS&E curricula, its effect is on a relatively small scale and reaches a limited audience. To transform CS&E to meet the challenges outlined earlier in this report, a broader and more scalable approach is needed.

The NSF Course, Curriculum, and Laboratory Improvement (CCLI) program focuses on improving the quality of STEM education for all undergraduate students. Only very few of its awards, however, even touch on aspects of CS&E curricula.

There are several excellent examples of CS&E programs at the undergraduate level abroad. A recent study [24] reported the strong impact of extending the CS&E curriculum into the undergraduate arena at leading universities in Switzerland and Germany. A positive influence was felt throughout the STEM undergraduate curriculum. The availability of computational and analytical courses attracted students from a wide range of departments.

The NSF Graduate Research Fellowships Program is an important source of support for graduate students in all non-biomedical fields of science. However, of the 1,248 awards made in 2009, only 82 were in the CISE cohort. Just 11 of those awards were related to scientific or parallel computing. Such limited and highly competitive opportunities provide little incentive for CS&E students to continue on to earn advanced degrees.

The NSF Integrative Graduate Education and Research Traineeship (IGERT) program provides multidisciplinary traineeship grants

in all areas supported by NSF. However, the number of these grants focusing on CS&E topics is extremely small.

NSF supports a limited number of computational science fellows through its Mathematics Research Training Group (RTG) program.

The Department of Energy Computational Science Graduate Fellowship program supports CS&E graduate fellows. In 2010 the program was able to add 20 new fellows bringing the total number currently being supported to about 80. Unfortunately the program is typically able to fund less than 5% of the total applicant pool and even more critically, the review process leads to an annual pool of around 60 highly qualified applicants ready to study computational science and engineering.

7.4 An Educational Call to Arms

To leverage and exploit the full potential of CS&E, new curricula must achieve the following:

1) Balance domain topics and mathematical and computational skills in a way that provides both depth and breadth;

2) Teach software engineering skills needed to write, modify, verify, and validate robust and efficient CS&E codes that will address community needs over the long term;

3) Teach underlying algorithms and their applications in a highly parallel multicore environment;

4) Teach the fundamentals of simulation and modeling over a wide range of scales and applications [20].

Another way to state the challenges before us is to ask: How do we 1) modernize the CS&E curriculum, 2) provide the needed depth and breadth in education and training reflective of the 21st century cyberinfrastructure, given typical curriculum constraints, and 3) grow and diversify the workforce?

Beyond traditional university experiences, learning opportunities must be provided to CS&E practitioners in the workforce so that they can stay current as computing architectures and paradigms continue to evolve. At the same time, opportunities must be created to support and nurture new computational scientists and engineers entering the workforce.

To meet the challenges, new approaches to education, training, and workforce development in CS&E are needed. These approaches are described below:

1) *New approaches to undergraduate and graduate CS&E education.* These approaches should include the development of new curricula, courses, and/or programs in CS&E that address the computational and analytical skills required in virtually all STEM disciplines. Courses should be carefully developed and well-tested, with the objective of making the materials available to all colleges and universities in a form that is easy to extend and modify.

Work on a few foundational undergraduate courses is urgently needed. At the same time, the issues of integrating CS&E more broadly into the undergraduate STEM curriculum are complex and require study. Attention to undergraduate CS&E is essential; most applicants to graduate school have not even heard of CS&E because of its absence in the undergraduate curriculum. Summer institutes emphasizing basic CS&E skills as well as research activities and REU/RET sites focusing on CS&E are recommended for undergraduates and for exceptional high school students.

2) *New virtual communities engaged in CS&E education.* A virtual community could develop and disseminate teaching materials, innovations, and best practices nationwide, thereby accelerating the development and modernization of the curriculum. This approach entails physical and virtual centers and schools and institutes leveraging expertise across multiple institutions, including national laboratories, and supercomputing centers.

Training in CS&E skills at all levels needs to be made available online and supported 24/7, making the training broadly accessible. This accessibility will also facilitate worker retraining for those computational scientists needing to keep up with new developments in computer architectures. Candidate topics for short courses include basic computer and programming skills, high performance computing skills such as programming for many-core and GPU, and basic data mining and data analysis skills.

3) *Institution-based traineeship grants that train graduate students and postdoctoral fellows in the multidisciplinary, team-oriented iteration between experiment, theory and computation*. This procedure is rapidly becoming a paradigm in critical STEM research areas and has long been a standard in government laboratories and industry.

This training could be done through institution-based grants large enough to develop a critical mass of collaborative students and faculty. Dual advising from multiple disciplines would tighten multidisciplinary links. Universities and colleges must work hand in hand with government laboratories and industry to create internship experiences that coordinate with and broaden thesis research. In some research areas, internships in experimental laboratories for computationally-oriented students, as well as internships in computational laboratories for experimentally-oriented students, can best develop the scientific and communication skills to excel in the 21st century research environment, including sustainable approaches to software engineering, verification, validation, and uncertainty quantification, and reproducibility.

4) *Coordination of the substantial resources of multiple agencies, government laboratories, supercomputer centers and industry*. Such coordination is essential for accelerating progress in CS&E education and removing a critical bottleneck in undergraduate, graduate, and postgraduate education. A pan-agency/lab program could match undergraduate and graduate students to industry co-op opportunities, "summer camps," and internships at supercomputer centers and elsewhere. As part of this effort, undergraduates and educators everywhere should be able to view the skills that are expected for qualified applicants. Such placement is currently performed on an *ad hoc* basis, relying on personal contacts that may or may not be available at any given institution or in any given situation. Such a program could serve as an information center for opportunities and fellowships for on-site, on-line, and virtual courses, ensuring that these resources are universally available and easy to locate.

5) *Transitional grants to foster a broader and more diverse workforce and to encourage the very best students to continue in CS&E careers.* Federal agencies must help facilitate the transition of exceptionally talented graduate and postdoctoral students in CS&E to permanent positions in academia as well as industry and government/national labs. New types of federal grants that are portable, flexible, tied

to the individual, and carry the recipient – with appropriate mentoring and checkpoints – through the equivalent of tenure, would demonstrate a long-term commitment to the CS&E discipline to universities and labs. By demonstrating the ability to generate long-term funding in CS&E to support one's research program from day one, computational scientists and engineers should pose less of a perceived "risk" to institutions trying to evolve structurally to better support 21st century cyber science research and education.

6) *Sustainable, permanent programs in CS&E that support CS&E as a discipline in its own right.* These programs are needed at all funding agencies to demonstrate a long term commitment to supporting CS&E research. This proof of commitment is essential to encourage new cyber science workers to enter the field of CS&E, and it is essential to support universities in the creation of new permanent positions and programs in CS&E. At NSF, OCI should establish a permanent program in CS&E as a core mission for the Office, in partnership with the Directorates.

The overall effect of the needed approaches to CS&E education, training and workforce development described above would be (1) to increase the size and diversity of the CS&E workforce; and (2) to modernize CS&E curricula with the knowledge and skills reflective of 21st century cyberinfrastructure. The impacts should be felt across the entire STEM undergraduate and graduate curriculum at colleges and universities nationwide. Note that each of the approaches outlined above pays particular attention to issues of scalability, accessibility, and engagement of government and industry.

7.5 Summary

A shortage at all levels of appropriately trained people in computational and analytical methodology is a major barrier to progress in most areas of science and engineering, and a serious workforce issue. A broad range of coordinated efforts could be initiated to address these problems. Such efforts should (1) increase the size and diversity of the CS&E workforce; and (2) modernize the CS&E curricula with the knowledge and skills reflective of 21st century cyberinfrastructure.

The impacts of broadening and modernizing our CS&E educational infrastructure will be felt across the entire STEM undergraduate and graduate curriculum. That infrastructure will need to be accessible to colleges and universities nationwide. In our recommendations, we have paid particular attention to issues of scalability, accessibility, and engagement of government and industry.

7.6 Recommendations

Our nation is losing its leadership position in CS&E among our principal competitors in the industrialized world. Much of the traditional compartmentalization of knowledge, both within our major universities, and to an extent within NSF itself, is not well suited for interdisciplinary research vital to CS&E. It is important that actions be taken by NSF to address those issues.

RECOMMENDATIONS:

NSF should support education, training, and workforce development through the following grants and new programs:

1) Educational excellence grants at the undergraduate and graduate levels, which include funding for the development of new, courses, curricula, and academic programs in CS&E that address the computational and analytical skills required in virtually all STEM disciplines. (i) Courses should be carefully developed and well-tested, with the objective of making the materials available to all colleges and universities in a form that is easy to extend and modify. (ii) Work on a few key foundational undergraduate courses is urgently needed. At the same time, the issues of integrating CS&E more broadly into the undergraduate STEM curriculum are complex and require study. (iii) Summer institutes emphasizing basic CS&E skills, as well as research activities, and REU/RET sites focusing on CS&E are recommended for undergraduates and for exceptionally talented high school students.

2) Support for the formation of virtual communities engaged in CS&E education, including virtual entities leveraging expertise across colleges, universities, national and government laboratories, and supercomputing centers. In particular, training, in the form of short courses in core skills at all levels should be available online and supported 24/7, making the training broadly accessible. Candidates for short courses should include (i) basic computer and programming skills; (ii) HPC skills: programming and multicore, many-core, GPU; (iii) basic data mining and data analysis skills.

3) Institution-based traineeship grants that train graduate students and postdoctoral fellows in the multidisciplinary, team-oriented iteration among experiment, theory, and computation that is rapidly becoming a paradigm in critical STEM research areas and that has long been a standard in government laboratories and industry. The grants should be large enough to develop a critical mass of collaborative students and faculty.

4) The creation of a pan-agency facility or program to coordinate training in CS&E education, including training for young scientists and graduate students in communicating their work to an audience of non-specialists, and which provides a service to match students to industry co-op opportunities and to summer institutes and internships at supercomputer centers and elsewhere, and which serves as an information clearing house for opportunities and fellowships for on-site, on-line, and virtual courses.

5) Grants that facilitate the transition of exceptionally talented graduate and postdoctoral students in computational science and engineering to permanent positions in academia as well as industry and government/national labs.

6) Sustainable, permanent programs in CS&E research and education at all funding agencies to demonstrate a long-term commitment to supporting CS&E as a discipline, thereby creating reliable partners for universities seeking institutional transformational change and for trained workers seeking careers in CS&E.

8 Grand Challenge Communities and Virtual Organizations

8.1 The Role of Virtual Organizations in Grand Challenge Communities

As noted in the Introduction, collaboration has long been an essential aspect of research. Grand Challenge Communities face special challenges and opportunities with respect to collaboration. A separate report to the Advisory Committee for Cyberinfrastructure from Group 2 of the Grand Challenges Task Force will address this history and argue for a more assertive NSF response to the broad challenge of improving the means for collaboration. Likewise, previous chapters of this report have explained the revolutionary effects on research from cyberinfrastructure in high performance computing, improved software, advanced models and algorithms, effective management of scientific data, new capabilities in visualization, and specialized education and training to enable effective development and use of cyberinfrastructure. This chapter addresses the role of cyberinfrastructure as a complementary and enabling asset in the coordination and execution of research activity, especially as captured in the idea of virtual organizations.

It might seem curious that a discussion of virtual organizations appears in a report coming from the NSF Office of Cyberinfrastructure, but technological innovation often gives rise to organizational and social innovation. An office charged with innovations in cyberinfrastructure should also support and engage the accompanying virtual organizations they foster. OCI's investment in virtual organizations is a down-payment on a larger set of investments that need to be made across NSF and in other federal funding agencies as new ways of doing research evolve. Cyberinfrastructure has changed the way we think about how research can be conducted. Computer power, data storage, and other elements of cyberinfrastructure have improved dramatically in a short period of time. Unfortunately, the social and institutional environments of universities, departments, laboratories, funding agencies, and so forth often evolve more slowly. New

Virtual organizations connect people across disciplinary, institutional, and geographic boundaries. Cyberinfrastructure can enable virtual organizations, potentially revolutionizing science and engineering work.

ways of organizing the assets brought to bear on Grand Challenges are necessary to optimize the *community* part of Grand Challenge Communities. At the moment, slow progress on these coordination and execution issues is a rate limiter to research progress.

Virtual organizations connect people across disciplinary, institutional, and geographic boundaries. Cyberinfrastructure can facilitate such connections through communications (*e.g.,* teleconferencing, email), shared resources (*e.g.,* data repositories), and tools (*e.g.,* workflow coordination systems). Cyberinfrastructure can also mediate collaborations by linking observational instruments, data streams, experimental tools, and simulation systems with individuals, who might be alone or in groups, but who as a community are distributed across the globe. Cyberinfrastructure can enable virtual organizations, potentially revolutionizing the way science and engineering are practiced. This is not primarily a technological revolution, although technology makes it possible. It is, rather, a *sociotechnical* revolution in that it involves representing interlinked social and technical changes in research across all fields of science and engineering. OCI cannot address all aspects of this revolution, but its position in the Office of the Director allows it to work across directorates to address some aspects. Certainly OCI should continue its leadership role in this area.

8.2 Examples of Virtual Organizations in Grand Challenge Communities

Four examples from the recent history of science and engineering research provide an indication of how important cyberinfrastructure can be in facilitating collaboration in different kinds of grand challenges.

8.2.1 ATLAS at CERN

High-energy physics has a long history of collaboration in creating and accessing unusual and expensive equipment such as particle accelerators. It has adopted modern cyberinfrastructure and evolved sophisticated coordination mechanisms that allow a distributed community of scientists to collaborate across long distances and over significant periods of time. The ATLAS (A Torroidal Lhc ApparatuS) Project at the Large Hadron Collider at CERN in Europe (http://atlas.ch/) is centered on a large, custom-made detector buried along a 27-km circular tunnel that accelerates hadronic elementary particles to nearly the speed of light. The particles collide within the detector, which measures the momentum and energy carried by the particles to aid in the search for the elusive Higgs particle, evidence of supersymmetry and other important aspects of contemporary high-energy physics. ATLAS will detect 100 interesting "events" per second out of about one billion that occur, and channel data into a sophisticated, tiered, distribution system. Tier 0 at CERN takes in the raw data, reconstructs the data in ways investigators can use, and sends the data to the Tier 1 sites - ten nodes in different countries - where they are distributed further to investigators through Tier 2 sites (http://cerncourier.com/cws/article/cern/31519).

ATLAS involves nearly 3,000 investigators from nearly 40 countries working in dozens of labs, institutes, departments, and universities. ATLAS started before the term *Virtual Organization* became popular, but it has been highly collaborative from the start through the cyberinfrastructure

of the ATLAS Collaboratory Project (http://atlascollab.umich.edu/). ATLAS, and the work of high-energy physics generally, demonstrates the long-standing importance of cyberinfrastructure-enabled collaboration in pursuit of Grand Challenges. The early development of the World Wide Web at CERN around 1990 is a part of this history, with consequences the entire world can now appreciate.

8.2.2 The George E. Brown Network for Earthquake Engineering Simulation (NEES)

NEES is a shared national network linking 14 research sites distributed across the United States with collaborative tools, data support, and earthquake simulation software (https://www.nees.org). Earthquake engineering experimental research was traditionally conducted at specialized facilities – shake tables, tsunami tanks, etc. – within the field's subdisciplines (geology, civil engineering, mechanical engineering, etc.). NEES employs cyberinfrastructure to facilitate new and more productive forms of collaboration, making better use of facilities and encouraging interaction among the field's specialties. This collaboration requires coordination across states, agencies, university systems, and departments to enable access to what were previously narrowly-held assets. Issues of structure, governance, funding, and ownership rights had to be negotiated and agreed upon for NEES to function.

NEES touches on two Grand Challenges. One is the need to improve the engineering of structures in a world subject to dramatic seismic events, such as the 2004 earthquake and tsunami in the Indian Ocean and the 2010 earthquake in Haiti, each of which killed more than 200,000, displaced many more, and caused billions of dollars in damage. Another is to bring together previously fragmented fields of research to achieve greater integration, enabling these fields to tackle previously elusive Grand Challenges. Many other research communities remain fragmented, moving only slowly toward coalescence.

8.2.3 The Community Earth System Model (CESM)

The CESM is a fully-coupled (atmosphere, ocean, land, biosphere, and cryosphere) global climate model with state-of-the-art computer simulations of the earth's past, present, and future climate states (http://www.cesm.ucar.edu/). It is sponsored by two different federal entities, the National Science Foundation (NSF) and the Department of Energy (DOE), with different mandates, policies, and procedures, but with overlapping interests. It is administered by the National Center for Atmospheric Research (NCAR). The CESM was built as a community of practice with the goal of collaborative learning and investigation of the earth's climate system. It is directed by a Scientific Steering Committee (SSC) of researchers from many institutions, led by a Chief Scientist, and advised by a board of scientific advisors also drawn from many institutions. Voluntary working groups involving yet more scientists from different institutions propose model components to the SSC. The entire community meets once a year at the CESM Workshop to plan work and deal with challenges. Face-to-face meetings are augmented by heavy use of teleconferencing and virtual meeting technology. The CESM has improved understanding in a number of disciplines related to climate change, and has informed policy through national and international assessments such as the Intergovernmental Panel on Climate Change.

The CESM is a Virtual Organization that embodies the current state of knowledge about the component processes of the earth system. A knowledgeable core team of scientists and software engineers work together to configure and validate the model in preparation for public release. The participants come together through the virtual organization because their own interests cannot be advanced except in conjunction with a state of the art, coupled model running on high performance computers. Unlike ATLAS, a physical instrument that measures physical phenomena, or NEES, a set of physical instruments that simulate physical phenomena, the CESM is an abstraction that attempts to capture and reflect an enormously complicated set of physical phenomena interacting with one another. This is a new and exciting frontier of research that cannot be done any other way.

8.2.4 TeraGrid

TeraGrid is one of the world's largest, most comprehensive distributed cyberinfrastructure facilities for open science research (https://www.teragrid.org/). It is an NSF-funded network of computational resources at 11 resource provider sites and 6 software integration sites distributed throughout the United States. It includes high performance computers, massive data storage systems, visualization resources, data collections, and tools connected by high-bandwidth networks and integrated by coordinated policies, operations, user support, education, outreach, and training. TeraGrid provides researchers with access to more than 60 petabytes of data storage and more than two petaflops of computing capacity that can be brought to bear on any science or engineering project. It supports complex modeling, simulation, and visualization in multiple scientific domains for multiple user communities, including chemistry, astrophysics, atmospheric science, biochemistry, biology, mathematics, earth sciences, electrical and communication systems, industrial engineering, materials research, mechanical engineering, medicine, meteorology, pharmaceutical science, physics, social science, and seismology.

Each resource provider secures its own funding and manages its own facilities and equipment, but also contributes to the common pool of computational resources, generally with support from NSF. Coordination of TeraGrid policy and planning, operation and user support, and software and services is the responsibility of the Grid Infrastructure Group (GIG). The GIG is led by the University of Chicago and includes members from the resource providers. Direction is provided by the TeraGrid Forum, consisting of the Principal Investigator of each resource provider and the GIG. In a manner similar to the CESM, each working group of scientists, managers, and technical professionals reports to a GIG member on issues of common concern and for making recommendations to the overall TeraGrid. Management and planning is coordinated via weekly and biweekly teleconferences and quarterly face-to-face meetings. Unlike ATLAS, NEES, and CESM, TeraGrid is a general-purpose cyberinfrastructure serving any domain of science. It operates as a Virtual Organization

TeraGrid is a general-purpose cyberinfrastructure serving any domain of science. It operates as a Virtual Organization to provide resources for investigators.

to provide resources for investigators.

8.3 Virtual Organizations in Grand Challenges of the Future

OCI's focus on virtual organizations recognizes the importance of cyberinfrastructure and collaboration as complementary assets, and attempts to accelerate the process of technological development and learning required to exploit the opportunities to improve productivity and effectiveness in research work. OCI draws upon its expertise in virtual organizations to accelerate the transfer of knowledge about successful and unsuccessful collaboration and coordination efforts. Deployment of this knowledge enables NSF to better support potential nascent Grand Challenge Communities.

The sociotechnical coordination of research is rightly seen as part of the "science" in the recently created program on the Science of Science and Innovation Policy, focusing on improving collaboration and stimulating creative potential [55]. Coordination, however, goes beyond policy and takes a central role in the routine activity at the heart of research. OCI's virtual organizations program recognizes this need, and should continue to refine its efforts through Virtual Organizations as Sociotechnical Systems (VOSS) program solicitations, while adding new activities to summarize and deploy best practices for virtual organizations in research [56].

8.4 OCI and Virtual Organizations

OCI has already taken a leadership role by creating NSF's first focused program on virtual organizations and by working with the SBE and CISE Directorates. This leadership role should continue, although, as noted in the larger report on collaboration, OCI cannot be expected to cover all of the topics in this broad area. OCI's continued leadership can help to catalyze the development, implementation, and evolution of a functionally complete national cyberinfrastructure that integrates physical, organizational, and cyberinfrastructural assets and services to support virtual organizations. OCI can also promote and support the establishment of world-class virtual organizations that are secure and efficient. Finally, it can support the development of common cyberinfrastructure resources, services, tools, and knowledge for effective and efficient, end-to-end cyberinfrastructure across all science and engineering fields.

There are two ways OCI can advance virtual organizations within NSF. First, OCI can expand its sponsorship of technological and organizational development relevant to virtual organizations within OCI and across NSF directorates. In addition to the VOSS solicitation, OCI should remain involved in NSF-sponsored development of collaboratories, digital repositories, observatories, science and engineering gateways, computational grids, and synthesis centers. OCI's active participation in the cross-directorate initiative Cyber-Enabled Discovery and Innovation (CDI) has helped promote innovation in computational thinking (http://www.nsf.gov/crssprgm/cdi/). In addition to virtual organizations, CDI embraces "understanding complexity" and moving "from data to knowledge." As CDI draws to a close, OCI will participate in a new cross-directorate program, tentatively titled Research Coordination Networks. This effort should continue the advance of virtual organizations across disciplinary, organizational, institutional, and geographical boundaries.

Second, OCI should gather what has been

learned about virtual organizations and bring best practices into the requirements and specifications of research and development. The best practices, as well as the technologies and other complementary elements of virtual organizations, can be incorporated by NSF into solicitations and reports, expectations of program officers, and the expertise of reviewers, panelists, committees of visitors, and PIs. This approach should stimulate the development of proposals that include sensible plans for virtual organizations as well as criteria for assessments of success. OCI can facilitate the creation and maintenance of resource repositories, workshops, and meetings to disseminate information about best practices in distributed, interdisciplinary, cyberinfrastructure-enabled virtual organizations.

8.5 Recommendations

The benefits of virtual organizations to scientific and engineering productivity are pervasive and difficult to single out, as are the issues relative to integrating them into a large-scale cyberinfrastructure. These and related issues are worthy of further study.

RECOMMENDATIONS:

The NSF should initiate a thorough study outlining best practices, barriers, success stories, and failures, on how collaborative interdisciplinary research is done among diverse groups involved in Grand Challenge projects.

The NSF should invest in research on virtual organizations that includes:

1) Studying collaboration, including virtual organizations, as a science in its own right;

2) Connecting smaller virtual organizations to the large-scale infrastructure by providing supplementary funds to such projects, supporting development of tools, applications, services, etc. with a mandate to disseminate those elements to other communities and users;

3) Investing in systematic, rigorous project-level and program-level evaluations to determine the benefits from virtual organizations for scientific and engineering productivity and innovation;

4) Encouraging NSF program officers to share information and ideas related to virtual organizations with training and online management tools.

9 Concluding Comments

Formidable science and engineering Grand Challenges that affect our nation's welfare, security, and competitiveness loom ahead that can be addressed by advances in CS&E enabled by advances in cyberinfrastructure. These advances will require the development of collaborative communities of researchers from diverse areas of science and engineering, and innovative virtual organizations, and this in itself will represent a challenging undertaking. The National Science Foundation, through the Office of Cyberinfrastructure, can play a fundamental role in addressing these challenges and advancing the frontiers of scientific discovery and enabling innovative advances in engineering. It is hoped that this study provides insight and recommendations that will be useful in structuring strategic programs within the Foundation that will aid in accomplishing these ends.

Appendix A: OCI – GCC's and VO's Workshops

August 25, 2009 - Workshop Attendees

Workshop Organizers

J. Tinsley Oden *(U Texas-Austin*)
John Leslie King *(U Michigan)*
Jon Bass *(U Texas-Austin)*
Omar Ghattas *(U Texas-Austin)*
Barry I. Schneider *(NSF)*

Universities

Klaus Bartschat *(Drake U)*
Donald Estep *(Colorado State U)*
Michael Gurnis (Caltech)
C. William McCurdy *(UC Davis)*
Linda Petzold *(UC Santa Barbara)*
Klaus Schulten *(U Illinois – UC)*
Cathy Wu (*U Delaware)*
Thom Dunning *(U Illinois – UC)*
Sharon Glotzer(*U Michigan)*
James Kinter *(IGES Inc.)*
Abani Patra *(U Buffalo)*
Tamar Schlick *(New York U)*
Victoria Stodden (Yale Law School)
Katherine Yelick *(UC Berkeley)*

Government Laboratories

John Drake *(ORNL)*

NSF

Paul Messina *(Consultant, OCI)*
Abani Patra *(OD/OCI)*
Barry I. Schneider *(OD/OCI)*
Susan J. Winter *(OD/OCI)*
Several other NSF participants.
Suzanne Iacono *(CISE/CNS)*
Edward Seidel *(MPS/OAD)*
Judith Sunley *(SBE/OAD)*

April 22-23, 2010 - Workshop Attendees

Workshop Organizers

J. Tinsley Oden *(U Texas-Austin*) *Chair*
John Leslie King *(U Michigan)* Co-Chair
Thom Dunning *(U Illinois - UC)*
Michael Gurnis *(Caltech)*
Linda Petzold *(UC Santa Barbara)*
Cathy Wu *(U of Deleware)*
Omar Ghattas *(U Texas-Austin) Co-Chair*
Barry I. Schneider *(NSF) Liaison*
Donald Estep *(Colorado St)*
Abani Patra *(U Buffalo)*
Victoria Stodden *(Yale Law School)*

Universities

Guy Almes *(Texas A&M)*
Jon Bass *(U Texas-Austin)*
Warren Bicknell Mori *(UCLA)*
James Brasseur *(Penn State)*
Hai-Ping Cheng *(U Florida)*
Thomas Cheatham III *(U of Utah)*
Peter Cummings *(Vanderbilt)*
John Drake *(U of Tennessee)*
Robert Fisher *(U of Massachusetts)*
James French *(U of Virginia)*
Gwen Jacobs *(Montana State U)*
George Karniadakis *(Brown U)*
Alexei Khokhlov *(U of Chicago)*
Rubin Landau *(Oregon State U)*
William Lester *(UC Berkeley)*
Philip Maechling *(USC)*
W. Richard McCombie *(CSH Lab.)*
Donald Pellegrino *(Drexel University)*
Karl Schulz *(TACC)*
Valerie Taylor *(Texas A&M)*
Renata Wentzcovitch *(U of Minnesota)*
Nancy Wilkins-Diehr *(SDSC)*
P. K. Yeung *(Georgia Tech)*
Lorena Barba *(Boston U)*
Jerry Bernholc *(North Carolina State U)*
George Biros *(Georgia Tech)*
Richard Brower *(Boston U)*
Ronald Cohen *(Carnegie Inst. of Washington)*
Michael Clark *(Harvard)*
Thomas Dietterich *(Oregon State U)*
Jacob Fish *(Rensselaer Polytechnic Inst)*
Geoffrey Fox *(Indiana U)*
Lincoln Greenhill *(Harvard)*
Lennart Johnson *(U of Houston)*
Daniel Katz *(U of Chicago)*
Jeongnim Kim *(UIUC)*
Alan Laub *(UCLA)*
Wing Kam Liu *(Northwestern U)*
Dimitri Mavriplis *(U of Wyoming)*
Richard Moore *(SDSC)*
Ralph Roskies *(PSC)*
Mark Shephard *(Rensselaer Polytechnic Inst)*
Homer Walker *(Worcester Polytechnic Inst)*
Phillip Westmoreland *(U of Massachusetts)*
Paul Woodward *(U of Minnesota)*
John Ziebarth *(Krell Inst)*

Government Agencies and Laboratories

Randy Avent *(OSD)*
Robert Bonneau *(AFOSR)*
Anne Chaka *(NIST)*
Frederica Darema *(AFOSR)*
Thomas Pinelli *(NASA)*
Sharon Welch *(LARC)*
Amber Boehnlein *(FNAL)*
Kimberly Budil *(DOE)*
Lee Collins *(LANL)*
Mark Pederson *(DOE)*
Taiching Tuan *(Army)*

Industry

Susan Fratkin *(Fratkin Assoc.)*
David Salzman *(LightSpin Tech.)*

NSF

Estela Blaisten-Barojas *(MPS/CHE)*
Clark Cooper *(ENG/EFRI/CMMI)*
Evelyn Goldfield *(MPS/CHE)*
Daryl Hess *(MPS/DMR)*
Bradley Keister *(MPS/PHY)*
Jacqueline Meszaros *(SBE/SES)*
Manish Parashar *(OD/OCI)*
Irene Qualters *(OD/OCI)*
Barry I. Schneider *(OD/OCI)*
Serdar Ogut *(MPS/DMR)*
Almadena Y. Chtchelkanova *(CISE/CCF)*
Cheryl Eavey *(SBE/SES)*
Horst Henning Winter *(ENG/CBET)*
Leland Jameson *(MPS/DMS)*
Fae Korsmo *(OD)*
Eduardo Misawa *(ENG/CMMI)*
Joy Pauschke *(ENG/CMMI)*
Thomas Russell *(OD/OIA)*
Edward Seidel *(MPS/OAD)*

Appendix B: ACCI Recommendation Letter for the Creation of a Program in CDS&E

Dear Dr. Bement,

At the May 2010 meeting, the National Science Foundation Advisory Committee for Cyberinfrastructure unanimously endorsed the following recommendation:

> **The National Science Foundation should create a program in Computational and Data-Enabled Science and Engineering (CDS&E), based in and coordinated by the NSF Office of Cyberinfrastructure.**

The new program should be collaborative with relevant disciplinary programs in other NSF directorates and offices.

Computational and Data-Enabled Science and Engineering (CDS&E) is now clearly recognizable as a distinct intellectual and technological discipline lying at the intersection of applied mathematics, computer science, and core science and engineering disciplines. It is dedicated to the development and use of computational methods and data mining and management systems to enable scientific discovery and engineering innovation.

CDS&E builds on the area of Computational Science and Engineering, growing out of scientific computation and the explosion of production of digital data. We regard CDS&E as explicitly recognizing the importance of data-enabled, data-intensive, and data-centric science. CDS&E broadly interpreted now affects virtually every area of science and technology, revolutionizing the way science and engineering are done. Theory and experimentation have for centuries been regarded as two fundamental pillars of science. It is now widely recognized that computational and data-enabled science forms a critical third pillar. CDS&E includes new methodologies for science and engineering that are indispensible to the nation's welfare, competitiveness, and standing in the international scientific community and global economy.

Computational and Data-Enabled Science and Engineering (CDS&E) is fundamentally important to the long-term NSF strategic initiative called CF21: Cyberinfrastructure Framework for 21st Century Science and Engineering. The NSF CF21 vision calls for a "comprehensive plan for education and outreach in computational science to support learning and workforce development for 21st century science and engineering."

NSF can make a strong statement that will lead the Foundation, researchers it funds, and US universities and colleges generally, by recognizing Computational and Data-Enabled Science and Engineering as the distinct discipline it has clearly become.

Approved
Arden L. Bement, Jr.
Director
National Science Foundation

05/27/2010
Date

1

Bibliography

1. Alcubierre, M., Brügmann, B., Diener, P., Koppitz, M., Pollney, D., Seidel, E., and Takahashi, R., *Phys. Rev. D*, 67, 084023 (2003), gr-qc/0206072.
2. Aprà, E., Harrison, R.J., deJong, W.A., Rendell, A. P., Tipparaju, V., and Xantheas, S.S., "Liquid Water: Obtaining the Right Answer for the Right Reason", *Proceeding of SC09*, Portland, Oregon, November 2009.
3. Bader, D.A., Roshan, U., and Stamatakis, "A., Computational Grand Challenges in Assembling the Tree of Life: Problems & Solutions", *The IEEE and ACM Supercomputing Conference 2005 (SC2005) Tutorial*, 2005. DOI 10.1.1.83.9548.
4. Berger, M.J., and Oliger, J., *J. Computational Physic*s, 53, 484, 1984.
5. *Board on Research Data and Information.* National Research Council. http://sites.nationalacademies.org/PGA/brdi/index.htm
6. Buras, R., Janka, H.,T., Rampp, M., and Kifonidis, K., *Astronomy & Astrophysics J.*, 457, 281 , 2006, astro-ph/0512189.
7. Burrows, A., Livne, E., Dessart, L., Ott, C.D., and Murphy, J., *Astrophysics Journal,* 655, 416, 2007.
8. Burrows, A., Dessart, L., Livne, E., Ott, C.D., and Murphy, J., *Astrophysics Journal,* in print, 2007.
9. *Challenges in Climate Change Science and the Role of Computing at the Extreme Scale*. DOE Workshop Report, November 6-7, 2008, Washington D.C. http://extremecomputing.labworks.org/climate/report.stm
10. *CIPRES: Building the Tree of Life: A National Resource for Phyloinformatics and Computational Phylogenetics.* http://www.phylo.org/
11. *Computational Science: Ensuring America's Competitiveness*. Benioff, M., and Lazowska, E.D., President's Information Technology Advisory Committee (PITAC), 2005. www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf.
12. CPATH 2009 Portfolio. http://www.nsf.gov/cise/funding/CPATH2009awards.pdf
13. *Cyberinfrastructure Vision for 21st Century Discovery,* NSF Cyberinfrastructure Council, March 2007.
14. *Dynamic Data Driven Applications Systems)*, Darema, F., Douglas, C., and Deshmukh, A., Eds., Report of the NSF Workshop Report, March 2000, Arlington, VA. National Science Foundation. Available online at http://www.nsf.gov/cise/cns/dddas/.

15. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Committee on Science, Engineering, and Public Policy, National Academies. The National Academies Press, July 2009.

16. *Exascale Computing Study: Technology Challenges in Achieving Exascale Systems*. Kogge, P., Ed., DARPA IPTO Report, September 2008.

17. *ExaScale Software Study: Software Challenges in Extreme Scale System*, DARPA Information Processing Techniques Office (IPTO), September 2009. http://users.ece.gatech.edu/mrichard/ExascaleComputingStudyReports/ecss%20report%20101909.pdf

18. Fryer, C.L., and Warren, M.S., *Astrophysics Journal,* 601, 391, 2004, arXiv: astro-ph/0309539.

19. GEO 600. http://www.geo600.uni-hannover.de/

20. Glotzer, S.C., Panoff, R., and Lathrop, S., "Challenges and opportunities in preparing students for petascale computational science and engineering", *Comping in Science and Engineering*, 11(5), pp. 22-27, 2009.

21. *Graduate Education in Computational Science and Engineering*. SIAM Working Group on CSE in Education, *SIAM Review*, vol. 43, no. 1, 2001, pp. 163–177.

22. *Harnessing the Power of Digital Data for Science and Society*. Report of the Interagency Working Group on Digital Data (IWGDD) to the Committee on Science of the National Science and Technology Council (NSTC), January 2009.

23. Hehre, W.J., Lathan, W.A., Ditchfield, R., M., Newton, M.D., and Pople, J.A., "Gaussian 70" (*Quantum Chemistry Program Exchange, Program* No. 237, 1970).

24. *International Assessment of R&D in Simulation-Based Engineering and Science*. S.C. Glotzer et al., World Technology Evaluation Center, 2009. http://www.wtec.org/sbes/SBES-GlobalFinalReport.pdf

25. *International Exascale Software Project Roadmap*, DOE ASCR, January 2010. http://www.exascale.org/mediawiki/images/a/a1/Iesp-roadmap-draft-0.93-complete.pdf

26. iPlant Collaborative Web Portal. http://www.iplantcollaborative.org

27. *Jim Gray presentation to the NRC-CSTB* (National Research Council-Computer Science and Telecommunications Board) in Mountain View, CA, on January 11, 2007. http://research.microsoft.com/en-us/um/people/gray/JimGrayTalks.htm

28. Johnson, C., "Top Scientific Visualization Research Problems", *IEEE Computer Graphics and Applications*, pp. 2-6, July/August, 2004.

29. Large Hadron Collider (LHC). http://public.web.cern.ch/Public/en/LHC/Computing-en.html

30. LIGO: Laser Interferometer Gravitational Wave Observatory. http://www.ligo.caltech.edu/

31. *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century*. National Science Board, National Science Foundation, 2005. www.nsf.gov/pubs/2005/nsb0540/start.jsp

32. *Mathematics for Analysis of Petascale Data,* DOE Workshop Report, 2008. http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/PetascaleDataWorkshopReport.pdf

33. Mavripilis, D.J., "High Performance Computational Engineering: Putting the E back in CSE", *Proceedings of the 21st International Conference on Parallel Computational Fluid Dynamcis (ParCFD 2009)*, Moffet Field, CA, May 2009, DEStech Publications, Lancaster PA, 2010.

34. Mészáros, P., *Reports of Progress in Physics,* 69, 2259, 2006, astro-ph/0605208.

35. *Modeling and Simulation at the Exascale for Energy and the Environment*. Stevens, R., T. Zacharia, T., and Simon, H., Department of Energy Office of Advance Scientific Computing Reserach, Washington, DC, Report on the Advanced Scientific Computing Research Town Hall Meetings on Simulation and Modeling at the Exascale for Energy, Ecological Sustainability and Global Security (E3), pp. 174, 2008. http://science.doe.gov/ascr/ProgramDocuments/Docs/TownHall.pdf

36. Moret, B.M.E., *Computational Challenges from the Tree of Life*. http://compbio.unm.edu (www.phylo.org/docs/alenex.pdf)

37. *NIH-NSF Visualization Research Challenges Report*. Johnson, C., Moorhead, R., Munzner, T., Pfister, H., Rheingans, P., and Yoo, T. S., Eds.; IEEE Press, ISBN 0-7695-2733-7, 2005. *http://www.computer.org/portal/pages/store/b2005/r0235.xml*

38. Pretorius, F., *Phys. Rev. Letters*, 95, 121101, 2005, gr-qc/0507014.

39. *Revolutionizing Science and Engineering through Cyberinfrastructure*, Atkins, D.E., et al., NSF, 2003. www.nsf.gov/od/oci/reports/toc.jsp

40. *SBE&S Vision Report, Cummings*, P.T., and Glotzer, S.C., in press.

41. Schnetter, E., Ott, C.D., Allen, G., Diener, P., Goodale, T., Radke, T., Seidel, E., Shalf, J., *Cactus Framework: Black Holes to Gamma Ray Bursts, in Petascale Computing: Algorithms and Applications,* D. Bader, Ed., CRC Press LLC (2007), arXiv:0707.1607.

42. See: http://www.gaussian.com/

43. See: http://www.linux.org/

44. See: http://www.exascale.org/iesp/Main_Page

45. See: http://www.ks.uiuc.edu/Research/namd/

46. See: http://www.ncrr.nih.gov/

47. See: http://www.bisti.nih.gov/.

48. See: http://www.scidac.gov/.

49. See: http://ngwww.ucar.edu/.

50. See: http://www.unidata.ucar.edu/software/netcdf/.

51. See: http://www.ccsm.ucar.edu/.

52. See: http://www.wrf-model.org/index.php.

53. See: http://cactuscode.org/community/projects/alpaca/

54. See: http://www.nsf.gov/news/news_summ.jsp?cntn_id=116336

55. See: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=501084&org=sbe

56. See: http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503256&org=NSF&sel_org=NSF&from=fund See also the report of the NSF Workshop Building Effective Virtual Organizations held in January 2008. http://www.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf

57. See http://twiki.ipaw.info/bin/view/Challenge/FirstProvenanceChallenge , the U.K. funded Taverna software package at http://www.mygrid.org.uk/ , the Pegasus system developed at the University of Southern California http://pegasus.isi.edu/ , and Galaxy software developed at Penn State University http://galaxy.psu.edu/ . Provenance and workflow tracking software is currently heavily focused on bioinformatics and biology research. See Microsoft's Trident Workbench for an oceanography example: http://research.microsoft.com/en-us/collaboration/tools/trident.aspx .

58. *Simulation-Based Engineering Science: Revolutionizing Engineering Science through Simulation*, Oden, J.T., et al., NSF Blue Ribbon Panel on SBES, 2006. http://www.nsf.gov/pubs/reports/sbes_final_report.pdf

59. Stodden, V., "Enabling Reproducible Research: Licensing for Scientific Innovation", *International Journal of Communications Law and Policy*, 13(1), 2009.

60. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information.* Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, February 2010. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

61. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Hey, T., Tansley, S., and Tolle, K., Eds., Microsoft Research, Redmond, Washington, October 2009.

62. *The Next Big Climate Change*. Editorial in *Nature*, 453, pp. 257 (15 May 2008) http://www.nature.com/nature/journal/v453/n7193/full/453257a.html

63. *Undergraduate Computational Science and Engineering Education*. SIAM Working Group on CSE in Undergraduate Education, SIAM Review, submitted.

64. VIRGO. http://www.virgo.infn.it/

65. *Visualization and Knowledge Discovery,* Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale, October 2007. http://science.doe.gov/ascr/ProgramDocuments/Docs/DOE-Visualization-Report-2007.pdf

66. *White Paper: U.S. Manufacturing – Global Leadership Through Modeling and Simulation*. Council on Competitiveness, 2009. http://www.compete.org/publications/detail/652/us-manufacturingglobal-leadership-through-modeling-and-simulation/

67. Wissink, A.M., Katz, A.J., Chan, W.M., and Meakin, R.L., "Validation of the Strand Grid Approach", AIAA Paper 2009-3792, *19th AIAA Computational Fluid Dynamics Conference*, San Antonio TX, June 2009.

68. *World Modeling Summit for Climate Prediction*. Shukla, J., Ed., Workshop Report, Reading, UK, January 2009.

http://www.nsf.gov/od/oci/taskforces