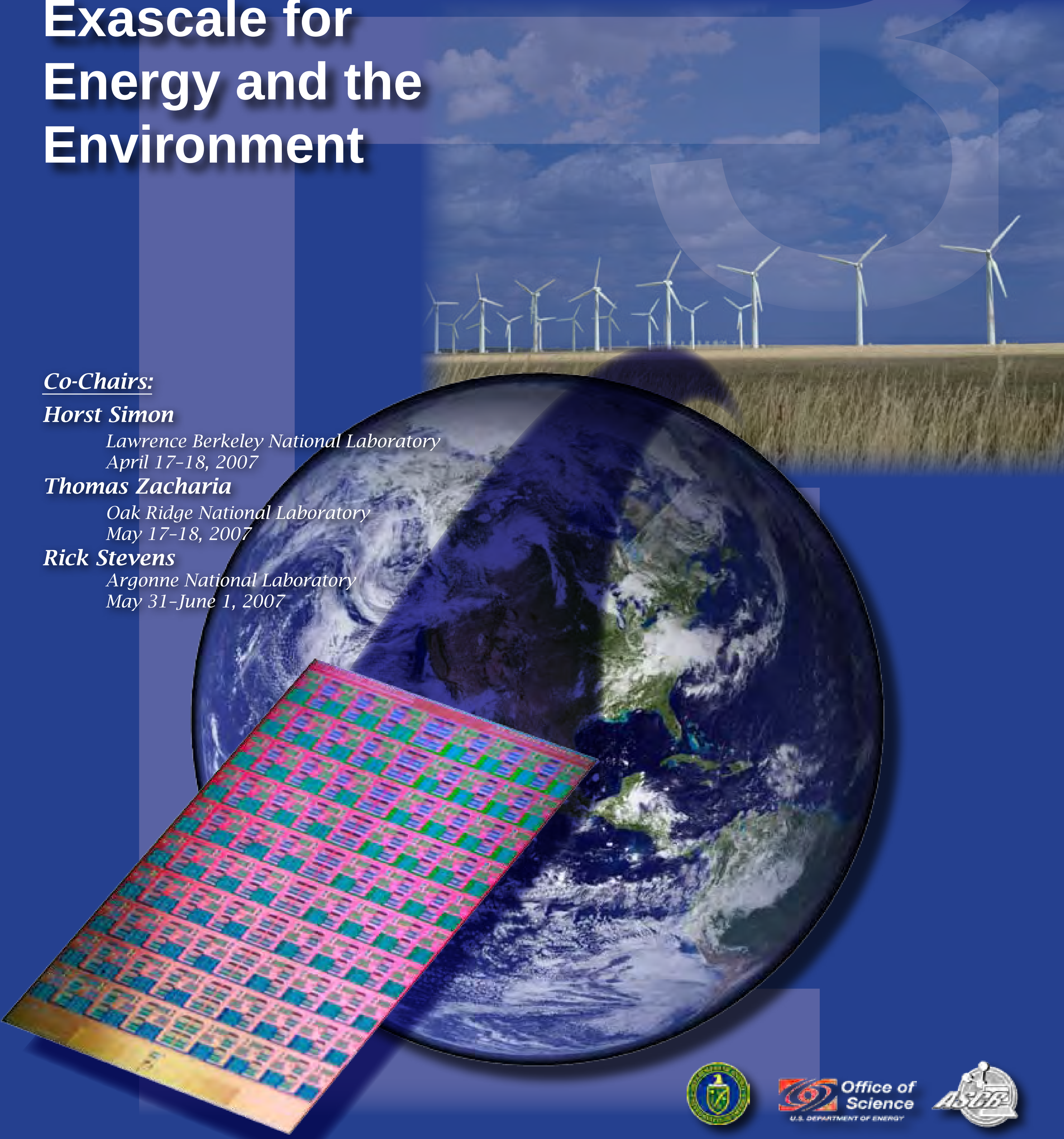
Modeling and Simulation at the Exascale for Energy and the Environment



Co-Chairs:

Horst Simon

Lawrence Berkeley National Laboratory
April 17–18, 2007

Thomas Zacharia

Oak Ridge National Laboratory
May 17–18, 2007

Rick Stevens

Argonne National Laboratory
May 31–June 1, 2007

Office of Science
U.S. DEPARTMENT OF ENERGY

On the cover:

Teraflops Research Chip - Intel Corporation

Wind Farm - Randy Hayes, Bureau of Land Management

Blue Marble - Reto Stöckli, NASA Goddard Space Flight Center Image

Modeling and Simulation at the Exascale for Energy and the Environment

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)

Co-Chairs:

Lawrence Berkeley National Laboratory: Horst Simon
Oak Ridge National Laboratory: Thomas Zacharia
Argonne National Laboratory: Rick Stevens

Office of Advanced Scientific Computing Research Contact: Michael Strayer

Special Assistance

Technical:

Lawrence Berkeley National Laboratory:
Deb Agarwal, David Bailey, John Bell, Wes Bethel, Julian Borrill, Phil Colella, William Collins, Nikos Kyrpides, Victor Markowitz, Juan Meza, Norman Miller, Peter Nugent, Leonid Oliker, Arie Shoshani, Erich Strohmaier, Brian Tierney, Lin-Wang Wang, Michael Wehner

Oak Ridge National Laboratory:
Eduardo D'Azevedo, David Bernholdt, John Drake, David Erickson, George Fann, James Hack, Victor Hazlewood, Al Geist, Igor Jouline, Douglas Kothe, Bronson Messer, Anthony Mezzacappa, Jeff Nichols, Stephen Poole, B. (Rad) Radhakrishnan, Nagiza Samatova, Srdjan Simunovic, Scott Studham, Jeffrey Vetter, Gilbert Weigand

Argonne National Laboratory:
Raymond Bair, Pete Beckman, Charles Catlett, Robert Edwards, Paul Fisher, Ed Frank, Ian Foster, William Gropp, Ahmed Hassanein, Mark Hereld, Paul Hovland, Robert Jacob, Kenneth Kemner, Veerabhadra Kotamarthi, Don Lamb, Ewing Lusk, Jorge Moré, Lois McInnes, Folker Meyer, Boyanna Norris, David Nowak, Michael Papka, Robert Ross

Administrative:

Lawrence Berkeley National Laboratory: Jon Bashor, Yeen Markin
Oak Ridge National Laboratory: Linda Malone, Debbie McCoy
Argonne National Laboratory: Kathy Dibennardi, Janet Werner, Cheryl Zidel

Publication:

Oak Ridge National Laboratory: Creative Media
Argonne National Laboratory: Joseph Insley

Editorial:

Oak Ridge National Laboratory: Agatha Bardoel, Bonnie Nestor
Argonne National Laboratory: Gail Pieper

This report is available on the web at http://www.sc.doe.gov/ascr/ProgramDocuments/ProgDocs.html

Contents

Executive Summary

Exascale Town Hall Meetings
Letter Report

Lawrence Berkeley, Oak Ridge, and Argonne national laboratories convened three town hall meetings aimed at collecting community input on the prospects of a proposed new Department of Energy (DOE) initiative entitled Simulation and Modeling at the Exascale for Energy and the Environment, or E3 for short.

The goal of the town hall meetings was to engage the computational science community in a series of broad and open discussions about the potential benefits of advanced computing at the exascale (10^{18} operations per second) on "global" challenge problems in the areas of energy, the environment, and basic science.

Approximately 450 researchers from universities, national laboratories, and U.S. companies participated at the three meetings held in April, May, and June 2007.

In addition to the scientific and engineering challenges and opportunities, the meetings also addressed needed advances in computer science and software technology, large-scale hardware, applied mathematics, and cyberinfrastructure and cyber security.

Here we summarize the major conclusions of the town hall meetings.

Feasibility of Exascale Systems

General-purpose exascale computer systems are expected to be technologically feasible within the next 15 years. These systems will likely have between 10 million and 100 million processing elements or cores. The major U.S. vendors of large-scale systems and processors (e.g., IBM, Intel, Cray, AMD) are in general agreement that these systems will push the envelope of a number of important technologies, including processor architecture, scale of multicore integration (perhaps into the range of 1000 cores per chip or beyond), power management, and packaging. The projected exascale systems themselves will have part counts comparable to those of today's largest systems (or slightly larger). Detailed cost studies have not been done, but the consensus is that costs will be comparable to those of the largest systems being contemplated today ($100 million to $200 million per system).

Significant challenges arise in accomplishing exascale computing, in areas that include architecture, scale, power, reliability, cost, and packaging. A major source of uncertainty is how quickly the general marketplace will be able to adopt highly parallel, single-chip, multicore systems in normal information technology (IT) products. The current belief is that the broad market is not likely to be able to adopt multicore systems at the 1000-processor level without a substantial revolution in software and programming techniques for the hundreds of thousands of programmers who work in industry and do not yet have adequate parallel programming skills.

Extrapolation of current hardware trends suggests that exacale systems could be available in the marketplace by approximately 2022 via a "business as usual" scenario. With the appropriate level of investments, it may be possible to accelerate the availability by up to five years, to approximately 2017.

Exascale systems will also require substantial investments in input/output (I/O) and storage research. The current trends in disk drives and other storage technologies are optimized for the consumer market and may not have the optimal ratios of capacity to bandwidth needed for large-scale systems.

Power efficiency is also expected to be a major problem, with the goal of an exaflops system at less than 20 MW sustained power consumption perhaps achievable. Driving the earlier availability of the systems will compromise the power efficiencies to some degree.

We note that Japan has outlined in its current petascale initiative a rough roadmap to the exascale that proceeds via three systems: a 10 petaflops (PF) system in ~2012, a 100 PF system in ~2017, and a 1000 PF system in the ~2022 timeframe. It appears possible for a U.S. computing program to maintain leadership during the next decade in this area – but only if increased investments are started immediately and are sustained over the long term.

Science and Engineering Opportunities

The three town hall meetings examined a range of applications that would be materially transformed by the availability of exascale systems. We highlight here several significant opportunities in the areas of energy, climate, socioeconomics, biology, and astrophysics.

Energy

Energy research offers significant opportunities to exploit computing at the exascale, in order to advance our understanding of basic processes in areas such as combustion, which would naturally lead to a design capability for improving the efficient use of liquid fuels, whether from fossil sources or renewable sources. First-principles computational design and optimization of catalysts will become possible at the exascale, as will *de novo* design of biologically mediated pathways for energy conversion.

Access to exascale systems and the appropriate applications codes could have a dramatic impact on nuclear fission reactor design and optimization and would help accelerate understanding of key plasma physics phenomena in fusion science critical to getting the most from the U.S. investment in ITER.

Exascale systems should also enable a major paradigm shift in the use of large-scale optimization techniques to search for near-optimal solutions to engineering problems. Many energy and industrial problems are amenable to such an approach, in which many petascale instances of the problem are run simultaneously under the control of a global optimization procedure that can focus the search on parameters that produce an optimal outcome.

Environment

Three broad areas relating to the environment were discussed: climate modeling; integrated energy, economics, and environmental modeling; and multiscale biological modeling from molecules to ecosystems.

Climate modeling. As the most mature of the three environmental application areas, climate modeling is expected to make good use of exascale systems. The impact of these systems will be threefold. First, they will enable the development of much higher resolution models that will advance our understanding of local impacts of climate change; second, they will enable the dramatic improvement of physical, chemical, and biological process representations in the climate models, which will more accurately reflect the real climate system; and third, they will enable a thorough exploration of the parameters that give rise to uncertainty in climate models via large-scale ensemble computations. Significant investments will be needed, however, to port and improve climate models for exascale architectures, including the explicit targeting of multicore in next-generation models and the development of an integrated climate research computing environment that will link climate modelers with climate data sources, collaborators, and university and laboratory resources.

Integrating energy, socioeconomics, and environmental modeling. There exists a considerable opportunity to couple detailed computer models of energy utilization and production with geospatialized socioeconomic models and, in turn, to couple these to an Earth systems model that captures the feedbacks to and from the environment from human activities. This integrated modeling suite would enable fundamental research into strategies for sustainable global economic development and would lead to exploration of alternative development paths and their impacts on global energy security.

Multiscale biological modeling. Large-scale computing is starting to have an impact in the biological sciences. Exascale computing will enable computational biologists to begin to build models that can bridge the space-time parameters that characterize important biological processes, including models of diverse microbial ecosystems from which we may gain considerable new biotechnology (bioenergy, carbon sequestration, environmental technology, and industrial processes). Bridging the scales from the molecular to the ecosystem offers many challenges for model developers, but it also provides many opportunities for coupling research in high-throughput genomics, proteomics, and bioinformatics to applications via exascale computing. This activity is key, for example, to accelerating the computing vision of programs such as DOE's Genomics:GTL initiative.

Astrophysics

Simulation opportunities in astrophysics include large-scale structure formation, galaxy formation, stellar evolution, supernovae, and compact objects. For example, in the Type Ia supernova problem, exascale computing will enable simulations with resolutions down to the Gibson scale (the length scale at which turbulent motion is effectively smoothed by the propagation of the nuclear flame) with definitive prescriptions for nuclear energy release and the associated nucleosynthesis.

Core collapse supernovae simulations with the spatial resolution required to properly model critical aspects of the explosion dynamics (e.g., the evolution of the stellar core magnetic fields and their role in generating the supernova) will require much higher resolution than today's terascale codes. These codes, in turn, will require exascale computing, particularly if a number of simulations are to be performed across the range of stellar progenitors and input physics. One such simulation is expected to take ~8 weeks, assuming 20% efficiency on an exaflops machine.

Advances in these areas will require the adaptation of existing, and in some cases the development of new solution algorithms for the underlying partial differential equations governing the evolution of these astrophysical systems and the codes that execute them, as well as the optimization of both as they advance to the exascale.

Computer Science and Applied Mathematics

To realize science at the exascale will require a concerted effort to couple advances in algorithms, programming models, operating systems, filesystems, I/O environments, and data analysis tools. In fact, exascale systems are likely to be so demanding that they will drive new working relationships between the disciplinary scientists and the computer science and mathematics communities.

Of great interest are methods that will enable the power of exascale computing to advance the use of mathematical optimization in many areas of science and engineering. Examples include the use of ensembles and "outer loop" optimization to iterate design parameters of new nuclear reactor designs that would simultaneously improve safety margins and lower cost, or to explore the parameter space of technology choices and how they might impact global energy security strategies.

Specific challenges that need to be overcome include development of scalable operating system services that can manage 10 million to 100 million cores, scalable programming models and tools that will enable developers to express orders of magnitude more concurrency in applications, and data storage environments that can scale to exabytes of capacity and sustained transfer speeds of terabytes per second. While reaching the needed scaling and performance goals will be a challenge, the community believes that it is possible and achievable on a schedule that would not limit the prospect of accelerating availability of exascale systems to 2017.

Cyberinfrastructure and Cyber Security

Large-scale computing resources are only a part of the overall computing environment needed to advance science. This environment also includes high-performance networking, mid-range and smaller clusters, visualization engines, large-scale data archives, a variety of data sources and instrumentation including emerging sensor networks, and the tens of thousands of workstations that enable access to and are the primary development machines. Complementing the hardware and networking is a vast software ecosystem that connects resources and enables them to work as part of a whole, spanning networking software, databases, security, and hundreds of domain-specific tools. This overall collection, commonly referred to as "cyberinfrastructure," will require investments to fully exploit the power and promise of exascale computing. The quality and robustness of the cyberinfrastructure will impact the productivity of the exascale computing resources. Additional investments in cyberinfrastructure and cyber security are needed to ensure that large-scale systems will be productive and secure. While extreme-scale systems are sometimes the targets of security attacks, they are generally well protected by layers of infrastructure. On the other hand, the general cyberinfrastructure that provides the rich computing environment surrounding the extreme-scale systems is often vulnerable. Clearly, we must be wise in our development of exascale systems to make balanced investments in the security of the overall scientific computing environment.

Conclusions

The broad computational science community has a golden opportunity to accelerate the availability of usable exascale systems. To take full advantage of this opportunity to deliver exascale computing by 2017 will require an integrated program of investments in hardware and software research and development, (R&D). Also required will be a tight coupling to a selected set of science communities and the associated applied mathematics R&D. In some cases, such as astrophysics and climate, the communities are well on the way to exploiting petascale systems. In other cases, such as socioeconomics and multiscale biology, there is great opportunity for acceleration. Computational science and engineering opportunities in energy are wide and deep and have an enormous potential impact on advancing energy technology and fundamental science. If acceleration is to be achieved—and there is every reason to both desire it and believe that it can be accomplished—then every minute will count, and even modest investments early in the cycle (e.g., 2008 and 2009) could have dramatic benefit and will reduce uncertainties moving ahead.

Rick Stevens, Argonne National Laboratory

Thomas Zacharia, Oak Ridge National Labortory

Horst Simon, Lawrence Berkeley National Laboratory

Introduction

The U.S. Department of Energy (DOE) Office of Advanced Scientific Computing Research (OASCR) has proposed a 10-year initiative on Simulation and Modeling at the Exascale for Energy, Ecological Sustainability, and Global Security (E3). This initiative, which is aligned with the strategic theme of scientific discovery and innovation in DOE's Strategic Plan, is designed to focus the computational science experience gained over the past ten years on the opportunities that will be introduced with exascale computing to ***revolutionize our approaches to global challenges in energy, environmental sustainability, and security***. A summary of the E3 initiative is presented in Appendix A.

Planned petascale and potential exascale systems provide an unprecedented opportunity to attack these global challenges through modeling and simulation. In combination with theory and experiment, computation has become a critical tool for understanding the behavior of the fundamental components of nature and for exploring complex systems with billions of components, including humans. Computing has already been used in partnership with theory and experiment to attack such problems as the time evolution of atmospheric CO_2 concentrations originating from the land surface, the activity of the cellulase enzyme on a time scale of 50 to 100 nanoseconds (ns), the stabilization of lifted flames in diesel engines and gas turbine combustors, and the behavior of superheated ionic gases in plasmas.

The deployment in this decade of several systems with peak performance in the range of 10^{18} operations per second (petaflops), enabling simulations sustaining hundreds of teraflops, should be followed in the next decade by systems with peak performance in the exaflops range and simulations sustaining a hundred or more petaflops. Exascale computers will have processing capability similar to that of the human brain and offer the potential to unravel scientific mysteries that we have not yet been able to address. Examples relevant to DOE missions include:

- Resolving clouds, forecasting weather and extreme events, and providing quantitative mitigation strategies
- Understanding high-temperature superconductivity
- Developing clean and efficient combustion systems for diesel and alternative fuels
- Developing a detailed understanding of cellulase enzyme mechanisms and creating more efficient enzymes for cellulose degradation through protein engineering
- Understanding the interaction of radiation with materials
- Advancing magnetic fusion through predictive capabilities with core-edge coupling, realistic mass ratios, and validated turbulence models for ITER
- Explaining and predicting core-collapse supernovae and putting theories of general relativity, dense equation of state (EOS), and stellar evolution to the test

Equally important, leading the development, acquisition, and deployment of exascale systems has the potential to ***make U.S. industry more competitive*** and to enable the solution of problems of national importance. In response to the OASCR initiative, Argonne National Laboratory (ANL), Lawrence Berkeley National Laboratory (LBNL), and Oak Ridge National Laboratory (ORNL) organized a community input process in the form of three town hall meetings (see Table I.1). The agendas of these meetings are provided in Appendix B. About 450 participants, listed in Appendix C, attended these three town hall meetings and contributed to this report.

Location	Date	Web site
Lawrence Berkeley National Laboratory	April 17–18, 2007	http://hpcrd.lbl.gov/E3SGS/main.html
Oak Ridge National Laboratory	May 17–18, 2007	http://computing.ornl.gov/workshops/town_hall/index.shtml
Argonne National Laboratory	May 31–June 1, 2007	https://www.cls.anl.gov/events/workshops/townhall07/index.php

Table I.1 Town hall meetings on Modeling and Simulation at the Exascale for Energy and the Environment

The goals of the town hall meetings were

- to gather community input for possible DOE research initiatives in high-performance computing (HPC), computer science, computational science and advanced mathematics, and to examine how these capabilities could be applied to global challenge problems;
- to examine the prospects for dramatically broadening the reach of HPC to new disciplines, including areas such as predictive modeling in biology and ecology, integrative modeling in earth and economics sciences, and bottom-up design for energy and advanced technologies;
- to identify emerging domains of computation and computational science that could have dramatic impacts on economic development, such as agent-based simulation, self-assembly, and self-organization;
- to outline the challenges and opportunities for exascale-capable systems, ultralow-power architectures, and ubiquitous multicore technologies (including software); and
- to identify opportunities for end-to-end investment in new computational science problem areas (including validation and verification).

Each town hall meeting was a day and a half in length and combined invited plenary talks and parallel breakout sessions. Breakout sessions at each meeting were organized and facilitated by a team of leading experts with representation from each of the three laboratories. At all three meetings, breakout sessions were focused on five application areas and four technical areas. The application areas and their central goals were as follows:

- Climate. Improve our understanding of complex biogeochemical (C, N, P, etc.) cycles that underpin global ecosystems functions and control the sustainability of life on Earth.
- Energy. Develop and optimize new pathways for renewable energy production and development of long-term, secure nuclear energy sources through computational nanoscience and physics-based engineering models.
- Biology. Enhance our understanding of the roles and functions of microbial life on Earth, and adapt these capabilities for human use, through bioinformatics and computational biology.
- Socioeconomics. Develop integrated modeling environments for coupling the wealth of observational data and complex models to economic, energy, and resource models that incorporate the human dynamic, enabling large-scale global change analysis.
- Astrophysics. Develop a "cosmic simulator" capability to integrate increasingly complex astrophysical measurements with simulations of the growth and evolution of structure in the universe, linking the known laws of microphysics to the macro world.

The four technical areas address the development and deployment of the tools needed to deliver scientific discovery at the exascale:

- Math and Algorithms. Advancing mathematical and algorithmic foundations to support scientific computing in emerging disciplines such as molecular self-assembly, systems biology, behavior of complex systems, agent-based modeling, and evolutionary and adaptive computing.
- Software. Integrating large, complex, and possibly distributed software systems with components derived from multiple application domains and with distributed data gathering and analysis tools.
- Hardware. Driving innovation at the frontiers of computer architecture and information technology, preparing the way for the ubiquitous adoption of parallel computing, power-efficient systems, and the software and architectures needed for a decade of increased capabilities, and accelerating the development of special-purpose devices with the potential to change the simulation paradigm for certain science disciplines.
- Cyberinfrastructure. Developing tools and methods to protect the distributed information technology infrastructure by ensuring network security, preventing disruption of our communications infrastructure, and defending distributed systems against attacks.

Each breakout session was tasked with addressing eight charge questions:

- What (in broad brush) is feasible or plausible to accomplish in 5–10 years?
- What are the major challenges in the area?
- What is the state of the art in the area?
- How would we accelerate development?
- What are expected outcomes and impact of acceleration or increased investment (i.e., what problems would we aim to solve or events we would cause to occur)?
- What scale of investment would be needed to accomplish the outcome?
- What are the major risks?
- What and who are missing?

This report provides detailed answers to these questions for each breakout topic; the major challenges for each application and technical area are summarized in Table I.2. The consensus at the town hall meetings was that all of these challenges, while formidable, can be overcome if action is taken immediately to accelerate the availability of usable exascale systems. An integrated program of investments in hardware and software research and development (R&D), carried out in partnership with key science communities and accompanied by applied mathematics R&D, can be expected to produce the transformational science and disruptive technologies needed to successfully attack global challenges in energy, the environment, and basic science.

Major challenges in exascale computing

Topic	Major challenges
Climate	• Integrating high-resolution Earth system models with massive assimilation of satellite and other data • Detailed modeling of controlled and modified ecosystems to fit the environmental envelope in which future climate changes will occur • Development of process-scale mechanistic models for biogeochemical, hydroecological, cloud microphysical, and aerosol processes • Rational design and analysis of computer experiments to navigate very large parameter space with very large outputs
Energy	• Combustion: – Predictive simulation capabilities that can accurately model combustion in new high-temperature, low-emission regimes – Robust and reliable ignition and combustion models for next-generation engines and power plants – Multiscale formulations that can exploit the specialized structure of typical combustion applications – Scalable algorithms for multiphysics reacting-flow problems – Improved discretization procedures – Management of software complexity – New tools for data management and information • Nuclear fusion: – Accelerated development of computational tools and techniques to extend the scientific understanding needed to develop predictive models – Advanced computations (in tandem with experiment and theory) to deliver the new science and technology needed to achieve continuous power with higher Q in a device similar to ITER in size and magnetic field • Solar energy: – Exploration of huge parameter spaces – Identification of the best materials and designs for device improvement and optimization, either through direct numerical material-by-design searches or through new understanding of fundamental processes in nanosystems • Nuclear fission: – Identification of fuel cycles that reduce generation of high-level radioactive waste – Reducing the time required for fuel development and qualification – New tools for assessing life-cycle performance, addressing safety concerns, and predicting fuel rod behavior in accidents – Accurate predictions of the behavior of transuranic fuel
Biology	• Model-driven high-throughput experimental data generation • Improving model development by incorporating genome-scale metabolic networks, regulatory networks, signaling and developmental pathways, microbial ecosystems, and complex biogeochemical interactions • New bioinformatics techniques to address the integration of genomics, proteomics, metagenomics, and structural data to screen for novel protein function discovery • Molecular modeling techniques that can address multiscale challenges

Table I.2 Major challenges in exascale computing

Topic	Major challenges
Socioeconomic modeling	• Comprehensive suite of validated models of unprecedented geospatial and temporal detail • Comprehensive error analysis • Leverage of state-of-the-art climate modeling activities (to include economic prediction models under alternative climate regimes, supported by basic research into spatial statistics, modeling of social processes, relevant micro-activity and biosphere coupling issues, and relevant mathematical challenges, such as multiscale modeling) • Novel, robust numerical techniques and HPC approaches to deal with the expected orders-of-magnitude increase in model complexity • Assembly and quality control of extensive data collections
Astrophysics	• Simulation of the formation of large-scale structures to understand the nature of dark energy • Detailed simulations of the formation of galaxies to compare with observational data, requiring dynamic ranges of order 10,000 in space and time • Full-scale simulation with validation quality of the helium shell flash convection zone in stars • Supernova models that include realistic nucleosynthesis studies • Accurate descriptions of binary systems (e.g., a black hole and a neutron star or two neutron stars)
Technical areas	
Math and algorithms	• Systematic approach for quantifying, estimating, and controlling the uncertainty caused by (e.g.) reduced models, uncertain parameters, or discretization error • Robust and reliable optimization techniques that exploit evolving architectures and are easy to use • Appropriate algorithms for novel optimization paradigms that can be implemented only at the exascale (e.g., hierarchical optimization problems over multiple time stages) • Handling of problems with hundreds of thousands of discrete parameters. • AMR for efficient solution of linear and nonlinear systems of partial differential equations (PDEs) • Dynamic load balancing • New data representations, data handling algorithms, efficient implementations of data analysis algorithms on HPC platforms, and representations of analysis results for massive data sets
Software	• Development and formal verification tools integrated with exascale programming models • New fault tolerance paradigms • Application development tools, runtime steering, post-analysis, and visualization • New approaches to handle the entire data life-cycle of exascale simulations (effective formats for storing and managing scientific data, automatically capturing provenance, seamlessly integrating data into scientists' workflow)
Hardware	• Performance per watt • Large-scale integration (packaging 10M to 100M cores with their associated memories and interconnects) • Integrated hardware- and software-based fault management • Integrated programming models
Cyber infrastructure	• Scalable and flexible resources for representing information and reducing information overload • Federated approach to: – Authentication and authorization – Creation and management of virtual organizations • Higher performance tools and techniques for data management and movement • Security products • Tools and techniques for system configuration, verification, troubleshooting, and management • Framework and semantics for integrating information in individual cyber security component systems for situational awareness, anomaly detection, and intrusion response • Data transfer tools that provide dedicated channels for control communication and graded levels of control

Table I.2 Major challenges in exascale computing

1 Climate

How can we improve our understanding of complex biogeochemical cycles that underpin global ecosystem functions and control the sustainability of life on Earth? The urgency of developing a science of global ecosystems is common for several key questions. The U.S. Climate Change Science Program and the Intergovernmental Panel on Climate Change (IPCC) have concluded that climate change will accelerate rapidly during the 21st century unless there are dramatic reductions in greenhouse emissions [Alley et al. 2007]. Assessments by the IPCC and the Military Advisory Board [CNA Corporation 2007] suggest that global warming could have serious implications for the natural and social fabric in many parts of the world. Fortunately, sensible policies to reduce greenhouse emissions could be formulated by using reliable climate forecasts and developing next-generation Earth system models (ESMs), including processes and mechanisms to represent the most likely mitigation strategies that depend on ecological and biological process over land, over oceans, and below the ground.

To develop the necessary forecasts, scientists must address two major challenges. First, how well can we forecast with increased certainty the committed climate change over the next few decades resulting from historical emissions? Second, how well can we forecast longer-term climate change (Figure 1.1), including interactions and feedbacks between all components of the ESM and at spatial scales of relevance to communities? The second question is particularly difficult to answer given our rather limited and rudimentary knowledge of biogeochemical cycles and feedbacks.

Meeting these challenges will require a qualitatively different level of scientific understanding, modeling capability, and computational infrastructure from that represented in the current studies of global average quantities [National Research Council 2001]. Achieving this capability may be impossible without a concentrated effort over the next 5–10 years to develop the detailed process models that are demanded by increased spatial resolution and driven by societal needs. We believe that accelerating scientific development, through targeted attack and application of exascale simulation, is the best way to make a difference in the limited time while key decisions must be made.

We have identified ten key scientific questions that address major unsolved issues and represent targets of opportunity for computationally intensive simulation. Resolution of these questions will yield a much better, more defensible scientific grounding for policy and political discourse. Dramatic advances in the science will be required, however, to provide robust answers and quantitative estimates of uncertainty. After describing the urgent questions, we discuss the scientific advances that are necessary to address these issues. Common to many of the scientific advances are more accurate multiscale models that integrate the physical, chemical, and biological processes in the climate system.

A new type of multidisciplinary research program is urgently required to advance the state of our knowledge, to address the attendant scientific challenges, and to project the future of Earth's environment from the local to the global scale. Such a program is beyond the scope of the current limited, piecemeal approach to climate modeling adopted by several U.S. agencies. Required over the next 5–10 years are focused and well-scoped investments in rapidly developing process-scale

A new type of interdisciplinary research is needed to project the future of the Earth's environment from the local to the global scale.

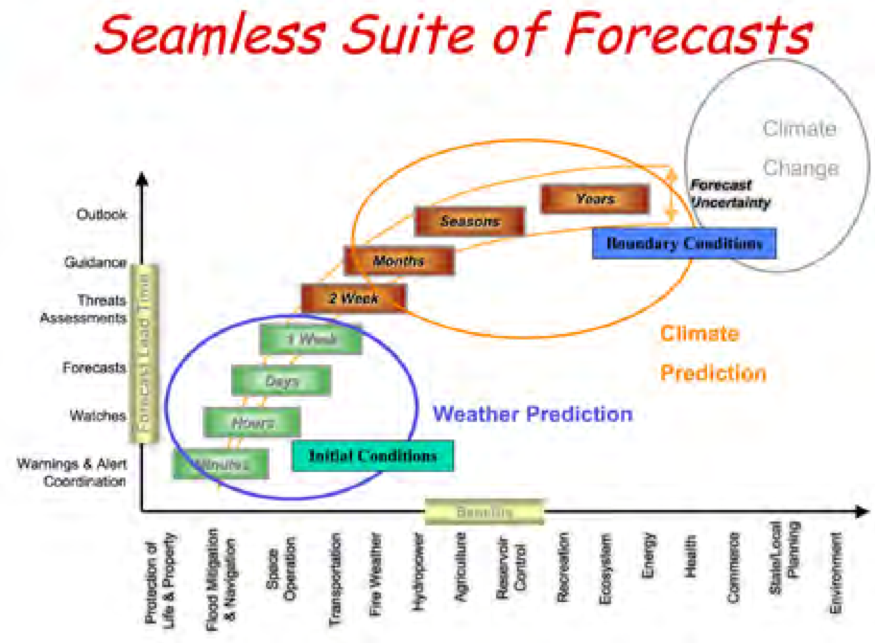


Figure 1.1 Relation between weather prediction and climate change studies. Climate prediction covers time scales of months to decades and has great relevance to threat assessment in hydropower, ecosystems, and energy sectors. (Image courtesy of Kevin Trenberth [NCAR]).

science related to biological and ecological processes of the Earth system; new methodologies and software tools that can integrate these branches of Earth system science with the existing ESMs; and significantly larger computing resources, such as those proposed by the exascale program. The opportunity to positively affect the outcome of the current global change debate is restricted by the current inability of the models to address these regional and local-scale impacts effectively. Significant investment over the next few years can lead to a quantitative impact on this process. Climate science is largely data limited, and the success of the research is contingent on basic measurements and observations necessary to validate, verify, and constrain ESMs. Quantifying the uncertainties in predictions is expected to require a new level of integration between modeling and observational science. New mathematical methods and algorithmic techniques will also be required to address the fundamental challenges of multiscale and multiphysics coupling. Even with exascale computing, approximations and assumptions must be made. Computing power has been and will continue to be a key factor in making these advances possible.

Quantifying the uncertainties in predictions will require a new level of integration between modeling and observational science.

1. Urgent Earth System Questions

Each of the ten questions in this section present significant scientific challenges. In some cases, these challenges can be overcome by basic research into processes, better observation networks, deeper theoretical understanding, and more advanced modeling approaches. In all cases, the path will be more direct and progress accelerated if we can take advantage of petascale and exascale computational power. As the demand is amplified for accurate and reliable predictions of the causes and effects of climate change, the best approach that scientists can take is to continue the development of comprehensive ESMs that can be used as scientific tools to determine the safe concentration levels for CO_2 and other greenhouse gases in the atmosphere. While the scientific community is engaged in expanding our theoretical knowledge and improving our observational depth, we are committed to exploiting our new understanding to address the societal challenges posed by climate change. This initiative will allow the science community to accelerate these efforts.

1.1 Development of Carbon Sequestration Process Models

Decisions on land use and carbon capture and storage technologies will have to be made over the next few decades. Can we develop new and coupled models representing the microbial, ecological, and physiological processes for methods that are currently under consideration in oceans, land, and subsurface?

If we had five years to come up with a sequestration strategy, we would have to use reduced-form models. Major factors are neglected in such models; more detailed models clearly are desired to take account of carbon allocation under large perturbations, of plant mortality, of species migration, and of change rates. The ability to do detailed and operational forecasting of the carbon cycle and climate in the 20- to 50-year range would put sequestration strategies on firm scientific ground and be a valuable tool in helping society adapt to decadal climate change and seasonal transients.

The predictability of short-term carbon–climate models must be rigorously assessed through evaluation with historical data sets. A significant emphasis of this theme will be to incorporate measurements and observations (Figure 1.2) to develop more mechanistic-based models of the various ecological and biological processes at scales ranging from a single tree or plant to scales of ecological systems. For example, starting with the year 1870 (preindustrial conditions), modeled carbon budgets driven by land use change and increasing atmospheric concentration can be performed with some certainty. Since predictability in the decadal range is expected to be low (based on theoretical results, Figure 1.3), data assimilation techniques for carbon will be required to constrain these hindcast predictions. Because of slow decomposition of frozen soils in high latitudes, carbon storage in soil and litter is greater in this part of the global ecosystem, almost twice as concentrated in the boreal and tundra regions as in temperate regions. With significant changes to the precipitation in high latitudes (up to 20% increase for IPCC scenarios), the possibility of abrupt changes and release of large stored carbon pools needs to be investigated. This effort requires model development and data collection to understand the processes that form the foundation for further model development work and ultimate integration into an ESM. We do not know the sign of the carbon flux signal for many parts of the Earth under climate change scenarios.

1.2 Characterizing and Bounding the Coupled Earth System

A systematic understanding of the mathematics of the climate system has yet to be discovered. Can we develop a theory of internal and forced modes of multiphysics, multiscale components that interact as a coupled system?

The paleo record provides evidence that the carbon cycle is well bounded, especially in the interglacial periods, but there are no asymptotic analyses of the cycle based on the mathematics of carbon–climate models. Clearly needed is a more rigorous systems approach, utilizing control theory for systems of partial differential equations (PDEs). Such a theory will allow model developers to determine whether important factors and processes are missing from current models and to pinpoint model components that contain errors when compared with the historical climate record.

Abrupt climate change, and the potential for rapid shifts from one climate equilibrium state to another, could be understood in the context of this system theory. The ability to develop accurate models that incorporate multiscale phenomena from process studies would be greatly advanced by such a theory.

1.3 Probability of Extreme Weather Events and Significant Shifts in Regional Climates

As climate change accelerates, questions arise regarding the frequency and occurrence of extreme weather shifts in regional climate patterns. How can the climate models be adapted to meet these challenges?

The study of extreme events using ESMs is just beginning, but this is just the kind of information that affects community needs. Some of the largest impacts of climate change are

Models development is critical: In many cases we do not know even the sign of the carbon flux signal for many parts of the earth.

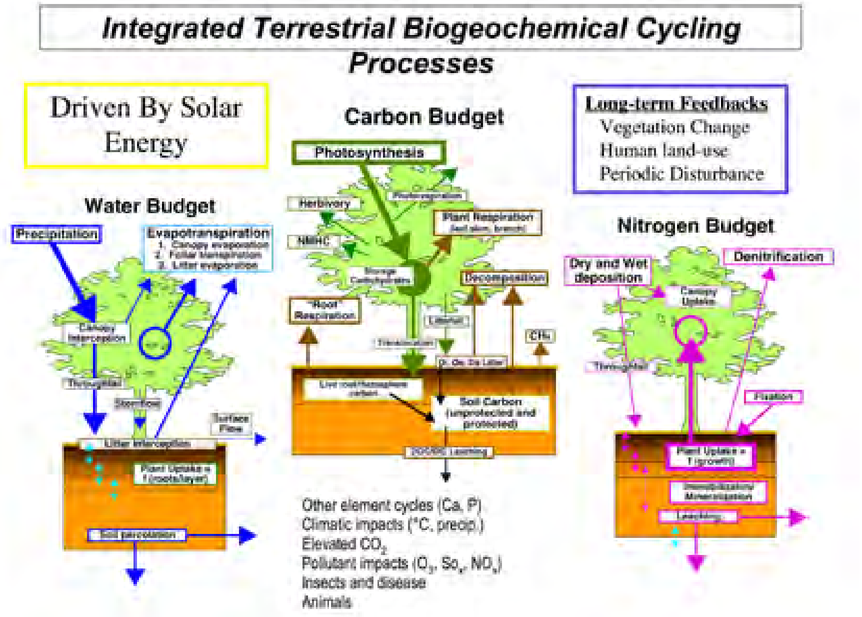


Figure 1.2 The terrestrial biosphere as a consumer and producer of chemicals in the atmosphere. As chemicals cycle through plants, soil, and atmosphere, the long-term feedbacks affect where they are stored. Nutrients such as nitrogen stimulate plant growth.

associated with changes in relatively rare but extreme localized phenomena, such as more intense hurricanes, violent rainstorms, flash floods, and heat waves, as well as low-frequency extremes such as droughts. More temporal and spatial specificity at scales relevant for agriculture, industry, and society is not yet feasible from a computational viewpoint. The ability of existing models to accurately simulate extreme temperature and precipitation events is severely limited by horizontal resolution constraints. Furthermore, since extreme events are rare by definition, adequate statistical characterization of the tails of the distribution of weather events is required to make quantitative statements about changes in their behavior. It is likely that downscaling methods will still be needed to reach the local scale, even with exascale computing power. An important part of this challenge involves engaging stakeholders to iteratively define the interface and the important metadata needed to interpret or interpolate between analysis tools, such as global information system (GIS) collections or GoogleEarth. How do we tell what is important? Can priorities be model-based?

Using models, we should be able to identify key triggering mechanisms for extreme weather and climate events and identify open scientific issues that introduce first-order uncertainties in climate forecasts. Since statistics are important for extreme events, a computational challenge arises in characterizing the tails of the distributions where extreme events occur.

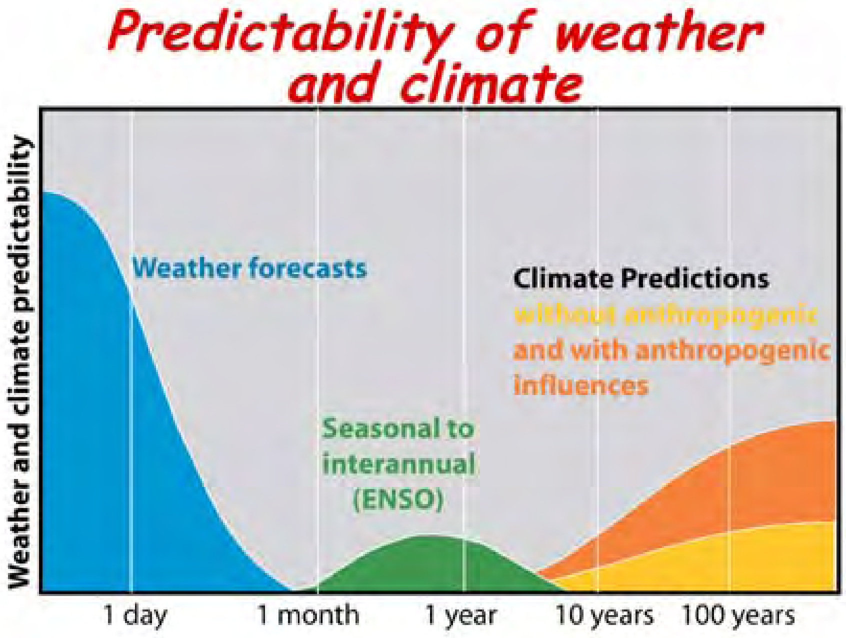


Figure 1.3 Predictability of weather and climate models: high on the short time scales and in the long, asymptotic scales. Since many of the questions to be answered are targeted to the 20–50 year range, the ability of models to provide reliable forecasts will be challenged. Image courtesy of Kevin Trenberth (NCAR).

With petascale computing, horizontal resolution can be increased to the 10- to 25-km scale, permitting reasonable simulation of tropical cyclones. Moderate increases in ensemble size, from the current state of the art of 10 realizations to about 50 realizations, should also be possible. Exascale computing will permit a combination of further resolution increases to better resolve individual storms and increased ensemble size to better capture extreme value statistics. This will offer a great advance in characterizing the uncertainty of climate models and provide the impacts community with reliable expectations of models.

1.4 Sustainability of the Tropical Rain Forest

Precipitation in the tropics is a leading-order factor governing the carbon cycle. What are the magnitude and stability of the carbon–climate feedback for tropical ecosystems?

The Amazon rain forest plays a pivotal role in the climate and, in particular, the carbon cycle. If this ecosystem were to collapse, a large amount of carbon from decaying plants would be released into the atmosphere Since climate change simulations indicate that precipitation will decrease in the tropics under warming scenarios and that less snowmelt will feed the Amazon River basin, a drier Amazon could represent a positive feedback for global warming. Attempts to forecast this situation highlight the possibility of large changes in the next 50 years.

New methods and models are urgently needed, with increased details and significantly more species diversity of plant and microbial life. Also needed are mechanistic process-based models of below- and above-ground ecology.

1.5 Stability of the Polar Caps and Greenland and Antarctic Ice Sheets

Melting or breakup of either the Greenland ice sheet or significant portions of the Antarctic ice sheet could cause a sea level rise of 6 m. What is the likelihood of this happening and on what time scale?

The past 15 years has seen an unanticipated acceleration of ice flow into the ocean from individual catchments of the Greenland and Antarctic ice sheets. If these accelerations are sustained, even larger portions of the polar ice sheets could become vulnerable to mass wasting, with potentially grave consequences for society. Sea level rise is likely the greatest uncertainty in evaluating climate change impacts.

Currently we cannot exclude the possibility of a >1 m cumulative sea level rise over the next century because of partial collapse of the major ice sheets, in addition to the 0.5-m rise expected due to thermal expansion. This problem is coupled not only to the climate system but also to energy and population security. Densely populated coastal areas would be severely impacted. Critical infrastructure—including oil refineries (80% of U.S. refining capacity is at ≤ 1.5 m above sea level), nuclear power stations, ports, and industrial facilities—is often concentrated around coastlines. Coastal biomes, many of which are implicitly part of coastal infrastructure designs, may be severely impacted worldwide.

Exascale simulation is a key tool in predicting the likely course of ice sheet dynamics. Previously unanticipated complexity in ice sheet dynamics is emerging, and new observational techniques are providing a wealth of data on past and present controls on ice sheet change. A new dynamic ice sheet model must incorporate new processes, including ice-stream flow regime change, production and transport of basal fluids, surface-melt to bed lubrication, fracturing, grounding line physics, and ice-shelf/ocean interactions. Such a model must represent scales appropriate for slow creep deformation and fabric formation within the vast ice sheet interior, fast plug flow and water redistribution beneath ice streams, and their tributaries, and flow acceleration and divergence into an ice shelf. Successive development will couple this ice sheet simulator to models of the dynamic earth, atmosphere, and ocean for predictions of future sea-level change. Systematic methods of constraining the model against the available datasets using inverse modeling have very high computational demands but will help yield robust results.

Exascale simulation is a key tool in predicting the course of sheet ice dynamics.

1.6 Release of Methane Hydrates

As warming occurs in the oceans and on land, frozen deposits of methane will be released into the atmosphere. What is the potential in the next 20–50 years for a sudden increase in warming as a result of melting of Arctic and ocean-shelf deposits of methane hydrates?

Historic records need to be further quantified, and new observations are needed to quantify the potential for a sudden release and positive feedback. High-latitude peatland regions are rapidly warming, and as permafrost melts, the shift from anaerobic to aerobic conditions will need to be part of ESM land surface schemes. Ocean-shelf methane hydrate deposits similarly will release methane as a threshold temperature is exceeded. Models of methane hydrate deposits need to be developed and coupled to both the land surface process models and deep ocean circulation models.

1.7 Sustainability of Sea Life

The increasing concentration of atmospheric CO_2 is changing the pH of the ocean. What level of change will trigger coral reef collapse, impacts on fisheries, megafauna changes, and a change in the ability of ocean organisms to take up CO_2?

The ocean removes a large percentage of CO_2 from the atmosphere. The increased levels of CO_2 in the ocean reduce the carbonate ions available for producing calcium carbonate shells. The result is that the skeletal growth rates of corals and some plankton can slow, and in extreme cases some shells may even dissolve. Research is needed to understand

Competition over water resources threatens security and political stability in many areas of the world: New models are needed for predicting river flow and underground water.

the ocean's role as a "sink" for CO_2 and to determine the potential impacts on the ocean food web.

1.8 Sustainability and Agricultural Ecosystems

The shift from agricultural or forest production to production of crops for biofuel could be a significant change in land use patterns in 10–20 years. How might this change the climate's hydrologic cycle and affect the potential for carbon sequestration of the biosphere?

Biofuels offer an attractive alternative fuel source that reduces the net emissions of CO_2 into the atmosphere and enhances our independence from declining and potentially unstable sources of petroleum. Development of other fuel sources also looks promising. The benefits and drawbacks need to be considered, however, in the context of a carbon sequestration strategy, pollution control, and the hydrological cycle. To this end detailed models of agricultural ecosystems are needed at the level of each crop and associated land use (see Figure 1.4). Existing models are fairly simple and parametric [Ma, Shaffer, and Ahuja 2001]. A significant investment and larger computational resources are urgently needed to advance the development of these models and their integration into an ESM.

1.9 Changes in Precipitation Patterns and Hydrology

Regional scale shifts in climate patterns could stress surface and groundwater resources and lead to a disruption of current levels of agriculture production and overall economic sustainability. What is the extent of these changes, where do they occur, how often, and what do we need in the ESM to predict these changes with some certainty?

The intelligence community is calling climate change a serious threat to global security. Competition over water resources under stress and regional scale shifts in precipitation patterns could further affect security and political stability in many areas of the world. An increase in confidence in ESM predictions at these scales for precipitation and hydrology is essential to effectively address these issues over the next couple of decades. Model development focused on methodologies for dealing with the stochastic nature of precipitation is urgently needed, as is development of more hydrological basin-scale based approaches for predicting river flow and underground water resources. New approaches with remotely sensed GRACE (Gravity Recovery and Climate Experiment) water levels and data assimilation techniques will reduce these groundwater model uncertainties.

1.10 Dynamical Linking of Socioeconomic and Climate Responses

At present, there exist one-way and loosely coupled flows of information between physically based ESM and socioeconomic model interfaces. How can we ensure two-way dynamic feedback between these models?

Current models are moving from prescribed emissions scenarios to dynamic emission scenarios, and ESMs are beginning to include dynamic feedback from impacts models [Foster 2007]. Bringing the needed feedbacks into a coupled ESM requires a methodology for incorporating social response as a function of climate into the emissions scenarios, incorporation of detailed local-to-regional socioeconomic phenomena, and decision-game theory mapped into these concepts.

2. State of the Art

The science surrounding the biogeochemical coupling of climate has become central to answering these questions as we learn more about how the coupled carbon cycle has changed in the fossil record, how it is changing now, and how it might change in response to global climate change. Addressing the science issues will require new observations and methods of analysis, new theoretical understanding of the carbon cycle, and new models of the Earth system that include the interactions between human society and the environment. These models play pivotal roles in interpreting the paleoclimate records, in synthesizing and integrating measurements to study the current carbon cycle, and in projecting the future responses of human society and the natural world to evolving climate regimes.

One of the most promising pathways to improving our understanding has been to develop models that represent the complexity of interactions in the Earth system as accurately as possible. Over the past 30 years, these models have advanced considerably in spatial and temporal resolution and in the representation of key processes. However, forecasts of environmental and societal responses to climate change remain highly uncertain. The principal challenges are quantifying the sources of uncertainty, reducing the level of uncertainty at all scales using observations and fundamental theory, and understanding the natural and anthropogenic feedbacks in the climate system. New, multidisciplinary teams of physical, biological, and social scientists could accelerate progress on these challenges with transformational levels of computing.

The carbon cycle has been characterized by using observations from ships, land surface sites, and aircraft. The amounts of CO_2 in and exchanges of CO_2 among the atmosphere, ocean, and land have been estimated to the first order. Representations of the carbon cycle have been introduced into a first generation of ESMs. In contrast to earlier atmosphere-ocean general circulation models (AOGCMs), ESMs can simulate the coupled physical, chemical, and biogeochemical state of the Earth system. Modern AOGCMs operate on terascale systems, realistically reproduce the historical record of global warming, and consistently attribute this warming to human-induced changes in atmospheric chemistry.

One method of assessing our state of understanding is to compare the process- and system-level simulations from multiple ESMs for a single scenario for anthropogenic emissions in hindcast and forecast modes for integration periods ranging from seasons to centuries. Recent studies indicate that the simulated carbon cycle interacts with climate change to increase, not decrease, the uncertainty in these forecasts. This uncertainty is caused by many factors, but one of the most important is the large range of projections for tropical precipitation. This illustrates that better understanding of biogeochemical cycles is, to a large degree, contingent on better understanding and simulation of the physical climate. Systematic error reduction of the physical climate system needs to progress concurrent with the advancement of ESMs with biogeochemical processes and ultimately socioeconomic/energy and emissions feedbacks. In addition, the simulated carbon cycle tends to amplify global warming, although this amplification is also quite uncertain. The feedback is caused by changes in the terrestrial carbon cycle that are difficult to test empirically with our limited observational network and limited process models. The feedbacks could become important and could therefore confound efforts to mitigate climate change in the latter part of the 21st century.

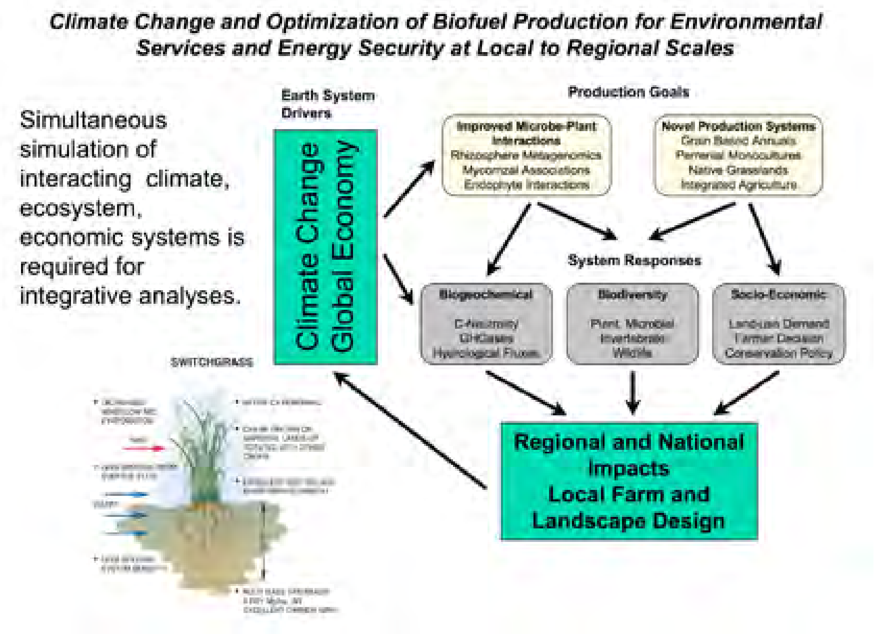


Figure 1.4 Biofuel production will entail land use changes that interact with the climate system. National, regional, and local impacts could be modeled in an exascale Earth system model.

Human society has been measurably perturbing the natural carbon cycle since the mid-18th century. Thanks to comprehensive data on the production and use of fossil fuels, we can quantify the emission of CO_2 from these fuels and its disposition throughout the climate system. It is unclear, however, whether we have socioeconomic models capable of hindcasting or predicting emissions of CO_2

Exascale computing will enable scientists to reduce the uncertainty in models of natural and anthropogenic feedbacks in the climate system.

with sufficient accuracy for policy formation. This lack of certainty arises especially because these models can be evaluated only by using historical data, whereas the economic transformations required to mitigate climate change are without historical precedent.

In summary, significant advances in understanding biogeochemical cycles will follow from

- integration of models and observations of the carbon cycle,
- process-level modeling of biogeochemical cycles across space and time scales,
- accurate models of the coupled physical and biogeochemical system, and
- robust economic models hindcasting and predicting climate-changing pollutants

Attaining these new capabilities requires new approaches that extend across the traditional disciplines of geophysics, biology, and ecology. Major advances are needed in observational, theoretical, and computational studies of our environment.

3. Major Challenges

Three major technical challenges face scientists in understanding biogeochemical cycles.

3.1 Integration of Models and Observations of the Carbon Cycle

Meteorological and oceanic analyses have become an important tool for studying the mean state and variability of the current physical climate. Such analyses are constructed by using a model that is adjusted by incorporating observations during its integration. These analyses have proved particularly useful for understanding the relationship between observations and the underlying dynamics of the climate system. It would be especially valuable to have a comparable analysis of biogeochemical cycles that could relate local and global biogeochemical processes.

No extant analyses encompass the physical, chemical, and biogeochemical processes in the climate system. Development of these analyses will require significant investment in assimilation systems for chemical and biogeochemical observations from *in situ* and satellite platforms. Also required will be considerably more advanced models to understand the error characteristics of the analysis system.

3.2 Process-Level Modeling of Biogeochemical Cycles

Simulation of biogeochemical cycles requires detailed understanding of terrestrial and oceanic ecosystems; the exchange of organic and inorganic carbon compounds with other parts of the climate system; and the fluxes of energy, water, and chemical compounds (e.g., nutrients) that affect these ecosystems. The critical nutrient cycles for ocean and land ecosystems span time scales ranging from a few days (e.g., nitrogen) to over 1000 years (e.g., iron). Modeling over these large time scales to fully evaluate the couplings between biogeochemical cycles and ecology will be a significant computational challenge. The spatial heterogeneity in the biosphere is a fundamental issue overlying much of this science. New models are needed to develop sensible volume/area/mass-averaged and mass-conserving idealizations that preserve the heterogeneity of the process and still allow for a degree of conceptualization. Other major challenges are the sophistication of the ecological representations, the effects of high-frequency spatial and temporal variability on the carbon cycle (e.g., fronts and eddies), and the behavior of the biogeochemical cycles in coastal zones [Doney 2004].

The ecosystem representations tend to be formulated as paradigms of ecological functions. The field certainly needs more mechanistic models of these ecosystems constructed at the level of individual organisms. It also needs much more detailed understanding of the nutrient networks and how these networks affect the carbon cycle. The effects of sharp gradients or rapid changes in the physical environment of the components of the carbon cycle are not well understood. With the advent of ultrahigh-resolution ESMs over the next decade, scientists should be able to probe the effects of rapid variability on scales much smaller than the mesoscale.

Coupling of biogeochemical cycles with ocean and land ecosystems requires simulation over time sales from a few days to thousands of years.

Moreover, the biogeochemistry in coastal zones has not been adequately studied. These regions have been challenging to simulate in global models with insufficient resolution to resolve the coastal regions, the discharges of river sediments into the regions, and other related features.

3.3 Accurate Models of the Coupled Physical and Biogeochemical System

Global models of the Earth system are irreplaceable tools for studying the past, present, and future climate. The accuracy of these models can, for some processes, be determined through comparisons with fundamental theory, with observations, or with benchmark computational models. For many processes—for example, the formation and evolution of clouds and convection—no practical fundamental theory exists. These processes are represented in AOGCMs and ESMs by using simplified statistical representations, or parameterizations. There is also no mathematical theory for the derivation of parameterizations from either observations or benchmark computational models. As a result, the parameterizations in ESMs represent a primary source of uncertainty, both in the reliability of the models as predictive tools and in the fidelity of models to the actual processes in nature. This uncertainty is manifest in the uncertainties regarding the sign of cloud feedbacks on climate change, the sign of convective feedbacks on water vapor, and so forth.

While it is relatively easy to evaluate the simulations produced by using parameterizations from observations and benchmark calculations, it has proved extremely difficult to determine how to improve the parameterizations based on these evaluations. It has also proved difficult to quantify accuracy—in a basic sense, it is not clear what level of accuracy is attainable. It is well known that weather cannot be predicted accurately beyond roughly one week because of the fundamental sensitivity of fluid evolution to the initial conditions of the fluid. However, there is no analogous theory for seasonal, internannual, decadal, or centennial prediction.

At a minimum, enhanced computing capability should make it possible to replace parameterizations selectively with computationally intensive representations at the limit of our present theoretical understanding. For example, with exascale computing it may be possible to replace conventional parameterizations of the carbon cycle (over limited domains) with mechanistic models that represent individual organisms. It should also become possible to replace conventional cloud parameterizations with models of cloud formation based on the fundamental physics of condensation.

Ocean models will make better use of the placement of grid points through unstructured and adaptive mesh technologies that allow for eddy-resolving simulations with dynamic coasts, sea-level rise, detailed boundary currents, and refinement of critical areas of the bathymetry such as sills and overflows. Details of tidal mixing—as tides move over ice melts and enhance melting—will couple with sea and land ice sheet models for accurate prediction of sea-level rise. At the petascale, we will simulate for centuries; at the exascale, the millennial time scales of the deep circulation will be simulated. A seamless suite of climate prediction capability would be a potential aim of the ESM in the future.

4. Feasible Objectives over the Next Decade

Despite the significant challenges outlined above, we are confident that significant progress can be made in biogeochemical simulation within the next 10 years.

4.1 Integrated Models and Measurements of Biogeochemical Cycles

Integration of models and observations of the Earth system appears feasible in the next 5–10 years. The integration should include new measurements of the carbon cycle from planned deployments of automated oceansondes and aerosondes and from new satellites such as the Orbiting Carbon Observatory and Earthcare. These new observational data streams will give total column CO_2 measure-

Global models of the Earth system are irreplaceable tools for studying past, present, and future climate.

A hierarchy of models is needed to represent the diversity and heterogeneity of ecological processes in agricultural systems.

ments. They will enable studies critical for detection and attribution of changes in the carbon cycle, such as the characterization of the natural variability in the coupled carbon cycle, the response of biogeochemical sources and sinks to natural variability in physical climate, and the ways in which natural disturbances such as fires and volcanoes perturb biogeochemical cycles.

The coupling of ocean pH with atmospheric CO_2 will allow a closer examination of the ability of sea life to adapt to and tolerate climate change. The complexity of the ocean ecosystem suggests that the carbon cycle is only the first step in coupling the terrestrial biosphere with climate. For example, isoprene emissions from Pacific Ocean algae appear to have an effect on cloud formation. Models of secondary organic aerosols (SOAs), when compared with the best field measurements, underestimate SOAs by a factor of 10. The treatment of cloud aerosol interactions, and particulates in general, will be important for predicting radiation changes as well as nutrient cycles for land and ocean ecosystems. These treatments require development of much better microphysical models of multicomponent aerosols for the full multidecade range of particle sizes observed in the atmosphere. Physically based and computationally demanding models based on the aerosol general dynamic equation should be reconsidered. A comprehensive methodology for generating SOAs from the potentially numerous organic compounds in the atmosphere from anthropogenic and biogenic emissions should be developed. This method will help in delineating the differences between the various degrees of biomass in a burning plume, a cause for much uncertainty for calculating radiative forcing in the current models. Cloud-resolving models at very high spatial resolution will need cloud condensation nuclei and droplet activation models that go beyond the current parametric representation and that can account for multicomponent aerosols with surfactants and with inert and hydrophilic particle nuclei. The impacts of biomass burning on air quality give urgency to addressing the scientific challenge of understanding these processes and ensuring that the model is doing the right thing for the right reasons.

4.2 Development of Next-Generation Ecological Models

Ecological models representing the diversity of plant life are under development [Stich 2003]. Most of these models are highly parametric, are primarily based on individual data sets, and tend to be site specific. The recent generation of dynamic vegetation models has started taking a more holistic approach to representing this diversity and heterogeneity by adopting macroscale aggregation based on plant functional types. However, agricultural ecosystems either are not present or find limited roles in these models. Since agricultural ecosystems are among the largest terrestrial ecosystems, better representation of these systems in terms of individual crops and climate zones is needed.

One possible solution is to build a hierarchy of models that can represent the diversity and heterogeneity of the ecological processes found in agricultural systems. Complexity could range from an agricultural crop monoculture to a diverse native prairie, and from these models one could develop reduced-form models with higher levels of abstraction. These reduced-form models could be functionally similar to the current generation of dynamic vegetation models (DVMs) with capability to both affect and respond to the dynamics of the more detailed models. Individual-based or agent-based modeling approaches could be targeted for developing these detailed models. Development of mechanistic process-based models would be needed for below-ground soil and microbe processes, in addition to the physiology of and competition among plant functional groups. Approaches such as genomic typing that are under consideration for representing microbial life would be evaluated and targeted for further development. If implemented, such a modeling approach would enhance our ability to plan for mitigation strategies such as carbon sequestration that involve the biosphere and land processes.

4.3 Better Theory for and Quantification of Uncertainty

Formulating a firm theoretical foundation for uncertainty quantification will require major new approaches to error attribution and new developments in the mathematical theory for complex model systems. The motivation for realizing such a foundation follows from the challenge to develop demonstrably more accurate models. If we understand the sources of uncertainty, we may be able to make models so good in particular areas that we are at the limits of what we can learn from observation. Conversely, where models are uncertain, we may be able to suggest observations or experiments that would significantly add to our knowledge of the climate system.

The propagation of uncertainty through a coupled model is particularly challenging because nonlinear and non-normal interactions can amplify the forced response of a system. New systematic theories about multiscale, multiphysics couplings are needed to better quantify relationships. Such theories will be important as ESM results are used to couple with economic and impact models. The science of the coupling and the quantification of uncertainties through coupled systems are necessary groundwork to support complex decisions that will be made over the next few decades.

5. Accelerating Development

Advances in our understanding require improved observations, theory, and computationally based models of the climate system. Here we focus on three areas that will accelerate such development: model development teams, close interaction with applied mathematicians, and high-end simulation. In each case the emphasis is on collaboration with ecologists and biologists, social scientists and economists, and applied mathematicians and systems experts.

5.1 Focused Model Development Teams with Dedicated Resources

At present, climate modelers develop ESMs in a mode suitable for large scientific enterprises. However, assessing the impacts and mitigation of climate change requires ESMs that have been designed from the outset to couple to models for ecology, biology, society, and the economy. The design and exploitation of these models would be greatly enhanced by direct collaborations between the climate community and ecologists, biologists, social scientists, and experts in public policy.

The design of mitigation strategies that adapt to the changing climate and our understanding of those changes requires new combinations of econometrics and game theory. The climate community should collaborate directly with mathematical economists to incorporate and study the behavior of interactive mitigation modules in ESMs.

Topics to be addressed by the development teams include the following:

- High-resolution ESMs with massive assimilation of satellite and other data
- Hierarchical unit testable models with requirements for accuracy in the ESMs
- Detailed modeling of controlled and modified ecosystems to fit the environmental envelope in which future climate changes will occur
- Greater scalability and identification of greater degrees of parallelism
- Process-scale mechanistic models for biogeochemical, ecological, and aerosol processes

5.2 Applied Mathematics and Computer Science Collaborations

The climate community needs to force much closer collaboration with applied mathematicians to address the complexity of climate models. Such collaborations could be useful in theoretical studies of climate models as dynamical systems, new approaches to quantify and reduce uncertainty, new methods to synthesize models and data, and techniques to parameterize very complicated processes.

Two cross-cutting developments are critically needed: (1) new applications of new algorithms in the physical climate model and (2) new software architectures and rapid devel-

Formulating a firm theoretical foundation for uncertainty quantification requires major new approaches to error attribution and in the mathematical theory for complex model system.

opment environments to facilitate code reformulation and refactoring.

5.3 High-End Simulation Capability

Individual ESMs in the next IPCC assessment will produce on the order of a petabyte of output. Data volume of this magnitude is already taxing traditional (and usually serial) analysis techniques and database systems. The new class of ESMs for the environment and society could produce truly prodigious amounts of model data. Extraction of information critical for impact studies (e.g., systematic shifts in precipitation extremes and natural modes of variability) will require new approaches in data archiving, data mining, and feature extraction. Figure 1.5 shows the balance of modeling investments that result from the availability of terascale, petascale, and exascale computers. Factors such as model complexity are traded for resolution, given limited computational resources.

Specific needs as we move toward the exascale include the following:

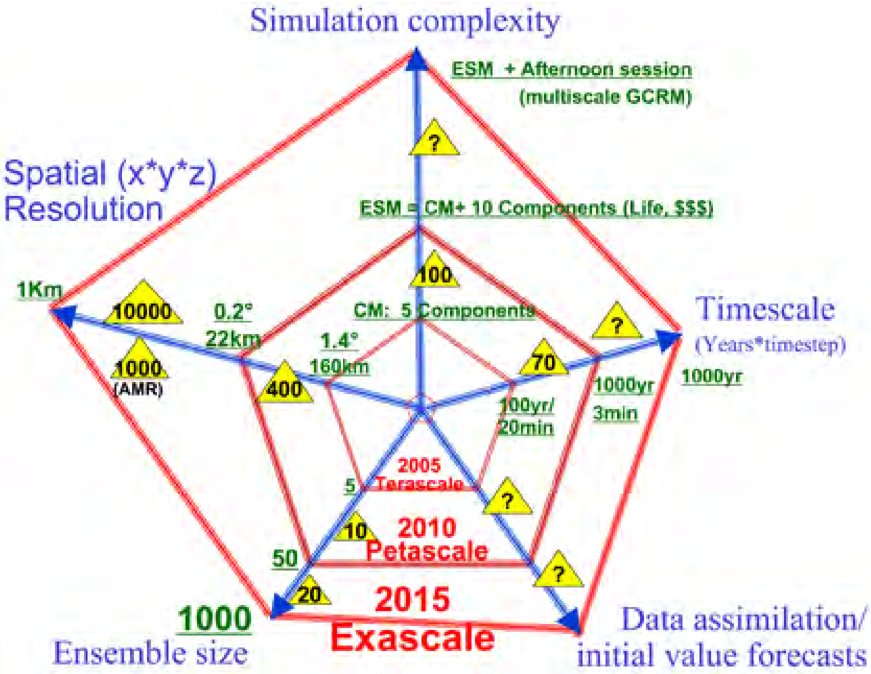


Figure 1.5 Investment of exascale and petascale computational resources in several aspects of a simulation: spatial resolution, simulation complexity, ensemble size, etc. Each red pentagon represents a balanced investment at a compute scale.

- Rational design and analysis of computer experiments to navigate very large parameter space with very large outputs
- Advances in analysis tools with parallelized capabilities, and the ability to explore the full climate solution space using climate experiments based on data mining, objective and repeatable metrics (e.g., Taylor diagrams), and expert pattern recognition and learning capabilities
- Increased computational capacity and capability with dedicated cycles for large climate change studies

Given the urgency of finding answers to key questions, and the added complexity of the modeling enterprise in the exascale environment, the staffing and training of the next generation of Earth and computational scientists are limiting factors. Exascale machines will require a new level of engagement from the algorithmic point of view. In Table 1.3, the shift in algorithmic focus is shown. Multiscale problems are forcing us to consider many new algorithmic approaches. Computer hardware components and the underlying operating systems will also be subject to unprecedented demands. The climate community will have to form even closer collaborations and alliances with hardware vendors and systems developers to address these major issues.

6. Expected Outcomes

The most important outcome of accelerated development and understanding will be quicker answers to the key questions that the climate science community is being asked. At the same time, we will be building momentum for stronger leadership and depth in climate research and creating the ability to produce reliable climate forecasts.

These activities should enable the development of entirely new methods for the attribution of errors in environmental simulation and understanding. These methods will be based on a hierarchy of ESMs running at various ultrahigh resolutions combined with a hierarchy of process-level models of varying complexity. It should become feasible to sep-

Code	Structured grids	Unstructured grids	FFT	Dense linear algebra	Sparse linear algebra	Particles	Monte Carlo	Data assimilation	Agents
CAM	X		X		X	X	X	X	
POP		X			X	X	X		
CLM		X					X	X	X
CICE		X			X	X	X	X	

Table 1.1 The "seven dwarfs" extended for atmosphere, ocean, land, and sea ice models. New developments are highlighted.

arate the uncertainties in the model into terms associated with the frontiers of understanding at the process and observational level. At present, the errors are frequently caused by the relatively simple reductions of these processes and observations incorporated in current AOGCMs. The accelerated development should make it feasible to separate model error into three categories:

- Asymptotic process uncertainties — errors remaining in the limit of the greatest process fidelity (e.g., incorporation of full-complexity cloud models) based on fundamental theory that can be constrained by observations. These uncertainties are also caused by the interactions of errors among process representations.

- Asymptotic scale uncertainties — errors in the mean state and uncertainties in its high-order statistics (e.g., extremes) remaining in the limit of highest possible spatial and temporal resolution. These are due to couplings between the processes, state, and dynamics out of the reach of modern observational systems.

- Asymptotic state uncertainties — errors in the constituents of the system (the mixture of condensed and gaseous species) remaining in the limit of the most detailed possible constituent treatments. Representative treatments include master chemical mechanisms and aerosol modules that track huge ensembles of individual aerosol particles.

It should also become feasible to attribute uncertainties in studies of impacts, adaptation, and mitigation to uncertainties in observation, theory, and computation. For example, this type of attribution will be facilitated by using massive ensembles of ESMs with perturbed physics coupled to models for ecology, biology, and society. The expected outcome will be twofold:

- Detailed and comprehensive information regarding the probabilistic risks of climate change for the environment and society

- Better engagement of stakeholder communities to define information interfaces that inform decision processes

7. Major Risks

The major risks to this enterprise are similar to those that have historically impeded better understanding of climate and biogeochemical cycles.

Lack of sufficient data to constrain key climate processes. One major risk is the lack of sufficient observational data to develop, test, and evaluate new process models. For example, there are essentially no routine observations of the vertical velocity of the atmosphere, and the observations that are available are collected at isolated surface sites. The absence of data on vertical velocities has made it much harder to understand and model the interactions of aerosols and clouds, the dynamics of the boundary layer, and the vertical mixing of chemical compounds.

Slow reduction of model uncertainty due to highly intractable or more complicated climate processes. It may prove quite difficult to reduce the uncertainties in hindcasting, forecasting, or present-day simulation of the coupled climate system. For example, the range of equilibrium climate sensitivity has remained essentially unchanged over the past

Exascale simulation offers the promise of reducing the range of equilibrium climate sensitivity, which has remained essentially unchanged for 30 years.

30 years of model development. While the transformation of computing over this time has enabled the development of much more realistic climate models, estimates of sensitivity from ensembles of AOGCMs have still not converged. Although this lack of convergence is usually attributed to the rapid increase in process complexity, it is caused primarily by uncertainties in basic processes that have been studied intensively for decades, including cloud evolution and convection.

Absence of adequate diagnostic frameworks connecting forcing, response, and initial conditions. The climate community has not developed methods to link errors in climate simulation to errors in process representation, forcing, or initial conditions. The major difficulty is in attributing error in highly nonlinear systems with huge numbers of degrees of freedom (e.g., AOGCMs). Conversely, the absence of a basic theory of error attribution complicates efforts to understand the connections between process-level realism and the fidelity of the entire model system. This risk is also related to the lack of proper diagnostic tools suitable for analysis of complex ESMs.

Overly complicated models. The rapid development of ESMs could produce models that are too complicated or too expensive for adoption by the academic, impacts, or mitigation communities.

References

R. Alley, T. Berntsen, N. L. Bindoff, Z. Chen, A. Chidthaisong, P. Friedlingstein, J. Gregory, G. Hegerl, M. Heimann, B. Hewitson et al. (2007), Climate change 2007: The physical science basis, IPCC, Working Group 1 for the Fourth Assessment, WMO.

CNA Corporation (2007), National security and the threat of climate change. National Research Council (2001), Improving the effectiveness of U.S. climate modeling, National Academies Press.

S. Doney, ed. (2004), Ocean carbon and climate change: An implementation strategy for U.S. ocean carbon research, U.S. Carbon Cycle Science Scientific Steering Group.

Ian Foster, ed. (2007), Exascale global socioeconomic modeling enabling comprehensive climate change impact and response analysis, DOE.

L. Ma, M. J. Shaffer and L. R. Ahuja (2001), Application of RZWQM for soil nitrogen management, pp. 265–301 in M. J. Shaffer, L. Ma and S. Hansen, eds., Modeling Carbon and Nitrogen Dynamics for Soil Management, Lewis Publ., Boca Raton, Florida.

S. Stich, B. Smith, C. I. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. Kaplan, S. Levis, W. Lucht, M. Sykes, K. Thonicke and S. Venevsky (2003), Evaluation of ecosystem dynamics, plant geography, and terrestrial carbon cycling in the LPJ dynamic global vegetation model, Glob. Change Biol. 9, 161–185.

2A Energy: Combustion

Combustion currently provides 85% of U.S. energy needs. Furthermore, because of large infrastructure costs, combustion will continue to be the predominant source of energy for the near and middle term. However, environmental, economic, energy security, and American competetiveness concerns, coupled with the specter of a diminishing supply of oil, are driving a shift toward alternative fuel sources. New combustion systems are needed to utilize these fuels with high efficiency while meeting stringent requirements on emissions. For example, new power plant concepts based on clean coal technologies, such as FutureGen [DOE Office of Fossil Energy 2004], require novel combustion systems that can burn either hydrogen or syngas. For transportation, new engine concepts will be needed to reduce emissions while simultaneously shifting to operate with alternative fuel sources. These new sources—whether oil shale, oil sands, biodiesel, or ethanol—all have physical and chemical properties that vary significantly from traditional fuels.

1. State of the Art

Land-based stationary gas turbines constitute a significant portion of the power generation industry. As part of the overall system design for next-generation power plants, there is considerable interest in developing clean and efficient burners for turbines that can operate with a variety of fuels such as hydrogen, syngas, and ethanol. Concepts based on lean premixed burners have the potential to meet these requirements because of their high thermal efficiency and low emissions of NO_x due to lower post-flame gas temperatures. However, combustion in this regime occurs near the lean flammability limit, making the flame susceptible to local extinction, emissions of unburned fuel, and large-amplitude oscillations in pressure that can result in poor combustion efficiency, toxic emissions, or even mechanical damage to turbo machinery. A fundamental understanding of the dynamics of premixed flame propagation and structure for a variety of different fuels is needed to advance combustion modeling capability and thereby achieve the engineering design goals for new power plants.

Transportation is the second largest consumer of energy in the United States, responsible for 60% of our nation's use of petroleum, an amount equivalent to all of the oil imported into the United States. Virtually all transportation energy today comes from petroleum. The nature of transportation technologies provides opportunities for significant (25–50%) improvements in efficiency through strategic technical investments in both advanced fuels and new low-temperature engine concepts [DOE Office of Basic Energy Sciences 2006]. Such enhanced efficiency will aid in energy conservation and minimize environmental impact. Methods involving low-temperature compression ignition (LTC) engines, such as homogeneous charge compression ignition, offer diesel-like efficiency with the environmental acceptance of current gasoline-fueled cars. These engines operate under high-pressure, low-temperature, dilute, and fuel-lean, oxygen-rich conditions compared to current designs. These new concepts rely on subtle control mechanisms that require a fundamental understanding of combustion science in these relatively uncharted regimes of combustion for their optimal implementation. Although LTC-based designs have shown promise in reducing energy consumption, pollutant emissions, and greenhouse gas emissions, the combination of unex-

To achieve the design goals associated with lean combustion, researchers need a fundamental understanding of the dynamics of premixed flame propagation.

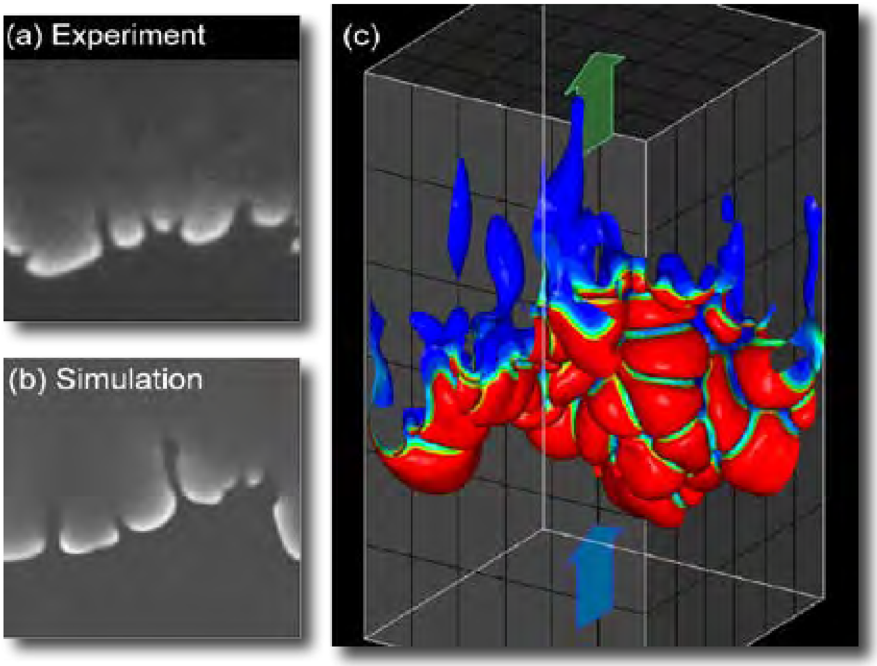


Figure 2A.1 Panel (a) Experimental measurement of the hydroxyl radical OH using planar laser-induced fluoresence and showing local extinction of a premixed hydrogen flame at ultralean conditions. Recent advances in simulation have made it possible to capture this phenomenon in idealized simulations. **Panel (b)** Slice through a three-dimensional simulation that captures the local extinction phenomena. **Panel (c)** View of the flame surface in the simulation.

plored thermodynamic environments and new chemical fuel properties results in complex interactions among multiphase fluid mechanics, thermodynamic properties, and chemical kinetics—the so-called aero-thermo-chemical interactions—that are not understood even at a fundamental level.

These new design concepts for both power generation and transportation will operate in combustion regimes that are not well understood. Effective design of these systems will require new computational tools that provide unprecedented levels of chemical and fluid dynamical fidelity. Current engineering practice is based on relatively simple models for turbulence combined with phenomenological models for the interaction of flames with the underlying turbulent flow. Design computations are often restricted to axisymmetric flows or relatively coarse three-dimensional (3D) models with low-fidelity approximations to the chemical kinetics. Although these approaches have proven extremely effective for traditional combustor design, a dramatic improvement in fidelity will be required to model the next generation of combustion devices. Next-generation, alternative-fuel internal combustion engines and power plants will operate in nonconventional, mixed-mode, turbulent combustion under previously unexplored aero-thermo-chemical regimes. Compared to current devices, combustion in these next-generation devices is likely to be characterized by higher pressures, lower temperatures, higher levels of dilution, and excess air (fuel-lean). In this environment, near-limit combustion sensitivities are amplified—for example, ignition, flammability, and extinction (see Figure 2A.1). These near-limit flame characteristics not only govern efficiency, combustion stability, and emissions but also determine the very existence of combustion in many situations.

Combustion processes in these environments, combined with new physical and chemical fuel properties associated with non-petroleum-based fuels, result in complex interactions that are unknown even at a fundamental level. These unknown parameters place new demands on simulation and severely restrict our ability to predict the behavior of these systems from first principles and our ability to optimize them. There is an urgent demand for high-fidelity simulation approaches that capture these aero-thermo-chemical interactions and, in particular, capture and distinguish the effects of variations in fuel composition. Future combustion technologies will require an unprecedented level of fundamental understanding to develop a new generation of predictive models that can accurately represent the controlling combustion processes for evolving fuel sources.

To achieve these goals, we need a deeper scientific understanding of the combustion processes and advanced modeling technologies to encapsulate that understanding in engineering design codes. Theory and experiment alone cannot address these issues. Theory cannot provide detailed flame structure or the progression of ignition in complex fuels, while experimental diagnostics provide only a limited picture of flame dynamics and ignition limits. For example, advanced nonintrusive laser diagnostics are extreme constraints at high pressure because of constraints on optical access and inherent limitations in the spectroscopy. Numerical simulation, working in concert with theory and experiment, has the potential to address the interplay of fluid mechanics, chemistry, and heat transfer needed to address key combustion design issues.

Recent developments in numerical methodology for combustion simulations in combination with new high-performance parallel computing architectures have enabled dramatic improvements in our ability to simulate reacting flow phenomena. We are now able to simulate realistic laboratory-scale gas-phase turbulent flames with high-fidelity models for chemistry and transport. This type of simulation, performed without incorporating explicit turbulence modeling assumptions, is referred to as direct numerical simulation (DNS); see Figures 2A.2–2A.3. DNS tools are currently being extended to treat multiphase flows and radiative heat transfer.

Scientists are also developing a new generation of engineering combustion design codes based on the concept of large eddy simulation (LES); see Figure 2A.4. This approach provides a more accurate model for turbulent flow than previous approaches and makes it possible to include detailed chemical kinetics models in engineering simulations.

2. Advances in the Next Decade

High-fidelity simulations of combustion phenomena based on DNS and LES approaches require high-resolution simulation of turbulent, reacting flows in three dimensions. Such simulations have benefited enormously from sustained growth in high-performance computing. DNS simulations are one of the key tools needed to study fundamental observations of the fine-scale turbulence-chemistry interactions in combustion; however, they are currently limited by computer power to moderate turbulence intensities and to relatively simple laboratory configurations. LES approaches provide a direct treatment of the large-scale flow dynamics, but physical modeling of the unresolved subgrid scales is required. LES can be applied to both laboratory and practical configurations and has the potential to include high-fidelity representations of the underlying physical processes in engineering design calculations.

Both types of simulations must be time-dependent and include accurate representations of underlying physical processes such as chemical kinetics and transport. Our current ability to perform these types of simulations relies on both high-performance parallel machines and new algorithmic technologies. New approaches based on high-resolution discretization approaches, adaptive mesh refinement, and multiscale formulations have led to significant improvements in the types of problems that can be simulated. Over the next decade, we anticipate that continued improvements in algorithm technology will enable scientists to model new classes of problems with increased physical and geometric complexity.

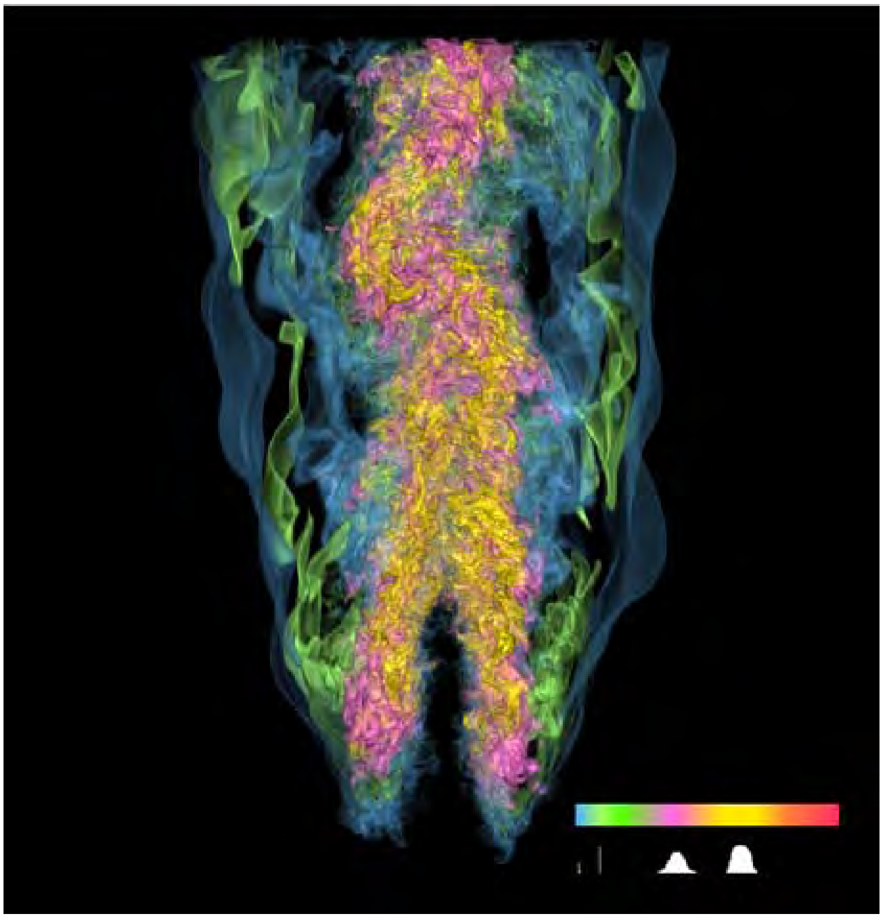


Figure 2A.2 Instantaneous image of the hydroperoxy radical (HO_2), a good marker for ignition, in a lifted turbulent H_2/O_2 jet flame at Re = 11,000 from a DNS simulation. The simulation had 1 billion grids and transported 14 variables requiring 2.5 million CPU hours on the Cray XT3 at ORNL. The stabilization mechanism of this lifted flame is due to autoignition upstream of the high-temperature flame base.

The other key factor that will influence combustion over the next decade is the continued development of new experimental diagnostic procedures. New laser diagnostic procedures are making it possible to probe turbulent flames experimentally in ways that elucidate the turbulent flame structure in much greater detail than has previously been possible. For example, scientists can now measure time-resolved velocity fields and flame locations to capture the interaction of the flame with turbulence. However, new quantitative diagnostic methods are needed at all scales, from individual reactive encounters, to controlled molecular ensembles, to *in situ* combustion chambers. Especially important will be the development of diagnostics applicable at high pressure, where spectral broadening interferes with current optical diagnostic techniques. *In situ* techniques are particularly challenging. At high pressures, spatial gradients are very steep, and new diagnostics need to be developed to resolve the spatial structure of the reaction fronts. Diffraction limitations may require new methods to capture these gradients. Propagating optical beams through high-pressure turbulent media and boundary layers is also extremely challenging. As these measurement techniques are improved, we will have much better characterization of detailed flame and ignition behavior, which will provide the data needed for accurate validation of new simulation capabilities, particularly when operating at pressure.

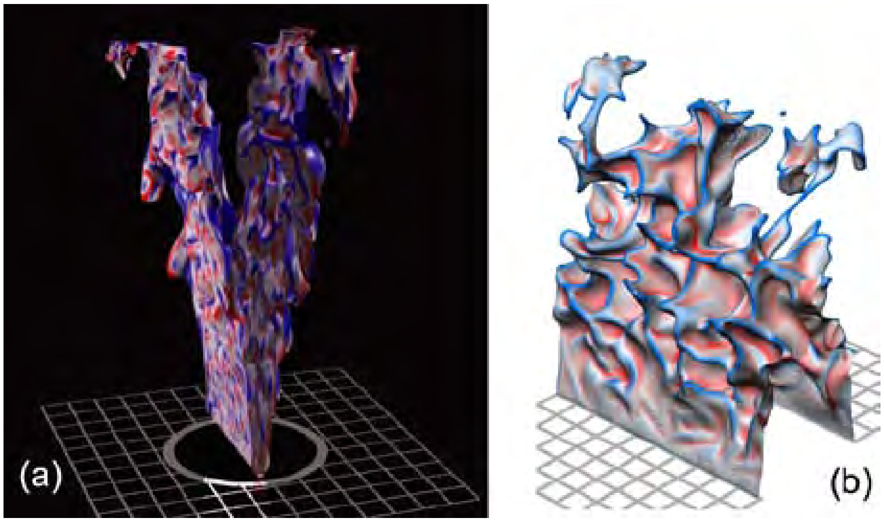


Figure 2A.3 New simulation approaches have made it possible to simulate laboratory-scale premixed flames with detailed chemistry and transport. Two examples are **(a)** a V-flame and **(b)** a slot Bunsen flame. The flame surfaces are colored by curvature to emphasize the wrinkling of the flame surface by turbulence.

3. Major Challenges

The next generation of combustion devices will need to operate at high efficiencies and low emissions with fuels such as hydrogen, syngas, or biofuels. These devices will need to operate in new combustion regimes that are fundamentally different from current engineering practice. Successful development of these new combustors will require new predictive simulation capabilities that can accurately model combustion in these new regimes. Advances in combustion simulation face a number of technological barriers.

One important scientific challenge is to develop robust and reliable ignition and combustion models adapted to the wide range of combustion regimes observed in next-generation engines and power plants, including propagation-controlled combustion, mixing-controlled combustion, kinetically controlled combustion, and combined mixed modes of combustion. Progress here depends on having kinetics and thermodynamic models for realistic fuel compositions at high pressures and low temperatures. Also needed are new strategies for chemical mechanism development and reduction that are both accurate and computationally tractable in multiscale simulations of combustion.

Although simulation methodologies are available for many of the computational problems identified, additional development is required to harness the power of new computer architectures for these problems. For example, multiscale formulations that can exploit the specialized structure of typical combustion applications are needed. Another critical area of research is scalable algorithms for multiphysics reacting-flow problems. Particular issues in this area include the development of scalable solver techniques for variable coefficient and nonlinear implicit systems and the development of improved load-balancing strategies for heterogeneous physics. Substantial increases in capability are also needed, requiring development of improved discretization procedures that not only provide better representations of the basic physical processes but also improve the coupling between these processes.

Another issue facing combustion science is management of software complexity. The simplest simulations are multiphysics algorithms that incorporate fluid mechanics, chemical kinetics, and transport. More complex problems will also require algorithms that treat particles, radiation, and multiphase processes and interfaces. Adaptive mesh and multiscale methodologies are often required to solve problems with the necessary fidelity. Additional issues are raised by the geometric requirements of realistic combustion devices that typically involve complex and, in the case of engines, moving geometries.

A major issue underlying the development of new combustion simulation methodologies is data management. The core simulation methodology discussed here will generate data at enormous rates. This volume of data poses two challenges. First, the data must be archived and software facilities created that allow simulation datasets to be accessed by the larger combustion community. This situation raises issues involving raw storage, software for data extraction, and data security. Second, for the simulations to have impact on design issues, we need to develop tools for extracting knowledge from simulation data and for encapsulating that data in engineering models that can be directly used for design optimization.

Combustion science and the supporting simulations rely on a diverse set of chemical inputs and models, and also produce new data and models. New data informatics approaches are needed to provide for the rapid collaborative development and exchange of chemical mechanisms, thermodynamics data, validated model descriptions, and annotated experimental data that will support the computational studies.

4. Accelerating Development

Developing predictive simulations tools for designing new combustion systems involves the integration of activities across a range of disciplines. Effectively harnessing the compute power of exascale computers to solve key combustion design issues will require a collaboration of computer scientists, applied mathematicians, and combustion scientists with expertise in different aspects of the combustion problem.

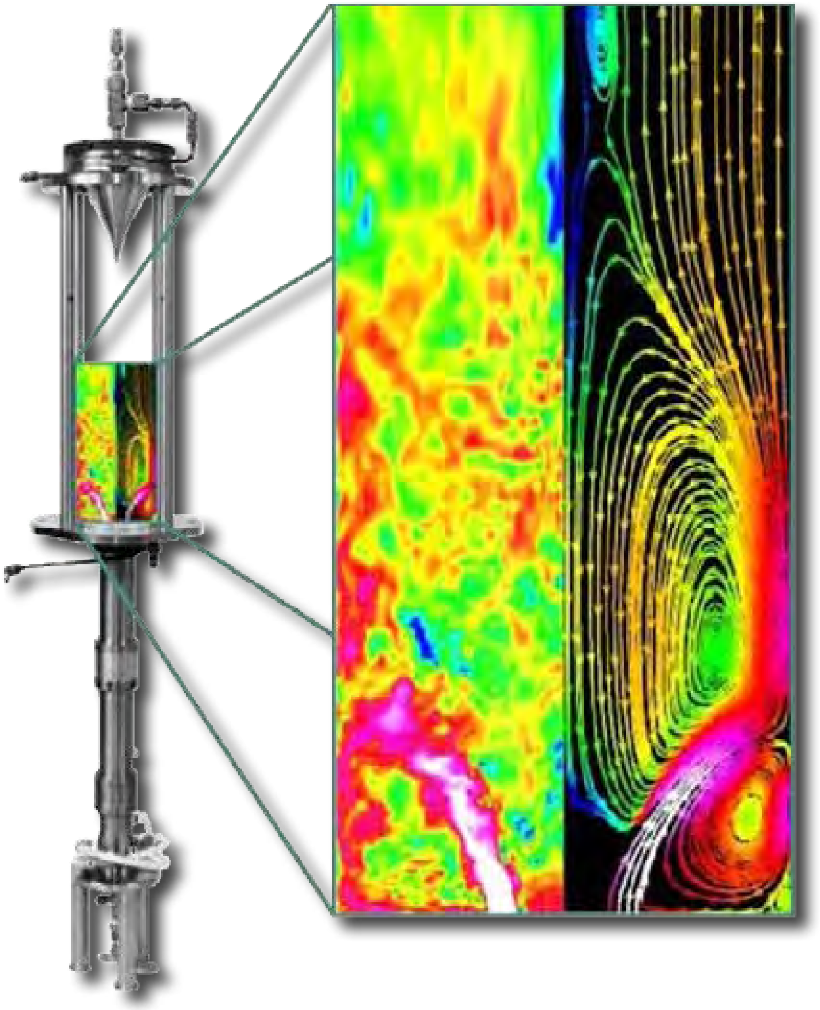


Figure 2A.4 New approaches to LES are enabling high-fidelity simulations of realistic combustion devices such as the high-swirl laboratory-scale annular combustor shown in the left frame. The middle image shows the instantaneous velocity field. The image on the right shows the time-averaged velocity.

Software development for combustion must provide not only the support needed to facilitate implementation of the numerical algorithms, but also the flexibility to integrate different physics modules without loss of performance. Users must be able to incorporate different chemistry and transport packages as required for different applications. At the same time, the overall implementation framework must support the programming models needed to exploit large numbers of processors while insulating the application scientist from the details of a particular architecture.

Applied mathematicians will need to develop new discretization procedures and associated solvers. There will be a pacing need to extend methodologies to more complex physical systems while at the same time developing the new algorithmic approaches needed to effectively solve the resulting systems on computers with large numbers of processors. A particularly challenging issue is to develop high-fidelity discretization approaches for moving geometries that do not sacrifice computational efficiency.

Successful application of the software tools to the design of new combustion systems will also require expertise from the combustion community in a range of topics. In addition to traditional roles in design and analysis of computational experiments, software development will require expertise in experiment, theory, and basic chemistry and transport. As the simulation tools are developed and validated, they will provide drivers for improvements in the fundamental chemistry and transport that determine the flame properties and the formation of pollutants. Establishing these types of linkages will require improved approaches for uncertainty quantification that can be integrated into the validation process. As we begin to explore new combustion regimes, experimentalists will need to work with computational scientists to define appropriate benchmarks for validating the software. Similarly, theorists will need to be involved in fundamental studies to provide the expertise needed to develop predictive engineering models from more fundamental computational flame studies.

5. Expected Outcomes

The development of predictive exascale combustion simulation methodology and the associated supporting software infrastructure will have enormous impact on the development of next-generation combustion systems. With these tools it will be possible to optimize the design of lean, premixed turbine combustors for stationary power generation. We will be able to simulate advanced engine concepts across a range of operating conditions and predict engine efficiency and emissions. Moreover, we will be able to evaluate new biofuels from the perspective of both combustion efficiency and pollutants. The requirement that new systems work in combustion regimes that are not understood at a fundamental level implies that exascale computing will play a deciding role in whether we are able to design these types of systems.

A family of new simulation codes is needed, with capabilities ranging from DNS studies to high-fidelity engineering design codes.

6. Required Investment

Developing the computational tools needed to design next-generation combustion devices is not simply a question of building an exascale simulation capability. Meeting the requirements for designing new systems will necessitate a family of new simulation codes with capabilities ranging from fundamental DNS studies to high-fidelity engineering design codes. In addition, capitalizing on the simulations will require a significant investment in software for managing and analyzing the simulation data.

7. Major Risks

Several components are key to the development of exascale simulation tools for designing the next generation of combustion systems. One is the management and analysis of simulation data. Combustion simulations at the exascale will generate data at an enormous rate. This data must be archived, and tools need to be developed to manipulate the data and extract fundamental knowledge about flame structure that can be used in the engineering design process.

Achieving this goal will also require that several facets of combustion science be brought together. Experimentalists, computational scientists, and chemists will need to work together synergistically to design validation experiments, assess simulation studies and relate uncertainty in chemical behavior to the overall fidelity of simulations. Theorists will need to work closely with experimentalist and computational scientists to develop new modeling paradigms that can be incorporated into new engineering design tools. This new level of collaboration between traditionally disparate activities within the combustion community will be a key component to successfully harnessing the power of exascale computing to solve major problems in combustion.

References

DOE Office of Basic Energy Sciences 2006. Basic Research Needs for Clean and Efficient Combustion of 21st Century Transportation Fuels (http://www.sc.doe.gov/bes/reports/files/CTF_rpt.pdf).

DOE Office of Fossil Energy 2004. FutureGen Integrated Hydrogen, Electric Power Production and Carbon Sequestration Research Initiative (http://www.fossil.energy.gov/programs/powersystems-/futuregen/futuregen_report_march_04.pdf).

2B Energy: Nuclear Fusion

Nuclear fusion, the power source of the sun and other stars, occurs when certain isotopes of the lightest element, hydrogen, combine to make helium. At the very high temperatures (100 million degrees Centigrade) needed for fusion reactions to occur, the electrons of atoms become separated from the nuclei. The resulting ionized gas is known as a plasma. Often referred to as "the fourth state of matter" plasmas comprise over 99% of the visible matter in the universe. They are rich in complex, collective phenomena and are the subject of major areas of research including plasma astrophysics and fusion energy science.

The development of fusion as a secure and reliable energy system that is environmentally and economically sustainable is a formidable scientific and technological challenge facing the world in the 21st century. In addition to being an attractive, long-term source of energy, fusion can have a major impact on climate change challenges, since it does not release CO_2.

Progress achieved in fusion energy to date—10 million watts (MW) of power sustained for 1 second with a gain (the ratio of the fusion power to the external heating power) of order unity—has led to the ITER project, an international burning plasma experiment supported by seven partners (including the United States) that represent over half of the world's population. ITER is designed to use magnetic fields to contain a plasma that will produce 500 MW of heat from fusion reactions for over 400 seconds with gain exceeding 10, thereby demonstrating the scientific and technical feasibility of magnetic fusion energy at a cost of about $10 billion. It is a dramatic step forward in that the fusion fuel will be sustained at high temperature by the fusion reactions themselves. Data from experiments worldwide, supported by advanced computation, indicate that ITER is likely to achieve its design performance. Indeed, temperatures in existing experiments have already exceeded what is needed for ITER.

While many of the technologies used in ITER will be the same as those required in an actual demonstration power plant, further science and technology advances are needed to achieve the demonstration power plant goal of 2500 MW of continuous power with a gain of 25 in a device of similar size and magnetic field. Accordingly, strong R&D programs are needed to harvest the scientific knowledge from ITER and leverage its results. Advanced computations in tandem with experiment and theory are essential. In particular, accelerated development of computational tools and techniques is needed in order to develop predictive models that can prove superior to extrapolations of experimental results. Essential to such development is access to leadership-class computing resources—both petascale and the projected exascale systems—that allow simulations of increasingly complex phenomena with greater physics fidelity.

1. State of the Art

Significant recent progress in particle and fluid simulations of fine-scale turbulence and in large-scale dynamics of magnetically confined plasmas has been enabled by access to terascale supercomputers and by innovative analytic and computational methods for developing reduced descriptions of physics phenomena spanning a huge range in time and space scales. In particular, the plasma science community has developed advanced codes for which computer runtime and problem size scale well with the number of processors on

Strong R&D programs will be needed to harvest the scientfic knowledge from ITER.

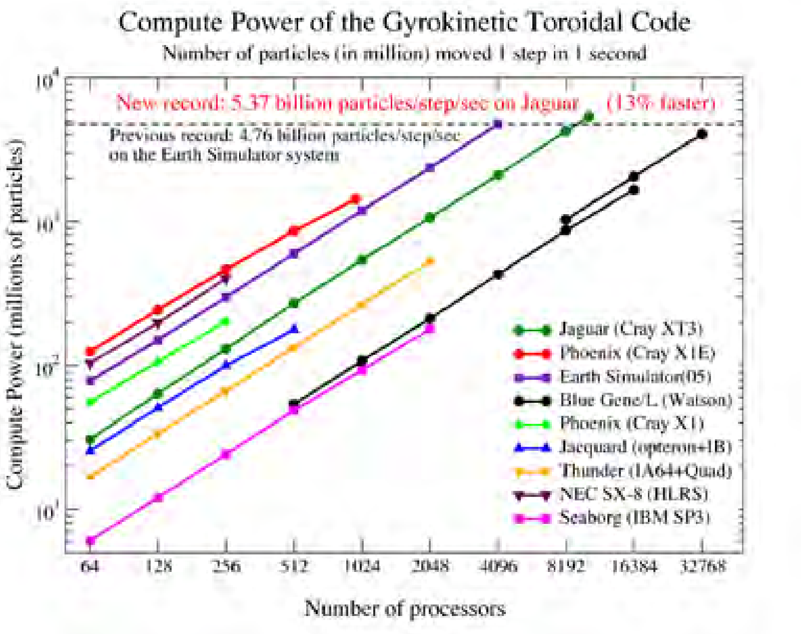


Figure 2B.1 Scaling of fusion turbulence codes on various supercomputers (courtesy of S. Ethier).

massively parallel machines (MPPs). A good example is the effective use of the full power of multi-teraflops MPPs to produce 3D, general geometry, nonlinear particle simulations that have enhanced scientific understanding of the nature of plasma turbulence in fusion-grade high-temperature plasmas confined in a doughnut-shaped (toroidal) configuration. These calculations, which typically involve billions of particles for thousands of timesteps, would not have been possible without access to powerful present-generation MPP platforms together with modern diagnostic and visualization capabilities to help interpret the results.

Realistic modeling of turbulence-driven heat, particles, and momentum losses is of interest for burning plasma laboratory experiments as well as for astrophysics and space and solar physics in natural environments [Tang and Chan, 2005]. Accelerated progress on this critical issue is especially important for ITER because the size and cost of a fusion reactor are determined by the balance between such loss processes and the self-heating rates of the actual fusion reactions [Batchelor et al. 2007]. Computational modeling and simulation are essential in dealing with such challenges because of the huge range of temporal and spatial scales involved. Existing particle-in-cell (PIC) techniques have demonstrated excellent scaling on current terascale leadership-class computers. For example, as illustrated in Figure 2B.1, the Gyrokinetic Toroidal Code (GTC) [Lin et al. 1998; Lin et al. 2000] scales well on virtually all of the current leadership-class facilities worldwide, using Message Passing Interface (MPI) and Open MPI [Ethier 2007; Oliker et al., 2007; Ethier et al. 2005, Oliker et al. 2005, Oliker et al. 2004]. As indicated, GTC scales to over 32,000 processors on Blue Gene (BG) Watson with better than 95% efficiency on the second core and achieves 96% efficiency on over 10,000 dual-core Opteron processors on the Cray XT3.

In common with general PIC approaches, the gyrokinetic (GK) PIC method [Lee 1983; Lee 1987] consists of moving particles along the characteristics of the governing equation—here the 5D GK equation. The equation increases in complexity because the particles are subjected to forces from an externally imposed (equilibrium) magnetic field and from internal electromagnetic fields generated by the charged particles. A grid is used to map the charge density at each point due to the particles in the vicinity. This is called the “scatter” phase of the PIC simulation. The Maxwell equations governing the fields (e.g., the Poisson equation in electrostatic simulations) are then solved for the forces, which are then gathered back to the particles’ positions during the “gather” phase of the simulation. This information is then used for advancing the particles by solving the equations of motion, or “push” phase of the simulation.

The original parallel scheme implemented in GTC consists of a 1D domain decomposition in the toroidal direction and a particle distribution within these domains [Ethier et al. 2005]. Each process is in charge of a domain and a fraction of the number of particles in that domain. Interprocess communications are handled with MPI calls. Particles move from one domain to another while they travel around the torus. Only nearest-neighbor communication in a circular fashion is used to move the particles between domains or processors.

This method scales extremely well to a large number of processors but eventually is dominated by communications as more particles move in and out of the shrinking domains at each timestep. This limit is never reached, however, because the number of domains is actually determined by the long-wavelength physics that we are studying. A toroidal grid with more than 64 or 128 planes, or grid points, introduces waves of shorter wavelengths in the system. These waves are damped by a physical collisionless damping process known as Landau damping. Using a higher toroidal resolution leaves the results unchanged; hence, GTC generally uses 64 planes for production simulations. This approach has limitations, however, because the local grid is replicated on each MPI process within a domain, leading to high memory requirements for simulating large devices.

Initial tests on the dual-core BG/L system were conducted in coprocessor mode, with one BG/L core used for computation and the second dedicated to communication [Ethier et al. 2007]. Additional tests were then conducted in virtual mode node, with both cores participating in both computation and communication. Results showed a per-core efficiency of over 96%. These results indicate that indirect addressing of the gather-scatter PIC algorithm is limited more by memory latency than by memory bandwidth, as the dynamic random access memory (DRAM) bandwidth is shared between the two cores.

Motivated by the strong shift to multicore environments extending well beyond the quad-core level, researchers are focusing on improving current programming frameworks for GTC, such as systematically testing a two-level hybrid programming method. In making full use of the multiple cores on a node, scientists are currently constrained to an MPI process on each core. Since some arrays get replicated on all these processes, the memory limit will be reached for the larger problem sizes of interest. If the hybrid programming method proves successful in initial benchmarking studies on modest multicore machines in Princeton, the plan is to test it on the quad-core leadership-class systems, such as Blue Gene/P at ANL and the Cray XT4 at ORNL.

The excellent scaling of fusion turbulence codes such as GTC on the most advanced leadership-class platforms provides great encouragement for being able to use petascale (and eventually exascale) resources to incorporate the highest physics fidelity. In general, new insights gained from advanced simulations provide great encouragement for being able to include increasingly realistic dynamics to enable deeper physics understanding of plasmas in both natural and laboratory environments.

2. Advances in the Next Decade

A computational initiative called the Fusion Simulation Project, led by DOE's Office of Fusion Energy Sciences with collaborative support from OASCR, is being developed with the primary objective of producing a world-leading predictive integrated plasma simulation capability that is important to ITER and relevant to major current and planned toroidal fusion devices. This initiative will involve the development over the next decade of advanced software designed to use leadership-class computers (at the petascale and beyond) for carrying out unprecedented multiscale physics simulations to provide information vital to delivering a realistic integrated fusion simulation modeling tool. Modules with much improved physics fidelity will enable integrated modeling of fusion plasmas in which the simultaneous interactions of multiple physical processes are treated in a self-consistent manner. The associated comprehensive modeling capability will be developed in close consultation with experimental researchers and validated against experimental data from tokamaks around the world. Since each long-pulse shot in ITER is expected to cost over $1 million, this new capability promises to be a most valuable tool for discharge scenario modeling and for the design of control techniques under burning plasma conditions.

The following are examples of expected advances needed to enable a comprehensive integrated modeling capability.

- Coupling of state-of-the-art codes for the plasma core and the plasma edge region

Delivery of a realistic, integrated modeling tool for multiscale physics simulations will require advanced software capable of running on DOE's leadership-class computers.

- Coupling of state-of-the-art codes for magnetohydrodynamics (MHD) and auxiliary heating of the plasma via radio frequency (RF) waves
- Development of more realistic reduced models based on results obtained from DNS-type major codes that use petascale capabilities
- Development of advanced workflow management needed for code coupling

3. Major Challenges

In order to achieve accelerated scientific discovery in fusion energy science (as well as many other applications domains), new methods must be develoed to effectively utilize the dramatically increased parallel computing power that is expected within the next decade as the number of cores per chip continues to increase.

Examples of outstanding challenges in the fusion energy science application area include the following.

- Efficient scaling of MHD codes beyond terascale levels to enable higher-resolution simulations with associated greater physics fidelity
- Efficient extension of global PIC codes into fully electromagnetic regimes to capture the fine-scale dynamics that not only is relevant to transport but also helps to verify the physics fidelity of MHD codes in the long-mean-free-path regimes appropriate for fusion reactors
- Data management techniques to help develop (and debug) advanced integrated codes
- Innovative data analysis and visualization methods to deal with increasingly huge amounts of data generated in simulations at the petascale and beyond

4. Accelerating Development

Future multicore processors will require new programming techniques, such as a hybrid framework incorporating both MPI and UPC.

As the future multicore processor chips are likely to support coherent shared memory and parallel computers are likely to support only message passing among nodes, it may be necessary to consider algorithms and programming techniques to construct parallel programs where the code that runs on an individual CPU chip uses the multithreaded, shared-memory programming model and uses the message-passing programming model (such as MPI or UPC) to communicate among the CPU chips in a parallel computer. The associated transition is to multicore programming characterized by very low latency but limited bandwidth to main memory.

In order to achieve the desired accelerated progress, several developments must be made:

- Compilers to decompose the code running on a single node into fine-grained computation tasks to utilize the collection of cores on a single chip
- Highly efficient runtime systems to schedule fine-grained tasks to optimize for available parallelisms and to maximize on-chip cache locality to overcome off-chip memory latency and bandwidth constrains
- Hybrid programming frameworks for MPI and UPC by incorporating the compilation and runtime systems with existing MPI and UPC programming environments.

5. Expected Outcomes

If proper investments in research efforts are made, specific state-of-the-art codes for fusion energy science (such as GTC) can be transformed to versions using a hybrid programming framework and then systematically exercised to demonstrate and to test the effectiveness of this proposed paradigm. If such methods for optimally utilizing multicore systems prove effective, then such codes can realize their high potential for achieving new scientific discoveries with exascale computing power.

Other expected outcomes if dedicated efforts are properly supported include the following.

- Clear demonstration of the ability to effectively integrate (or at least couple) advanced codes to deliver new physics insights

- Significant progress in the ability to reliably predict the important edge-localized mode (ELM) behavior at the plasma periphery

- Significant improvement in the physics fidelity of "full-device modeling" of ITER, along with significant progress in achieving reliable predictive capability for ITER

6. Required Investment

The new Fusion Simulation Project will require on the order of $25 million/year over the course of the next 15 years and more. In addition, research progress enabled by ultrascale compute power will demand much greater computer time. For example, a single GTC simulation carried out at present to investigate the long-time evolution of turbulent transport requires around 100,000 cores for 240 hours, or 24 million CPU hours. Since the current version of the leading plasma edge code XGC requires roughly the same amount of time, actual coupled simulations of the core and edge regions could demand approximately 50 millon CPU hours. If additional dynamics (such as the modeling of RF auxiliary heating) are included, then computational resources at the exascale will be essential.

7. Major Risks

Without the dedicated investments described in the preceding sections, the leading fusion codes (especially the MHD codes, with their scaling challenges) run the risk of not being able to effectively utilize the large number of processors at the exascale. The development of effective mathematical algorithms for integration and coupling is very difficult and may be hard to achieve within the next decade. Moreover, if the fusion energy science applications are able to effectively utilize only a small fraction of the cores on a CPU, major efforts will be needed to develop innovative methods for per-processor performance.

8. Benefits

Reliable full-device modeling capabilities in fusion energy sciences will demand computing resources at the petascale range and beyond to address ITER burning-plasma issues. Even more powerful exascale platforms will be needed to meet the future challenges of designing a demonstration fusion reactor. With ITER and leadership-class computing being two of the most prominent missions of the DOE Office of Science, full-device integrated modeling, which can achieve the highest possible physics fidelity, is a worthy exascale-relevant project for producing a world-leading, realistic predictive capability for fusion. This should prove to be of major benefit to U.S. strategic considerations for energy, ecological sustainability, and global security.

Hybrid programming frameworks promise enormous benefits, including greater physics fidelity and more reliable, predictive capability.

References

D. A. Batchelor, M. Beck, A. Becoulet, R. V. Budny, C. S. Chang, P. H. Diamond, J. Q. Dong, G. Y. Fu, A. Fukuyama, T. S. Hahm, D. E. Keyes, Y. Kishimoto, S. Klasky, L. L. Lao, K. Li, Z. Lin, B. Ludaescher, J. Manickam, N. Nakajima, T. Ozeki, N. Podhorszki, W. M. Tang, M. A. Vouk, R. E. Waltz, S. J. Wang, H. R. Wilson, X. Q. Xu, M. Yagi and F. Zonca (2007), Simulation of fusion plasmas: Current status and future direction. *Plasma Sci. Technol.* 9:312–387.

S. Ethier, W. M. Tang and Z. Lin (2005), Gyrokinetic particle-in-cell simulations of plasma microturbulence on advanced computing platforms. *J. Phys. Conf. Series* 16, 1–15.

S. Ethier, W. M. Tang, R. Walkup and L. Oliker (2007), Large scale gyrokinetic particle simulation of microturbulence in magnetically confined fusion plasmas, *IBM J. Res. Dev.*, accepted for publication.

W. W. Lee (1983), Gyrokinetic approach in particle simulation. *Phys. Fluids* 26(2) 556–562.

W. W. Lee (1987), Gyrokinetic particle simulation model. *J. Comput. Phys.* 72, 243–269.

Z. Lin, T. S. Hahm, W. W. Lee, W. M. Tang and R. B. White (1998), Turbulent transport reduction by zonal flows: Massively parallel simulations. *Science* 281, 1835–1837.

Z. Lin, T. S. Hahm, W. W. Lee, W. M. Tang and R. B. White (2000), Gyrokinetic simulations in general geometry and applications to collisional damping of zonal flows. *Phys. Plasmas* 7(5) 1857–1862.

Leonid Oliker, Andrew Canning, Jonathan Carter, John Shalf and Stephane Ethier (2004), Scientific computations on modern parallel vector systems. In *Proc. SC2004*, Pittsburgh, PA..

L. Oliker, A. Canning, J. Carter, C. Iancu, M. Likewski, S. Kamil, J. Shalf, H. Shan, E. Strohmaier, S. Ethier, and T. Goodale (2007), Scientific application performance on candidate petaScale platforms. In *Proc. IPDPS'07*, Long Beach, CA.

L. Oliker, J. Carter, M. Wehner, A. Canning, S. Ethier, B. Govindasamy, A. Mirin, D. Parks, P. Worley, S. Kitawaki, Y. Tsuda (2005), Leading computational methods on scalar and vector HEC platforms. In *Proc*. *SC2005*, Seattle, WA.

W.M. Tang and V. S. Chan (2005), Advances and challenges in computational plasma science. *Plasma Phys*. *Control*. *Fusion* 47(2) R1–R34.

2c Energy: Solar

Solar energy, either in the form of photovoltaic or solar chemical fuel generation, can be the ultimate renewable energy solution for our energy/global warming crisis. Key scientific challenges and research directions that will enable efficient and economical use of the abundant solar resource have been identified (DOE Office of Basic Energy Sciences 2005); the need for new computational and modeling tools to meet the challenges of solar energy research is widely recognized.

1. State of the Art

More than 30 years were needed for the relatively simple thin-film crystal/multicrystal Si solar cell to reach its current efficiency of 24%. In order to develop next-generation solar cells based on new materials and nanoscience fast enough to reduce the global warming crisis, a different paradigm of research is essential. Exascale computation can change the way the research is done—either through a direct numerical material-by-design search or by enabling a better understanding of the fundamental processes in nanosystems that are critical for solar energy applications.

Investigations include finding the right materials for hydrogen storage; identifying the most reliable and efficient catalysts for water dissociation in hydrogen production; determining an inexpensive, environmentally benign, and stable material for efficient solar cell application; understanding the photo-electron process in a nanosystem; and hence helping to design an efficient nanostructure solar cell. In all of these areas, the possible exploratory parameter spaces are huge. This situation on the one hand provides ample opportunity and potential for device improvement, but on the other hand presents a tremendous challenge to find the best material and design.

2. Major Challenges

In computational materials science and nanoscience, three major challenges exist.

- Developing appropriate numerical approximations and models for accurately calculating the corresponding physics properties
- Integrating the diverse models and computational approaches and programs used to calculate different parts and aspects of a complex system, hence enabling the simulation of the whole system and process
- Calculating the large-scale systems (nanosystems containing tens of thousands of atoms) dynamically for a long period of time (nanoseconds or microseconds)

The computational physics and chemistry communities have been addressing the first challenge since the invention of quantum mechanics. Although the many-body Schrödinger's equation is well known and exactly describes almost all phenomena in materials science, the direct accurate solution of that equation is almost impossible. The reason is that a system with N electrons must be described by a many-body wavefunction in N-dimensional space. That makes the needed numerical coefficients scale as N^N. A direct solution of such a problem might be possible only with a quantum computer. The most common approximation is to describe the many-body wavefunction and the Schrödinger's equation with an N single-particle wavefunction. This is exemplified by the currently popular density functional theory (DFT), where a direction calculation scales as N^3, instead of N^N.

Developing next-generation solar cells based on new materials and nanoscience requires a new paradigm exploiting exascale computing.

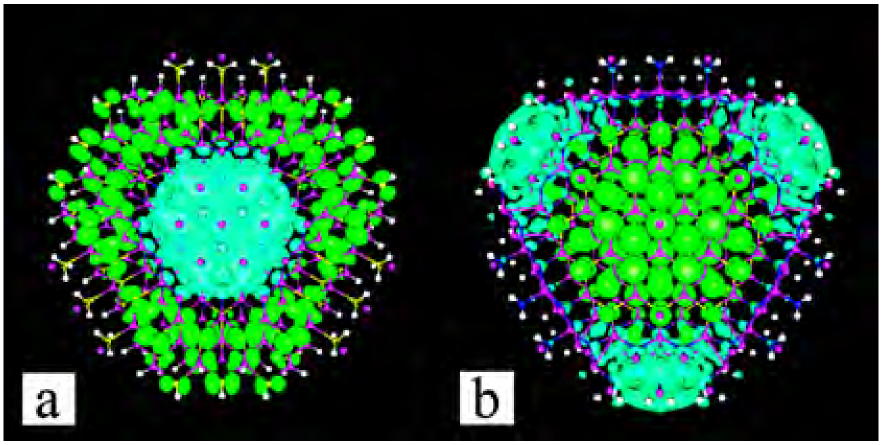


Figure 2c. 1 A core/shell coaxial cable structure can be useful for future solar cell applications by separating the photon-generated electrons from the holes. Here the electron (green) and hole (cyan) states are shown in nano-coaxial wires, with (a) GaN(core)-GaP(shell) and (b) GaP(core)-GaN(shell). Fast and accurate calculations including the electron-hole correlation effects, the carry transports, the carrier cooling, and other dynamical effects are critical for solar cell simulations.

The second challenge focuses on software development and integration. It also presents a computational hardware requirement more on the capacity side than on the capability side. At the least, an integrated but heterogeneous computing platform with different emphases on speed, memory size, input/output (I/O), and communication may be needed, rather than a single homogeneous system. Most experimentally measured physical properties (e.g., solar cell efficiency and a nanosystem synthesis) are results of the combination of different physical processes. At present, we have different methods and codes to calculate each individual property and process. But we lack a flexible tool to integrate these properties and processes or to easily replace one model or algorithm in one part of a calculation by another model. Clearly needed is the development of a common framework and a common community code.

The third challenge calls for both algorithm development and exascale computers. Many of the critical processes in solar energy applications are poorly understood. For example, it is not clear how a nanocontact between metal and semiconductor works. Also unknown is the mode of hole carrier transport in a solar cell using organic materials. How a photon-generated exciton dissociates itself into an electron and hole in a nanosystem is another critical issue that needs to be better understood. Moreover, researchers have little insight into the pathways in water splitting. Understanding these critical phenomena will help scientists tremendously in rationally designing new solar cells and solar chemical cells (see Figure 2c. 1).

3. Advances in the Next Decade

DFT can describe many properties accurately, including atomic structures and binding energies. Thus, it is useful in the search for hydrogen storage materials and catalysts. Although it gives the wrong band gap, with some corrections it still can be used to study optical properties and electron-phonon interactions and hence also for solar cell simulations. Other methods, such as the coupled cluster method in quantum chemistry and the GW method in materials science, can provide more accurate calculations for chemical binding and band structures, respectively.

New codes will also be needed, and these and existing codes must be integrated for maximum usefulness. Specifically envisioned in the next decade is a new community code that is both highly flexible and modular, enabling different research groups to contribute different modules. This community code must be more than a collection of codes as in NWChem, and it should go beyond the modulation, exemplified in the ABINIT project.

Another key factor in advances in the next decade is the availability of exascale computers. Computation can potentially provide a direct way to reveal the secrets of diverse processes involved in nanoscale interactions. Indeed, computation is key because probing some of these processes experimentally is extremely difficult. Because of the $O(N^3)$ scaling for DFT, even with petascale computers one can probably calculate only the electronic structures of a system with 50,000 atoms for a given atomic configuration. If many timesteps are needed to simulate a dynamic process, this direct approach will become unfeasible. Thus, linear-scaling (to the size of the system) approaches become necessary. Fortunately, as a result of the near-sighted feature of quantum mechanical effects, such linear-scaling algorithms based on domain decomposition are possible and effective for the large systems discussed here.

4. Accelerating Development

With linear-scaling algorithms and exascale computing, scientists will be able to simulate a whole nanostructure device.

With linear-scaling algorithms and exascale computing, we should be able to simulate a whole nanostructure device—from photon absorption and exciton generation, to exciton dissociation and carrier collection in a nanosize solar cell. This simulation can be done by following a time-dependent Schrödinger's equation (e.g., time-dependent DFT, TD-DFT) and at the same time doing Newtonian dynamics for the atoms. Although some uncertainties remain about how to do this exactly (for quantum state collapsing), such a simultaneous dynamical simulation for the electrons and atoms will help to reveal the electron-phonon interaction, the nonradiative carrier decay, and cooling, thereby helping scientists to understand carrier transport and collection. Such simulations can also help us to understand the electrocatalysis of water splitting and to figure out the dynamic pathway of the water splitting process.

5. Expected Outcomes

Carrying out simulations for experimental-sized nanosystems is a tremendous computational challenge. In addition to the problem of size scale is the long time scale. The typical carrier dynamics takes tens of nanoseconds, while the typical time step needed for a TD-DFT is usually in the order of 10^{-3} femtoseconds (Fs), which is a thousand times shorter than a time step for atomic molecular dynamics due to the small mass of an electron. Thus, the number of the timesteps is on the order of 10^{10}. Currently, one can do tens of femtosecond simulations for small systems containing about a hundred atoms (e.g., on the Earth Simulator) based on the direct TDDFT formalism. Thus, there remains a gap of about 10^5 in time scale and 10^2 in size scale. Both algorithm development (both in linear scaling and in accelerating the dynamics) and exascale computers are needed to close this gap. But the benefit will be tremendous for understanding the photon-electron process in solar-cell-related applications. The lack of such understanding is the current bottleneck in developing more efficient nanostructure solar cells.

6. Required Investment

Traditionally, code developments have not been supported by either OASCR or DOE's Office of Basic Energy Sciences (BES). Rather, they are often supported within an individual investigator's projects as a side product for studying a specific physical phenomenon. Hence, community code development and integration have been severely limited. Arguably, more recent federal efforts such as the Scientific Discovery through Advanced Computing (SciDAC) projects have attempted to address this limitation, but currently these efforts are either too general on the computer science side or too narrow in specific physics-oriented projects. Further investment is needed to ensure that nanoscience and materials science research are supported at a level needed to solve key problems in solar energy simulation.

Reference

DOE office of Basic Energy Sciences 2005. Basic Research Needs for Solar Energy utilization: Report of the Basic Energy Sciences Workshop on Solar Energy Utilization, April 18-21, 2005. (http://www.sc.doe.gov/bes/reports/files/SEU_rpt.pdf).

2D Energy: Nuclear Fission

Commercial nuclear power plants (NPPs) generate approximately 22% of the electricity produced in the United States. There is growing interest in operating the existing fleet of NPPs beyond their original design lifetimes, in constructing new NPPs, and in developing and deploying advanced nuclear energy systems to meet the rising demand for carbon-free energy; this situation presents significant opportunities for the application of petascale and exascale computing. The new GNEP (Global Nuclear Energy Partnership) Program [DOE Office of Nuclear Energy, Science, and Technology, 2006] seeks to bring about wide-scale use of nuclear energy while at the same time decreasing the risk of nuclear weapons proliferation and effectively addressing the challenge of nuclear waste disposal. GNEP aims to advance the nonproliferation and national security interests of the United States by reinforcing its nonproliferation policies and reducing the spread of enrichment and reprocessing technologies, and eventually eliminating excess civilian plutonium stocks that have accumulated. The stated goals of GNEP are:

- Expand nuclear power to help meet growing energy demand in an environmentally sustainable manner;
- Develop, demonstrate, and deploy advanced technologies for recycling spent nuclear fuel that do not separate plutonium, with the goal over time of ceasing separation of plutonium and eventually eliminating excess stocks of civilian plutonium and drawing down existing stocks of civilian spent fuel;
- Develop, demonstrate, and deploy advanced reactors that consume transuranic elements from recycled spent fuel;
- Establish supply arrangements among nations to provide reliable fuel services worldwide for generating nuclear energy, by providing nuclear fuel and taking back spent fuel for recycling, without spreading enrichment and reprocessing technologies;
- Develop, demonstrate, and deploy advanced, proliferation resistant nuclear power reactors appropriate for the power grids of developing countries and regions;
- In cooperation with the IAEA (International Atomic Energy Agency), develop enhanced nuclear safeguards to effectively and efficiently monitor nuclear materials and facilities, to ensure commercial nuclear energy systems are used only for peaceful purposes;

A recent BES report [DOE Office of Basic Energy Sciences 2006] reviewed the status and basic science challenges, opportunities, and research needs for advanced nuclear energy systems, with specific attention to the role of predictive modeling and simulation (M&S) in addressing the difficulties posed by the radioactive materials and harsh environments found in these systems. The conclusions drawn in this report were similar to those of the town hall meetings:

- Computational M&S offers the opportunity to accelerate nuclear energy development by simulating complex systems to evaluate options and predict performance, thus narrowing the technology path and optimizing testing requirements.
- Today's high-performance computational systems are capable of modeling complete

Nuclear power plants are becoming increasingly attractive because of the possibility of offering carbon-free energy.

reactor systems and related technologies; the availability of exascale systems will enable high-fidelity M&S that can further improve the performance of existing reactors and have a significant positive impact on both the design and the operation of future reactors.

Simulation has the potential for addressing the critical needs of advanced nuclear energy systems by providing the tools necessary for safety assessments, design activities, cost, and risk reduction. One can, for example, imagine virtual prototyping of reactor cores yielding data that leads to more accurate identification of design margins, allows early experimentation with novel design concepts, and ultimately significantly reduces plant certification timelines. In other areas, such as advanced fuel fabrication, atomistic fuel simulations could ultimately make it possible to target a small subset of promising candidate fuel types for further experimentation, greatly reducing the number of experiments to be performed. A simulation-based methodology is within reach with exascale computers.

1. State of the Art

Modeling and simulation (M&S) has always been an integral part of nuclear engineering analysis, safety, and design. Computational analyses have been used to predict detailed quantities that could not be readily measured *in situ*, for example, the aging of structures, power distributions in cores, transient safety behavior, etc. Existing (legacy) tools for material property determination, spent fuel reprocessing, fuel performance, reactor safety and design, and waste forms and storage based on a large experimental database will be insufficient. Experimental testing will be extremely expensive, protracted, and in some cases unfeasible. Furthermore, the existing experimental database is insufficiently documented and often has inadequate precision to support a modern validation process. Complementing or replacing testing with high-fidelity computer simulation will make it possible to collect simulated data that can, in conjunction with a sound experimental validation program, be used to understand fundamental processes that affect facility efficiency, safety, and cost.

Material Property Determination: Basic material properties is the area with the most potential for progress with the greatest return on investment. Properties including nuclear (neutron and gamma reactions), thermophysical (e.g. thermal conductivity, phase diagrams), mechanical (e.g. tensile properties, fracture toughness) and chemical (e. g. corrosion rates) have to be determined under static and dynamic conditions. The fuel process selection is an example that illustrates the possible gain. In the design of a reactor, fuel definition along with the choice of coolant, is always the first step that then determines the subsequent components of the system. The traditional approach requires fabrication of samples or pins of the new fuel, measurements of physical and mechanical properties, and finally neutron exposure to high fluence under relevant operating conditions (e.g. temperature, stress constraints, interaction with coolant, etc.) with subsequent characterization. This approach requires a great expense of money and time (several years). In some cases, the fuel form may have become obsolete or irrelevant as a result of programmatic considerations by the time the experimental evaluation is complete. A similarly long process is required for the structural materials involved in the fuel pin cladding and other critical in-core components.

Spent Fuel Reprocessing: Reprocessing was abandoned in the 1970s as an option within the current nuclear fuel cycle, so there is great opportunity for development of advanced processes. Current reprocessing models provide only qualitative descriptions of process behavior. Empirical models of chemical behavior for major components are used to provide overall descriptions of various reprocessing strategies. These empirical models are based on benchtop experiments, and usually assume chemical equilibrium conditions are met instantly. Even then, current models are unable to answer many questions involving phase equilibria, such as precipitation from solution or determining oxidation states, where multiple possibilities exist. Very few reaction rate constants are known, and wherever transient conditions are simulated, they are usually just assumed or selected heuristically.

Fuel development and performance evaluation is currently an empirical process that takes decades. For acceptance, new fuels must be fabricated, placed in a test reactor over multiple cycles, tested under multiple accident scenarios, undergo post-irradiation examinations, and finally be placed in an operational reactor for several cycles. Fuel performance simulation tools can help to accelerate current fast and thermal reactor fuel qualification practices by helping decrease the time required to qualify new fuels, with the goal being a reduction by a factor of three in the current 10- to 15-year qualification time. The tools must accommodate all relevant fuel types in both normal operating (quasi steady state) and transient conditions.

Fuel performance simulation tools could reduce by a factor of 3 the time needed to qualify new fuels.

Current state-of-the-art models of nuclear fuel performance empirically capture various phenomena that occur in a fuel rod during nuclear reactor operation. Currently, these models are mostly limited to axi-symmetric geometries. In this simple axi-symmetric setting, the empirical models capture power generation, dynamic fuel/cladding gap, thermal analysis solution, simplified neutronics, fission product generation and propagation in a fuel rod, localized chemical state, fuel/clad swelling/creep, fuel cracking, fuel contact and mechanical interaction with the clad, and clad interaction with adjacent structures like grid spacers. A simple thermal hydraulics model is generally assumed for flow channel heat transfer. Because current models are based on empirical curve fits to fuel behavior in common nuclear reactor environments, they cannot be trusted to predict the behavior of fuels under conditions outside their narrow range of calibration. Transient fuel performance simulators are employed to predict the rod thermal/mechanical behavior under design basis accidents (DBAs), which for the light water reactor (LWR) fuel rod is a reactivity insertion accident (RIA). Survival of the cladding and maintenance of the core coolability are the primary issues for these scenarios. The time frame for these codes is on the order of milliseconds to seconds rather than the days to months for the quasi-steady-state simulators.

Reactor safety and design simulation tools require models for thermal hydraulics, neutronics, and structural mechanics. Such "reactor core codes" have been in existence for decades, but need improved physical, numerical, and geometric fidelity. Codes developed at the time used lumped parameter models for predictions of neutronics, thermal hydraulics, and structural mechanics quantities. These simple codes were calibrated against a very large experimental database, developed over the years for specific projects by calibrated against principally integral data. Existing reactor core codes, for example, employ the traditional single-channel or sub-channel approach to model reactor core heat transfer. Traditionally, separate thermal hydraulic code systems are used to execute design and safety calculations.

Neutronics modeling has traditionally relied on both stochastic (Monte Carlo) simulations and deterministic transport and diffusion theory approaches. Monte Carlo techniques incorporate the basic physics at the level of stochastic particle tracking with the general system geometry and material cross sections governing the particle track histories. Monte Carlo offers the strong conceptual advantage of keeping a close (essentially exact) correspondence between computational and physical geometric and cross section energy dependence models. Nevertheless, Monte Carlo can become computationally impractical for several different classes of problems. These include calculations of small reactivity coefficients, some types of sensitivity/uncertainty propagation studies, time-dependent solutions, and some types of burn up calculations.

Structural mechanics software development has been driven by a breadth of applications over the last 30 years which include automotive, aerospace, national defense, civil infrastructure, and, in the 1970's and 80's, nuclear reactor technology. These developments have led to a number of finite element-based computer programs that have relatively mature element technologies and solution algorithms. Existing software that has application relevance in the nuclear fuel cycle area can generally be divided into three categories: linear finite element programs, implicit time integration nonlinear finite element programs, and explicit time integration nonlinear programs.

2. Major Challenges

Significant exascale simulation challenges are present in each component of the fuel cycle. Opportunities exist in the areas of material property determination, spent fuel reprocessing, fuel performance, thermal and fast reactors, and waste forms and storage, to name a few. These are briefly discussed in the following.

Material Property Determination. With modern methods and powerful computing tools, one can foresee the opportunity to employ computational simulations to advance the evaluation and selection of advanced fuels and structural materials. For example, first principles methods can now realistically be used to determine fundamental material properties and support the development of new interatomic potentials that can be used in Molecular Dynamics (MD) simulations involving millions of atoms. These MD simulations can be used to study defect properties and to determine parameters such as the atomistic reaction rates that are required in coarser scale simulations of degradation mechanisms that take place over long times. These parameters can be employed in reaction rate theory or Kinetic Monte Carlo (KMC) models of microstructural evolution. MD-based dislocation dynamics simulations can also provide the fundamental dislocation-defect interaction parameters required for continuum 3D Dislocation Dynamics (DD) simulations. The 3D DD simulations can be used to obtain needed information on the constitutive behavior of the materials which is required for use in macroscopic methods such as finite element models. Taken together, this family of multiscale simulations can provide predictions from the atomistic and microscopic through to the mesoscopic and macroscopic level. An initial goal should be to establish the required degree of accuracy and practical limitations at each level of simulation (e.g. *ab initio*, atomistic, mesoscale, continuum). This will provide a basis for predicting the expected impact of an advanced simulation program to reduce both the absolute development time and the related uncertainties as part of the overall fuel and materials development effort. The initial technical objectives should include a strategy for determining the best approach for integrating the various multiscale components, i.e. when to use parameter passing and when models should be more tightly coupled.

Spent Fuel Reprocessing. Modeling of a reprocessing plant involves many complicated steps, each of which requires knowledge of several areas of physics, chemistry, or engineering. Fuel disassembly involves mechanical processes (chopping, filtering) and/or chemical dissolution in strong acid. The fuel solution is then passed through many stages of solvent extraction in order to separate several fission product and actinide streams. Several separate solvent extraction processes are required to accomplish this separation, each using different additives and components in the organic phase, as well as different acid concentrations in the aqueous phase. Safety considerations require: monitoring of volatile fission product and organic gaseous releases, careful evaluation of component inventories in each stage and even in piping, to avoid costly shutdowns, or repairs which must be performed remotely, strict attention to nuclear criticality safety in actinide solutions with widely varying component inventories, control systems which are based on realistic models of processes, not generalizations or even intelligent assumptions. The hazards of plant operation involve radioactive materials, toxic materials, strong acids, highly flammable materials, and highly volatile materials. Thus, detailed accounting of all components through all process stages is of utmost importance.

To support both detailed design and safe operation, the improvement of reprocessing models requires improved chemistry modeling, including both equilibria and kinetics. New fuel materials and requirements arising from nonproliferation concerns demand the use of modern sophisticated modeling tools for the design and optimization of a process consisting of several major steps, each of which presents its own chemical and physical complexity. These steps include fuel dissolution in acid, dissolved fuel treatment in a series of solvent extraction processes, and fabrication into fuel or waste forms. None of these steps is adequately simulated in a production sense, and some require experimentation to understand or confirm their chemical behavior. Advanced simulation can be used to help understand and optimize these processes, as well as integrate their behavior into the overall model in areas such as:

- Unit operations of chemical separations and materials processes based on first principles: extraction, evaporation, dissolution, etc.
- Three-dimensional, transient behavior of multiphase, multi-species reactive processes
- Turbulent, multiphase fluid flow and interfacial phenomena
- Predicting physical/chemical properties
- Development of separating agents/processes with functionalities specified by: affinity for target species; viability of synthesis; interaction with other materials; chemical and radiolytic stability; etc.
- Basic knowledge for process development, including: equilibrium partitioning in complex mixtures, transport and kinetics in non-equilibrium, multiphase systems, and improved process model solution algorithms.

Added value can come from the optimization of the fuel cycle as a whole, the possibility of detecting diversions, criticality problems, or possible effluent composition deviations outside specifications.

In addition to improved chemistry, additional elements of reprocessing systems must be modeled which are currently not even considered. Fluid dynamics must be considered in piping as well as in process equipment. Effects of control systems on component inventories, and vice versa, are necessary to adequately understand inventories throughout the plant. Interfaces with nuclear criticality calculations are important for both design and safe operation. Balance-of-plant modeling (such as volatile releases to the environment) must be included. As reprocessing gains in popularity, new and improved processes will almost certainly develop. Therefore, it is essential that computer codes be flexible, adaptable, and modular. They must not only be compatible with other codes which perform related or concurrent calculations, but they must be designed to function within larger systems of codes.

The simulation challenge is to couple these specific computational models into a coherent package that is considered a unified simulation with a distributed, modular modeling structure that allows interactive selection of specific models at levels of detail and breadths of scope required for analysis. The model structure must support three levels of coupled engineering detail:

- A *discrete event model* that provides throughput analysis, scheduling impacts, output compositions, and serves as the backbone to call supporting simulations at other levels of detail not captured in discrete event simulations.
- *Chemical process models* to represent modular, exchangeable simulations of specific unit operations or groups of operations. This allows a balance between simulation speed and required levels of rigor.
- *Specialized models* to address specific behavior accurately, including phase equilibrium calculations, computational fluid dynamics, specialized data retrieval, detailed adsorption models, and even detail as fine as computational chemistry if necessary.

Fuel Performance. Nuclear fuel assemblies must perform in aggressive environments characterized by stress, heat, corrosion, and irradiation, all of which lead to progressive degradation of the mechanical and physical properties of fuel cladding materials and other structural components.

The process by which nuclear fuel undergoes change in a nuclear reactor core is inherently multiscale. Materials issues to be considered include thermal and mechanical properties, swelling, microstructural phase changes and crack formation, as well as the effects of point defects and fission products. Reliable predictions of fuel behavior are essential to preventing fuel failures and to improving the economics of reactor operations.

Fuel performance modeling places an emphasis on the detailed understanding of the thermal, mechanical, physical and chemical processes governing fuel rod behavior during normal reactor operation and under accident conditions. Fuel rod performance codes are used extensively in research, by fuel vendors, and by licensing authorities for the prediction of fuel and cladding performance. The simulation tool should consist of a clearly defined mechanical-mathematical framework into which physical models can be easily incorporated. Besides its flexibility for fuel rod design, the code will be utilized for a wide range of different situations, as given in experiments, under normal, off-normal and reactor accident conditions. The time scale of the problems may range from milliseconds to years. All important physical models are included in the fuel performance code: i.e. models for thermal and irradiation-induced densification of fuel, fuel swelling due to solid and gaseous fission products, fuel creep and plasticity, pellet cracking and relocation, fission gas release, oxygen and plutonium redistribution within the fuel, volume changes during phase transitions, formation and closure of center void and treatment of axial forces (between the fuel and cladding), cladding creep and cladding/coolant interactions (such as oxidation), etc. Additionally, the code must have access to a comprehensive material database for oxide, mixed oxide, carbide, nitride, and inert matrix fuels, Zircaloy (and advanced zirconium alloys) and steel claddings (with the capability to add new fuel forms and advanced claddings). Also, interfaces (or subcoding) must be available for thermal/hydraulic and neutronics feedback to the fuel performance models.

Fuel rod simulators must be able to predict the thermal-mechanical-chemical response of the fuel rod throughout its irradiation life-time, including fuel rod failure mechanisms. Requirements for fuel performance simulation tools include

- Must be coupled with the nuclide inventory, which is required for the determination of the chemical/phase state of the fuel;
- Able to predict chemical species evolution and transport (in addition to FP);
- Able to predict the spatial distribution and chemical composition of separate metal and oxide phases;
- Able to predict fission product concentrations at the fuel-to-cladding gap and chemical interactions with the clad;
- Able to predict (via the transport models) the accumulation of volatile species (noble gases, Am, Cs, iodine) in the fuel pin free volume (especially the plenum);
- Can allow for chemical composition feedback into the fuel physical properties;
- Can allow for independent updating with most recent models and experimental results; and
- Can be directly validated using simple experiments and PIE results.

To meet these requirements, a high-resolution 3D spatial representation of the fuel and cladding that allows for non-symmetrical power generation and clad/fluid boundary conditions will be needed. In particular, resolving the thermo-mechanical response of the fuel and cladding on a sub-pellet level will be paramount. Coupling with other neutronics packages will also be necessitated, as will reaction-diffusion chemistry for fuels, cladding, coolant, and plenum gases. Direct coupling with databases of thermo-mechanical and chemical properties and models of irradiation effects on the properties will also be desirable.

The drivers for exascale platform requirements in fuel performance can be quantified. For example, a high-resolution simulation of a bundle of 40 fuel rods (~300 million elements with a 10-μ scale size), with only thermomechanics and no coupling to other multiphysics models, is estimated to require about a half day on a 20 PF platform. A more rigorous simulation of a single fuel pellet (~1 billion elements with a 1 μ scale size), again with only thermomechanics, would require approximately one day on a 1 PF platform. Incorporating all relevant additional physics, such as neutronics, fluid flow, and fundamental materials science, will easily multiply these requirements by a factor of 1000, bringing high-fidelity fuel performance simulation requirements to the exascale.

Reactor Safety and Design. Thermal hydraulic computational tools will be required to perform both design and safety analysis, allowing reactor and subsystem designs to incorporate a "safe from birth" philosophy. Long term focus is on the development of a next generation thermal hydraulic computational code that provides a completely integrated design and safety analysis architecture, that operates on leadership computing platforms, that seamlessly interfaces with the multiple physics necessary to perform high fidelity reactor and process analyses (neutronics, structural, fuels, chemistry, etc.), and that is experimentally validated to perform licensing analysis. Design and safety engineers will form the user base for these codes, and will greatly benefit from faster, more integrated, and higher resolution thermal hydraulic analysis systems. Expected code capabilities required to support the design and licensing of the experimental test facilities include: transient single and two phase fluid flow analysis, the capability to quantify the effect of both code and input uncertainty levels on design and safety limits, the capability to interface with neutronics and structural (and/or other as necessary) analysis codes, the ability to perform multidimensional thermal hydraulic analysis for selected components or component regions, and the verification, validation, and documentation packages required for licensing. Longer term requirements include the ability to analyze multiple coolants ranging from gases to liquid metals, the ability to use unstructured grids for three-dimensional thermal hydraulic analysis of any component, additional multi-physics interfaces including fuel analysis, coolant chemistry, etc, the capability of executing these codes on exascale platforms, and a code structure that allows the methodology to be physically validated.

Detailed neutronics analyses required to support the design, licensing, and operation of many of the components of the fuel cycle range from the physics analysis of reactors, criticality safety of the separations and fuel fabrication facilities, and radiation shielding and dose assessment. Accurate neutronics methods will be used by facility designers and regulators to assess the efficiency and safety of all nuclear systems and to identify areas where existing experimental data are insufficient. In addition, improvements in modeling capabilities can reduce unnecessarily conservative margins to ensure economical operation. A complete neutronics capability requires a tight coupling of several independent components that have traditionally been developed in a layered approach. The standard components of a complete neutronics suite of simulation codes include several computational tools and drivers, along with the ability to propagate uncertainties in the basic nuclear data and biases of each computational tool through the analysis.

The key components of a complete neutronics capability include: processing of data from evaluated nuclear data files to provide accurate, problem-dependent cross sections that account for material temperature and resonance shielding; radiation transport methods (Monte Carlo and deterministic) for steady-state and quasi-static time-dependent simulation of neutrons and photons within a system and the determination of the criticality condition; complete tracking of isotopic inventory changes due to fuel depletion, actinide transmutation, fission product buildup and decay, and associated radiation source terms; and integration of neutronics components with other physics packages, such as thermal-fluid dynamics, fuel performance, chemistry, and structural mechanics. This comprehensive suite of simulation tools is required to provide analysis for the design, construction, and operation of near-term facilities (short- and intermediate-term objectives) and must consist of the current best-in-class, qualified simulation tools. However, advanced optimization techniques, improved solution fidelity, and multi-physics coupling will provide substantial impact on the viability of commercial-scale facilities.

For structural mechanics, there is a compelling need for development and implementation of advanced material constitutive models that can accurately represent the time dependent behavior of materials in extreme environments. These should address the effects of high radiation levels, extreme temperatures, and chemical interactions on material behavior. Advanced materials models should account for the fully 3D, multi-axial states of stress

Difficult analyses such as the structural performance of the system under seismic loads are now possible with advanced simulations.

both at low strain rates (normal operations), and at high strain rates (accident scenarios). The development of macroscopic, continuum-based phenomenological models must progress in parallel to fundamental materials science research aimed at understanding microscopic material behavior in extreme environments. Recent developments in solid/structural mechanics have moved towards a merging of capabilities from traditional solids hydro-type codes, which have pushed the computational technologies for representing extreme deformations and flow of materials, with traditional structural type elements. Such codes have been developed in frameworks that prevent mesh tangling at extreme ranges of response. This has begun to open up substantially the types of problems that can be modeled for extremely nonlinear accident scenarios. It would be very desirable to move towards a single program that can solve multiple problems associated with slow (static and quasi-dynamic) phenomena associated with operations, slow accidental events like earthquakes, and also accurately simulate extreme accidents associated with very rapid transients such as pipe breaks.

Waste Forms and Storage. Different incoming waste streams will likely be sent to a geologic repository relative to the waste streams (mostly commercial LWR spent fuel) currently planned. This new waste steam, consisting mainly of fission-product wastes from recycling, will ultimately require that the repository be analyzed for the purposes of regulatory compliance, and perhaps have its design updated to take advantage of the significantly reduced loadings (heat loading, mass loading, and much shorter overall half lives) that the new waste stream will represent. If spent fuel in the current incoming waste stream is replaced by reprocessed fission-product waste in glass (or another solid waste form), the opportunity arises for a redesign of the repository, where multi-faceted simulations can more effectively capture the major processes for a suite of new design concepts. This will allow optimization of the safety (represented by dose to a human receptor on the surface in the distant future), the cost, and the repository volume within Yucca Mountain. For a new design, the opportunities for improvement are many. They include the configuration of tunnels within the mountain, the configuration of waste packages within each tunnel, and the waste loading per package. Difficult analyses such as the structural performance of the system under seismic loads, the interactions of incoming water with the drift walls including capillary forces, and fracture flow and transport, are all amenable to advanced simulation that analysts would not have considered even a decade ago, because the computer power required to execute such analyses was far beyond what was available.

3. Advances in the Next Decade

The stated long-term simulation goal for nuclear energy via a GNEP-based Program [DOE Office of Nuclear Energy, Science, and Technology, 2006] is the development of an architectural model that will facilitate the predictive modeling of the entire fuel cycle from mining through final disposition of waste material, taking into account interacting factors such as market forces, socio-political effects, and technology risk. This architecture must incorporate a comprehensive suite of simulation tools for the design, analysis and engineering of next-generation nuclear energy systems with enhanced safety, reduced environmental impact, optimal deployment of facilities, and reduced construction cost. The scope of these tools is daunting:

- Integrated 3D reactor core simulations with rigorous propagation of uncertainty
- Coupled thermal hydraulic and primary loop simulation
- Advanced fuel design and performance
- Fuel behavior engineering
- Advanced secondary loop and balance of plant engineering and analysis
- Advanced fuel cycle design
- Separations facility engineering optimization
- Repository design including seismic, geological, chemical, and thermal modeling and simulation

- Overall nuclear energy systems model development suitable for alternative economic analysis.

More broadly, exascale M&S objectives have the potential for a significant impact on nuclear facility design and operations:

- to study fuel cycle and resource issues (once-through, plutonium recycle, actinide burning, waste disposal, etc.);
- to develop and optimize nuclear plant and facility designs (systems design, physical layout, materials requirements, cost, economics);
- to demonstrate fundamental safety issues (defensive systems, passive safety, establishing the licensing basis);
- to avoid (reduce) costly prototype construction and operations (critical assemblies, prototypes, intermediate-scale plants);
- to accelerate optimum fuel selection (fuel irradiation is almost certainly required); and
- to characterize SNF-related requirements (on-site storage and criticality, shipping cask designs, and repository requirements);.
- optimization of performance in mixed utility electrical grids (cycle length, fuel resource requirements, economics, outages; analysis of hundreds of thousands of core loading options);
- demonstration of cycle-specific safety requirements (static and transient calculations: tens of thousands of coupled neutronics/thermal-hydraulic/systems computations);
- support for reactor operators (optimize startup and power maneuvers: thousands of calculations per cycle);
- on-line monitoring functions (real-time surveillance of safety margins: thousands of calculations per cycle); and
- operator training (real-time simulation on full-scope simulators).

4. Accelerating Development

Modeling and simulation of advanced nuclear fuel cycles will require a hierarchy of models of vastly different physical systems across a wide range of space-time scales, from detailed molecular dynamics of new materials to systems level simulation of the entire cycle. The final goal will be optimization in the presence of modeling and input uncertainty in order to design safe, reliable, economical, and socially acceptable end-to-end solutions for nuclear energy production. While there have been many advances in fundamental enabling technologies in mathematics and computer science in the past, additional research and development will undoubtedly be required to tackle a problem of this scale. At each level, new enabling technologies will be required to enhance predictive capability, understand and propagate uncertainties, model unresolved physics, and couple multiple physical models at a wide range of space-time scales. Likewise, new research and development is required to analyze, visualize, and optimize the results of such large simulations, and to do so in a way that is useful to designers and decision makers, who must be fully aware of the limitations of the computational predictions and the uncertainties inherent in the simulated results, due to the inevitable uncertainties of input parameters and modeling assumptions. Associated with this is the stringent need to establish careful protocols for simulation code verification and validation. These tools must be uniformly accessible within an integrated computational environment that reduces time-to-simulation, provides compatible geometry representations, and allows for a hierarchy of model fidelity running on workstations to state-of-the-art parallel computer architectures.

Multiphysics Coupling. Predictive simulation of each process within the fuel cycle requires accurate solutions to multiple, simultaneous nonlinear physical processes. Traditional simulation approaches to this problem involve segmented solution techniques whereby the simultaneous physics are assumed to proceed in a sequential, loosely-coupled manner. Such a solution approach is not nonlinearly consistent, is prone to numerical errors (particularly sensitive to time step size), and in some cases does not even converge (exhibits zeroth order errors). Such an approach is in general not predictive. Fully-coupled, nonlinearly-consistent multi-physics time integration algorithms solve this problem. Multi-physics coupling efforts will also provide common computational tools and interface structures, through executables and/or modules, to unite the varied temporal and spatial discretizations of each physics package to provide a consistent basis for analysis and information propagation. These products and interfaces will be used by facility designers to integrate multiple state-of-the-art physics simulation packages to provide a generalized coupling. This work will not only pay off in more accurate, predictive simulation results, but in many cases result in increased efficiency (time to solution), by virtue of being able to take a larger integration time step per given level of desired temporal accuracy.

The algorithmic and software coupling of multiple physics modules and codes will provide analysis capabilities for the design, construction, and operation of near-term facilities and must integrate the current best-in-class, qualified simulation tools. However, utilizing high-performance computational systems, high-fidelity simulation packages may be tightly-coupled to provide a fully-integrated facility simulation that will provide substantial impact on the viability of commercial-scale facilities. Associated strategies:

- Utilizing and developing advanced computational algorithms for fully-coupled, nonlinearly-consistent multi-physics time integration techniques;

- Providing the improved algorithms via standardized software interfaces for the communication of information from each physics package;

- Developing generalized interpolation, integration, extrapolation routines to ensure compatibility of solution with different geometric representations and time-scales;

- Integrating tightly-coupled physics packages into a unified module to provide an efficient solution on high-performance computing architectures;

Optimization and Confidence Analysis. Sensitivity analysis is used to determine the change in a computed result due to a change in some input parameter used in the calculation. Sensitivity and uncertainty (S/U) analysis is important not only for performing methods/data V&V studies, but also for certifying that a proposed system design satisfies all performance and safety specifications. This is especially significant because the new reactor and fuel processing/reprocessing systems have not been previously characterized by measurements. Specifically, S/U methods can provide the following types of information:

- Improved physical insight into the underlying phenomena governing the system of interest, by indicating relationships between variables. This is often useful for guiding design modifications and interpreting hypothetical accident scenarios.

- Realistic, best-estimate design margins that can reduce the tolerances obtained from bounding-analysis.

- Quantitative ranking of important modeling/data parameters that impact the calculated results. This analysis can identify major sources of uncertainties and can determine the required measurement accuracy in the input data necessary to achieve a desired accuracy in the computed results.

- Rapid (but approximate) evaluation of how design perturbations affect computed output parameters. This can be coupled to a system optimization algorithm.

Although developed most extensively for criticality safety and reactor physics analysis, sensitivity techniques can also be applied to other types of calculations, including shielding evaluations for reactors, reprocessing, and transportation systems, fuel depletion studies of actinide burning, core lifetime, and proliferation indicators, ex-core fuel cycle parameters such as source term activities/decay heat, and safety analysis involving reactor kinetics with thermal hydraulics feedback.

5. Risks

Nuclear energy does not have the same drivers, from an M&S perspective, as nuclear weapons, climate change, or the aircraft industry (to name a few). Unless M&S activities are driven by requirements set by industry and regulatory agencies, approaches in which "better" can become the enemy of "good enough" may prevail and defocus efforts. The oft-quoted statement that "all models are wrong, some models are useful; the only way to tell the difference is to get some data" applies to the role of M&S in the nuclear fuel cycle. V&V activities will be crucial.

6. Expected Outcomes

Enhanced use of M&S for nuclear energy will lead to improvements in knowledge and reduction of uncertainties that will produce cost savings in current reactor operations and substantially reduce the cost of future reactors. Such improvements could also provide a basis for innovative designs that will reduce the need for excessively conservative (and costly) system specifications, improve efficiency and performance, enhance safety and reliability, and extend operating lifetimes.

For the existing fleet of NPPs, simulation can reduce margins, thereby increasing performance. For example, for 100 reactors with an operating cost of $1 million/day, a 5% power increase results in an annual return of $2 billion, for 20+ years. For advanced new reactors, optimized designs can be developed without any "learning curve" (thus increasing years in operation). For the current fleet, for example, the difference between 30 years at 60% capacity and 10 years at 90% capacity can be viewed as unrealized potential equating to $328 billion (in 2007 dollars), over 30 years.

As an example, for the three areas of reactor and fuel cycle technology, the principal benefits are as follows.

- Exascale M&S could have a substantial impact on **spent fuel reprocessing** in areas such as the simulation of separations; the development of new separation agents, addressing whether these agents can be "engineered" to provide the desired results; the performance of full-scale plant simulations using first principles (some codes are adequate, while others need work); and the integration of multiple codes. Exascale computing is expected to reduce R&D cost and time, improve and accelerate the design process, support process scaleup, lower facility costs, and provide opportunities for major change. The biggest payoff for M&S in reprocessing, however, is expected to result from advances in the modeling of waste forms, which should make it possible to delay or avoid the construction of additional repositories for SNF.
- The application of exascale computing to **fuel performance** should reduce the time needed for fuel development and qualification and lead to reliable assessments of life cycle performance, predictions of fuel rod behavior in a design basis accident (DBA), and predictions of transuranic (TRU) fuel behavior.
- Benefits of exascale M&S for **reactor analysis and design** include eliminating unrealistic assumptions that drive to more conservative designs and thus higher cost, helping to achieve higher power efficiencies, reducing learning curves to efficient operation, improving safety posture, optimizing the design of the power grid and the fuel cycle, and supporting better (more efficient) operations, including inline monitoring and operator training.

References

DOE Office of Basic Energy Sciences (2006), *Basic Research Needs for Advanced Nuclear Energy Systems: Report of the Basic Energy Sciences Workshop on Basic Research Needs for Advanced Nuclear Energy Systems*, July 31–August 3, 2006 (http://www.sc.doe.gov/bes/reports/files/ANES_rpt.pdf).

DOE Office of Nuclear Energy, Science, and Technology (2006), *Global Nuclear Energy Partnership Technology Development Plan*, Global Nuclear Energy Partnership Technology Development Program Integration Office, January 9, 2007.

3 Biology

Microbial life is responsible for all major processes on Earth, from growth of corn in a field to deposition of carbon under the oceans. Understanding the role of microbes in these processes is the first step in manipulating them, enhancing the positive aspects of microbial biology to enhance all life on earth. Our current understanding of microbial life comes almost entirely from experimental observations. Biology must capitalize on emerging technologies to combine computational analyses and modeling into a greater understanding of microbial systems and their interactions within a community. Such understanding will pave the way for the new scientific frontiers of synthetic biology and will be central to ecological sustainability in the future.

Biological experimentation has revealed the intricate interplay between proteins and their ligands: crystal structures and genetic tests have demonstrated atomic and molecular-level interactions driving reactions that raise or raze complex macromolecules essential for life. Within cells, biopolymers and molecular complexes are constructed through regulated pathways. As we enhance our rudimentary understanding of the connections between the systems that form a cell, and their temporal and spatial separation, we move toward modeling whole microbial cells. Modeling steady-state cellular growth will expand and adapt into real-time continuous culture modeling. In nature, cells do not live in isolation. Combinations of microbial cells form communities, and coordinated microbial actions affect local and global environments.

The genomics period started in the mid-1990s with the first complete microbial genome sequences, and progress towards more rapid and cheaper sequencing has continued unabated. As microbial sequencing continues to exponentially produce the blueprint of cultivated microbial life, comparative computational genomics tools are enhancing our ability to model individual proteins, groups of proteins, and microbial cells. These analyses are helping to unravel novel metabolic and regulatory pathways that have not been characterized previously. Moreover, the recent advent of environmental sequencing—understanding the molecular makeup of all organisms in an environment simultaneously—has revolutionized our comprehension of microbial diversity on Earth. To this wealth of information, high-throughput experimental capabilities, such as mass spectroscopy and chip technologies, are providing information that adds meaning to the sequence.

HPC in biology is accelerating the transition from hypothesis-driven to design-driven research at all scales. Computational simulation of biological systems is beginning to drive the direction of biological experimentation and the generation of insights. Computational biologists thus are proving to be at the vanguard of the revolution in biological sciences.

1. State of the Art

Microbial life affects every known physical and geochemical process on the planet. If we are to manipulate and control the pathways and mechanisms, we must understand the roles of microbes in these critical processes.

Determining how a protein is folded *in vivo* and then how that protein binds its ligand or substrate is central to understanding the function of that protein. For the majority of proteins that are identified, the ligands are unknown or assigned based on sequence or

Understanding microbial systems will pave the way for new scientific frontiers of synthetic biology and will be central to ensuring ecological sustainability in the future.

With exascale computing, scientists can run models that fully represent the complex, spatially heterogeneous, multiscale, multimodel processes of microbial systems.

structural similarity to other proteins. Therefore, the modeling process may begin with docking or protein:ligand interaction studies to identify suitable substrates from the entire complex chemical universe. A subset of these searches is critical to drug discovery—finding enzyme inhibitors that block specific functions. Searching through portions of chemical space and comparing those 3D chemical structures with 3D protein structures requires massively parallel computations. However, once a protein:ligand interaction has been demonstrated computationally, genetically, and structurally, modeling the temporal variations in specific atom-level interactions remains beyond the realm of current computational capacity.

New techniques are providing high-throughput structure prediction and characterization [Goens et al. 2004,] but exascale computing is required to predict structures for all identified proteins. The complexity of this problem is reduced through identification of "protein families" —groups of similar (but not identical) proteins that are common to different microorganisms. Clustering, similarity, alignment, and maximum likelihood trees and phylogenetic methods will be leveraged to reduce the complexity biological experimentation, to predict functions, and to design experiments that can confirm or refute those predictions. With the advent of technology and price points that enable complete sequencing of all cultivated microbial life, these computations will become ever more complex. These three intertwined problems—alignment, phylogenetic trees, and structure predictions—are at the heart of protein function prediction and currently operate at the terascale. For example, there are approximately 10,000 proteins in families. To generate 100 trees per family, at approximately 1 day per tree, requires 10^6 CPU-days with current computational platforms. These phylogenetic trees will improve the design of structure prediction and testing models that will in turn lead to enhancements in the phylogenetic methods. Advances in algorithms for string matching, similarity searching, and identification of similar proteins that are required to reduce the computational load for these approaches all depend on high-performance, integer-based computations.

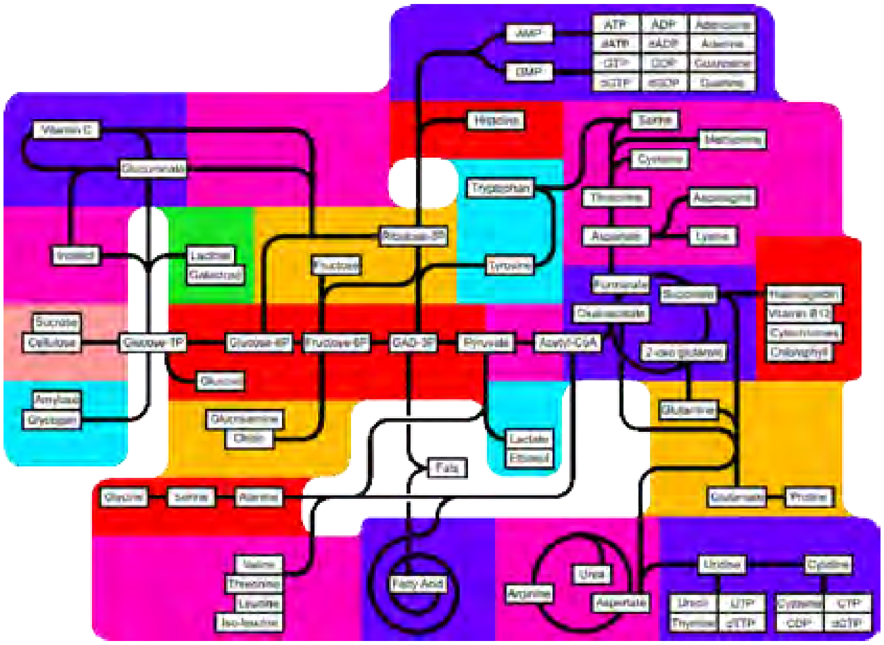


Figure 3.1 Central metabolism in all organisms consists of a set of interconnected pathways. Although the components of most of the pathways are known, modeling microbial metabolism, signal transduction, and regulation currently can handle steady-state processes. Exascale modeling will allow real-time fluxes to be simulated, leading to design driven experimentation. (Image from: http://en.wikipedia.org/wiki/Metabolic_pathway)

Systems biology aims to develop validated capabilities for simulating cells as spatially extended mechanical and chemical systems in a way that accurately represents processes such as cell growth, metabolism, locomotion, and sensing. Initially this work will be developed in the much simpler, and better understood, bacterial realm rather than at the level of complex multicellular systems. Currently, the state-of-the-art methods for modeling bacteria are flux-balance analysis and regulatory and signal transduction network analysis. Metabolic pathways can be extracted automatically from an annotated genome, and derivation of stoichiometric metabolic networks suitable for flux balance methods is becoming automated. More realistic models of bacterial growth in defined conditions will soon replace these simplistic models of cells in a single state. Two developments render such an effort feasible with exascale computing: (1) first principles–based algorithmic approaches to fully represent the complex spatially heterogeneous, multiscale, and multimodel processes characteristic of microbial modeling; and (2) the wealth of new experimental techniques that can provide the basis for validating the models generated, including high-throughput methods for acquiring genomics and proteomics data, and high-resolution imaging techniques that can, for example, track the locations of individual molecules in a cell. The development of simulation models will provide a theoretical framework that will transform these massive data streams into design-driven research priorities.

Unlike other modeling efforts, microbial-sized multiscale models may enable design-driven research that can be rapidly tested by experimental biologists. This combination of modeling, prediction, and testing will generate a positive feedback loop, continually enhancing the models that are generated. This new paradigm, with simulation driving the hypotheses to be tested by the biologists, will place computational biology firmly at the helm of biological discoveries.

As combinations of microbial models are combined into communities, it will be possible, initially at an entirely different scale, to model whole communities not only of microbes but of their associated Eukarya: plants and animals. To understand and manipulate carbon fluxes in terrestrial or aquatic environments, to enhance energy production or carbon sequestration, will require not only understanding individual isolated organisms but also modeling the entire balance and flow of carbon, nitrogen, phosphorus, and oxygen through the life cycle of the ecosystem. New models will be needed that extend to microbial eukaryotes such as the fungi most critical for nitrogen fixation in the rhizosphere. In addition, microbial ecology will need to learn from traditional macro ecology and models of different ecosystems. Studies in environmental microbiology are just beginning to unearth the fundamental links between micro- and macroecology that will be critical to understanding microbes' roles in the world.

2. Advances in the Next Decade

Microbial life is so pervasive, and so essential for human life and sustainability, that we need to understand the connections from individual proteins through whole cells and into ecosystems and environments. Development of tools for simulation and modeling of microbial life using exascale computing will impact virtually every aspect of human interaction with nature. Tools developed for exascale computing models of microbial life at each scale will be used to simulate the effects of perturbations. These models will provide unparalleled ability to produce hypotheses and to design experiments from the computational framework, to test the models through interactions with biologists, and to refine the models based on the results of the experiments.

Models and simulations will be generated at the molecular dynamics, systems biology, and ecosystem levels. A challenge will be to integrate these models from different scales into unified systems that will stimulate quantitative biology. The exascale computational framework will accelerate the transition of computing in biology to design-driven research. Enhancing our understanding and modeling of microbial life will benefit a broad range of application areas:

- ***Energy.*** Microbial modeling at the exascale will identify new proteins, pathways, subsystems, and manipulations that will be used as biocatalysts and to derive biofuels, and to accelerate the development of biofuel cells and direct solar energy cells.

- ***Environmental remediation.*** Microbial modeling at the exascale will provide the ability to engineer individual organisms to more effectively transform unwanted compounds to harmless metabolites, as well as the ability to engineer multispecies communities for efficient functional purposes through design-driven research.

- ***Industrial-scale microbiology.*** Microbial modeling at the exascale will produce recipes of mutations that transform wild microbial strains to production strains (for production of fine chemicals, pharmaceuticals, and next–generation green feedstocks). Complete metabolic and functional models, design-driven predictions, and experimental confirmation pro-

Development of tools for modeling microbial life at the exascale will
provide unparalleled ability to simulate the effects of perturbations
on microbial systems.

Simulating 100 ms of a protein folding is expected to require 3 years on a petaflops architecture; with exascale computing, scientists will be able to simulate the entire self-assembly process in microseconds.

viding positive feedback are all required to ensure high-quality products.

- ***Carbon sequestration.*** Microbial modeling at the exascale will provide the ability to manipulate microbes to sequester carbon from the environment via mineralization. Without understanding the very nature and forces driving the fixation of CO_2, we will not be able to reduce the concentrations of atmospheric greenhouse gases.

- ***Sustainability.*** Microbial modeling at the exascale will yield models of complex, productive, and sustainable environments, like the prairie grass that does not require exogenous fertilization. These ecological models provide the key to sustainability and security into the future.

Enhancements in our understanding of biology at each scale—from atomic, through genomic and cellular, to ecosystems—are driving the need for high-performance biological computing. These approaches and algorithms bring their own challenges and problems, in particular the need for multiscale modeling. However, computing at the exascale level will provide unique opportunities to reveal intimate knowledge about the smallest, and yet the most important, members of the biosphere.

3. Major Challenges

The *technological challenges* for microbial multiscale modeling are vast, but not insurmountable. The central technological challenge is to advance modeling to the point where it can produce a steady flow of predictions that are fed into wet-lab operations for functional characterization. These characterizations will feed back to the modeling to provide near real-time assessment of the quality of the model. In essence, this approach will construct a growing body of models that are continuously calibrated against the parallel growth in phenotypic data.

Not all the challenges of moving into the exascale computational era are technological, however. Significant *biological hurdles* must be overcome before exascale computing can be efficiently applied to multiscale microbial systems. For example, determination of the structures of diverse proteins, including traditionally intractable structures such as membrane proteins, is critical for understanding the flow of compounds within and between cells, and essential for protein-scale modeling components. Increased genome sequencing capacity—initially sufficient to handle sequencing all known microbial genomes—is essential to capture the diversity of life on earth. However, at projected sequencing rates of hundreds to thousands of genomes per year, DNA availability, logistics, and manual curation of sequenced genomes are likely to pose more of a hurdle than sequencing capacity itself. As we move toward exascale computing, technological advances will likely be capable of delivering a whole genome as a routine microbial assay, a fundamental requirement for understanding ecosystems-level biology.

The biotechnological advances will demand similar *computational* advances. Current biomolecular modeling and simulation capabilities used for self-assembly in molecular biology rely heavily on coarse-grained techniques and empirical approximations for particle-particle interactions. The simulation is also necessarily limited in length, and current scales do not capture the long time periods of the self-assembly process. For example, the IBM Blue Gene project estimates that to simulate 100 microseconds of a protein folding will require about 3 years of computation on petaflops architecture. With exascale computing, the entire self-assembly process will be simulated over micro- or millisecond time scales, leading to new scientific insights and design-driven research. Combining higher-fidelity and longer-time simulations will enable research into the parameter space that affects individual protein:protein and protein:ligand interactions (temperature, pressure, different connection and surface capping molecules, etc.) or even the use of optimization procedures to design new structures with desired properties. The challenges for other types of computational biology (such as homology searching) are even more overwhelming as a result of high I/O and memory requirements and traditional dependence on shared-memory resources.

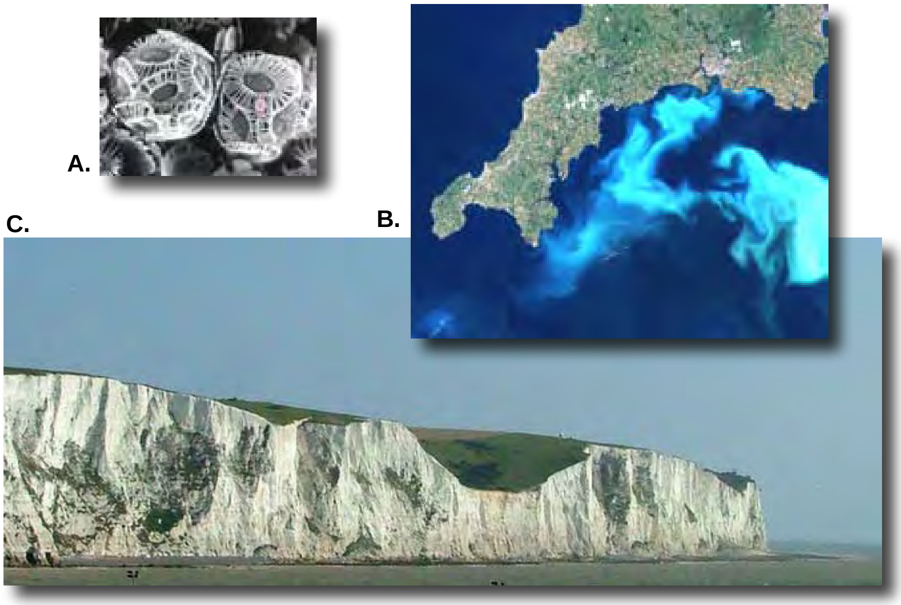


Figure 3.2 Microbes are the original carbon sequesterers. Microbial cells (Emiliana huxleyi; A) are approximately 5 µ across, yet when they bloom in the ocean the vast numbers of cells are visible from space (B; an E. huxleyi bloom off the southwest coast of England). Over millennia, these microbial blooms are deposited on the ocean floor, storing carbon as calcite, and hence removing CO_2 from the atmosphere (C: The white cliffs of Dover were created by microbes, especially cocolithophors). Design-driven biology based on exascale computing could identify how to enhance the growth of cocolithophores, increasing deposition of carbon on the ocean floor. (Images A and B from http://en.wikipedia.org/wiki/Emiliana_huxleyi. Image C from http://en.wikipedia.org/wiki/White_cliffs_of_dover.)

Improving the consistency and accuracy of protein functional annotations through the elucidation of metabolic pathways or processes as subsystems that cover almost all of the machinery embodied in the newly sequenced genomes will be central to unraveling the metabolism within each microbe and the roles of these microbes in their environments. These improved annotations will also feed into systems biology approaches to understanding the holistic nature of the cell. However, modeling and simulation provide only a glimpse of a narrow local view of each process without interaction between modalities and scales. The approach to addressing all of these challenges is development of exascale algorithmic and software tools for representing the various macroscopic subsystems, combined with the application of these tools in various combinations to simulate specific problems in systems biology. The design of the tools and their application requires both biologists and mathematicians, with theory and experiments informing the design of models for specific problems and the factorization of the algorithmic tools required into reusable software components. The feedback between biological experimentation and mathematical tool development is an ongoing process that will yield robust representations of systems biology across multiple specific problems. For example, we have neither the understanding of how to attack the hybridization of stochastic and deterministic algorithms in a single simulation, nor any ideal what the biological design-driven outcomes of such a synthesis might be.

4. Accelerating Development

Parallel advances in biological and computational sciences will be required to realize multiscale microbial modeling at the exascale. Advances will be required at each individual layer from atomic protein models through whole cell modeling to ecosystem-scale models.

Biomolecular dynamical models need to adapt to varying parameter space, local variations of constituents, and effects of nearby proteins. Entirely new models that can leverage exascale computing to realize micro- or millisecond-scale protein simulations are essential for this modeling to succeed beyond the nanosecond scale.

Technologists are driving the sequencing and annotation of thousands of single-celled organisms (including Archaea, Bacteria, and Eukarya). By the time exascale computing arrives, the majority of diverse microbial organisms will be completely sequenced. Improvements to terascale and petascale integer-based computations will be essential to ensure that computational efficiency scales with technological efficiency. The federation of molecular databases will result in kernel databases that are well annotated and maintained and support massive amounts of

Exascale algorithms and software tools will be critical for representing macroscopic subsystems, a key step to understanding the holistic nature of the cell.

New biomolecular models must be able to leverage exascale computing to realize microscale or millisecond scale protein simulations.

sequence data. These databases will have to address data centralization versus distribution issues common to other data-intensive scientific endeavors. However, these databases, together with an enhanced computational framework that accelerates the modeling and characterization of heretofore-unknown biochemistry, will arise simultaneously with our genome-level understanding of microbial cells.

The new computational framework for simulating complete cells will need to incorporate protein functions from these federated databases, regulation and cellular states (experimentally tested by microarrays), and compound concentrations and stoichiometries required to model flows through individual cells, groups of cells, and complex ecosystems. In addition to these separate developments, each of which will occur in parallel, a complex integration layer that transcends the separate domains will also be needed. This multiscale modeling will be essential for understanding how proteins interact with other proteins and/or their ligands in one cell and how these interactions might affect the metabolism of a cell elsewhere in the ecosystem. This new approach to computing in biology will be the driving force behind the determination of the role of proteins, pathways, microbes, and communities in complex biological processes.

5. Expected Outcomes

The generation of predictive capabilities is at the heart of exascale microbial modeling and will drive the transition to design-driven biological sciences. However, existing research infrastructure is not able to support the transition to exascale computing. Therefore, additional investments are needed in key target areas to ensure that computing in biology is appropriately situated to take advantage of the exascale computing opportunities.

- ***Novel molecular dynamics algorithms for microsecond simulations of proteins and ligand interactions.*** Enhancing the algorithms used for molecular dynamics approaches will include extending them out to microsecond or millisecond timescales for all proteins encoded by all microbes (bacteria and viruses) sequenced:
  - Protein structure prediction and classification
  - Prediction of interacting protein partners
  - Prediction of protein-protein complexes
  - Refinement of function prediction (e.g., specifics of substrates the protein may bind)
  - Prediction of structure-function changes caused by single amino acid (aa) changes or indels [e.g., nonsynonymous single nucleotide polymorphisms (SNPs)]
- ***Predictive capabilities of metabolic models.*** Automatic generation, testing, and verification of models for keystone species for given environments and for the 50 species with the most wet-lab data will lead to the framework for high-throughput generation of accurate flux-based models for all sequenced microbes. These models will predict the state of the cells and the transitions between states, while allowing comparison of those predictions to measured variables to assess the efficacy of the model.
- ***Modeling and simulation of less complex microbial communities.*** Several low-complexity microbial systems have been well studied at the molecular and sequence level. These include the acid mine drainage system, the enhanced biological phosphate removal system, the Soudan mine, and the solar saltern crystallizer ponds. Modeling these environments will move us towards understanding complex environments at the exascale.
- ***Algorithms, architectures, and methods for integrating networks.*** The integration of transcriptome, metabolome, genome, and other data requires new algorithms that will enable the adoption of exascale technology by computational biologists. Some of these algorithms may require novel computer architectures that go beyond the limits of floating-point-centric,

memory-limited cluster computing that are suitable for the physical sciences.

6. Required Investment

In order to enable exascale systems biology, funding must be available for generation of experimental data (both large-scale data and painstaking measurements of critical parameters), development of experimental methods to measure biochemical parameters in a high-throughput fashion, data standardization [Klipp et al. 2007], and database integration, as well as experimental validation and verification of model predictions.

7. Major Risks

A major risk is that biologists are not sufficiently trained to leverage high-performance computational methods and resources. A critical component of facilitating exascale microbial modeling is to cross-train microbiologists with state-of-the-art modeling tools and applications.

A related risk is the failure to lower the barrier for access to exascale computers. Accessibility of simulation software for microbiologists is a constant problem. Failure to address this problem will impede the adoption of exascale computing.

The generation of predictive capaabilities is at the heart of exaascale microbial modeling.

References

S. S. Goens, S. Botero, A. Zemla, C. E. Zhou and M. L. Perdue (2004), *J. Gen. Virol*. 85, 3195.

E. Klipp, W. Liebermeister, A. Helbig, A. Kowald and J. Schaber (2007), *Nature Biotechnol*. 25, 390.

4 Socioeconomic Modeling

Decades of research have deepened scientific understanding of the impact of greenhouse gases on climate, as documented, for example, in the assessment reports of the Intergovernmental Panel on Climate Change (IPCC). Central to this research are Earth system models (ESMs) incorporating detailed representations of the atmosphere, ocean, cryosphere, and biosphere and their interactions, largely through chemical systems. Model predictions are now considered to be reasonably reliable on global and continental scales, and research is now being directed towards improving predictive capability at regional and subregional scales.

As the scientific community's confidence in these model results grows, discussion of impacts and of potential adaptation and mitigation strategies also grows. Modeling tools can play a vital role in developing the scientific knowledge base required to understand these issues. In developing such tools, we must recognize that the impact of climate change on society and the ultimate effectiveness of specific responses depend critically on human actors, who will determine, for example, how energy supply and demand evolve over time and how and when different "solutions" are deployed and applied. Hence, we must model human responses if we are to understand the likely effectiveness and impacts of different responses and thus help to sustain a prosperous and secure society.

Integrated modeling of the social, economic, and environmental system with an extensive treatment of couplings among these different elements and consequent nonlinearities and uncertainties is a scientific and computational grand challenge. Existing integrated models incorporate treatments of economic impacts and impacts on human well-being, while bottom-up economic and energy models describe, for example, the greenhouse gas emissions of different industry sectors and mitigation costs. However, computational limitations have prevented any existing model from including substantial regional and sectoral disaggregation, a dynamic treatment of world economic development and industrialization, and detailed accounting for processes such as technological innovation, industrial competition, population changes, and migration. Other than emission scenario drivers for climate models, feedbacks from human activities to climate change are generally not addressed, and it has not been feasible to explore the uncertainty range within climate and economic model domains with Monte Carlo or more general Bayesian methods. The lack of computational power has also limited the ability to apply the best statistical techniques to existing data.

The emergence of petascale and (within the next ten years) exascale computers makes it possible, in principle, to attempt a detailed and fully integrated treatment of these diverse factors. By allowing for far more detailed treatments of the various components and feedbacks among these different components, and issues of uncertainty and risk, exascale computers have the potential to transform understanding of socioeconomic-environmental interactions.

Such detailed and integrated models can allow for the quantitative study of key questions, including the following:

- How will climate change impact energy demand and prices?

Integrated modeling of the social, economic, and environmental system with an extensive treatment of coupling of these different elements and consequent nonlinearities and uncertainties is a scientific and computational grand challenge.

- How will nonlinearities, thresholds, and feedbacks in the coupled climate-economic-energy system impact both climate and energy supply?

- How will different adaptation and mitigation strategies affect energy supply and demand, the overall economy, the environment, individual products and services, public health, and the vulnerability of the U.S. economy and infrastructure?

- How can computational approaches help us to identify and visualize good strategies for R&D, policy formulation, and technology adoption under conditions of future uncertainty, but with future information feedback opportunity?

- How are answers to these questions influenced by other processes, such as population growth and demographic change; economic growth and development, particularly in Asia; immigration; and technological change (such as teleconferencing technology)—all of which will affect transportation and land use patterns, human behavior, and business conditions?

- How will the success of potential solutions be affected by technical issues vs other social, political, regulatory, and market factors, such as barriers to entry for new technology providers?

DOE's leadership role in high-end computing and its strong interest and expertise in climate change make it natural for DOE to take a leadership role in an R&D program aimed at building detailed, integrated socioeconomic-environmental models designed for exascale computers. Building on DOE expertise in climate and energy system modeling, and bringing to bear the latest methods in economics, quantitative techniques in behavior and decision theory, and HPC tools, such a program can develop tools to deepen the understanding of technical, economic, and social issues that underpin the climate change challenge.

With expertise in high-end computing and climate change, DOE is in an excellent position to lead the development of detailed, integrated, socioeconomic-environmental models for exascale computers.

Such a program should aim to create a high-performance, high-fidelity modeling framework that incorporates detailed treatments of the various components listed above and that allows for computational investigations of both individual elements of the socioeconomic-environmental system and the entire coupled system. The program must address a wide range of methodological and computational problems, including data management and acquisition, parameter estimation, technology characterization and forecasting, socioeconomic model specifications, regional and subregional disaggregations appropriate for ESMs, and the need to quantify the degree of uncertainty in forecasts. This system will capture feedbacks between socioeconomic systems and climate and will allow for quantitative evaluation of potential strategies designed to reduce greenhouse gas emissions with minimal societal costs.

In a series of staged R&D steps, the program can:

- Enhance existing socioeconomic models to provide far greater geographical and temporal resolution of important processes, interactions, and feedbacks.

- Incorporate important socioeconomic elements previously treated as exogeneous, such as economic development, exchange rates, and foreign industrialization.

- Invest in multidisciplinary tool development including algorithms, validation, data analysis techniques, and uncertainty analysis.

- Extend large-scale socioeconomic modeling to new domains, such as analysis of market barriers to new technologies and health impacts of climate change.

- Integrate existing and new data sources to allow for rigorous model validation.

- Integrate global ESMs with global socioeconomic models that highlight human and geophysical interactions.

In the following, we describe this need and our approach in more detail. We also mention other potentially important applications of the proposed socioeconomic model, which can be both coupled with ESMs and used in complex computational settings such as game theory and stochastic dynamic programming

1. Importance of Global Socioeconomic Modeling

Earth system modeling has progressed to a point where there is considerable confidence in predictions of continental- and global-scale climate changes over the next 100 years [IPCC 2007]. We are thus understandably interested in determining likely impacts of climate change on ecosystems, human society, human health [McMichael et al. 2003] and well-being, the economy, and national security and in understanding the effectiveness of potential adaptation and mitigation strategies.

One set of questions that we must answer to address these issues concerns the geographical and temporal distribution of climate change. For example, will climate change increase or decrease the level, variability, and timing of rainfall and temperature in agricultural regions? Will it increase or decrease the frequency of hurricanes over low-lying coastal regions or extend the scope of vulnerable regions? To answer these and other related questions, DOE has defined as a primary goal over the next five to seven years the improvement of the performance of ESMs on regional and subregional scales.

A second set of questions, equally central to understanding climate change impacts and responses, concerns climate system-human system interactions (see Figure 4.1). Such interactions can be both important and complex, as the following examples demonstrate.

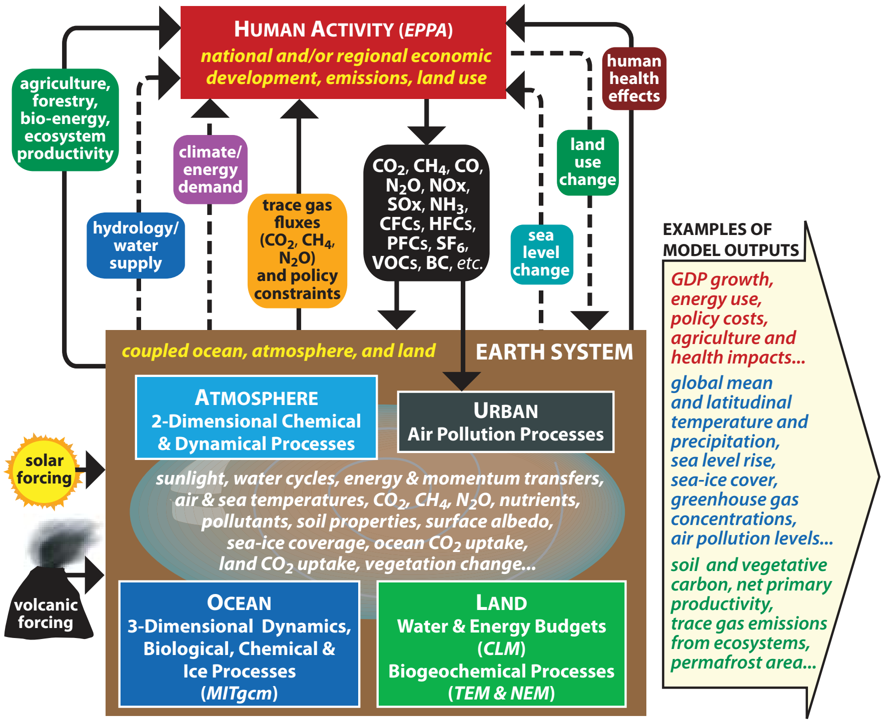

HUMAN ACTIVITY (EPPA)
national and/or regional economic development, emissions, land use
agriculture, forestry, bio-energy, ecosystem productivity
climate/ energy demand
hydrology/ water supply
trace gas fluxes (CO_2, CH_4, N_2O) and policy constraints
CO_2, CH_4, CO, N_2O, NOx, SOx, NH_3, CFCs, HFCs, PFCs, SF_6, VOCs, BC, etc.
land use change
sea level change
human health effects
coupled ocean, atmosphere, and land
EARTH SYSTEM
ATMOSPHERE 2-Dimensional Chemical & Dynamical Processes
URBAN Air Pollution Processes
sunlight, water cycles, energy & momentum transfers, air & sea temperatures, CO_2, CH_4, N_2O, nutrients, pollutants, soil properties, surface albedo, sea-ice coverage, ocean CO_2 uptake, land CO_2 uptake, vegetation change...
OCEAN 3-Dimensional Dynamics, Biological, Chemical & Ice Processes (MITgcm)
LAND Water & Energy Budgets (CLM) Biogeochemical Processes (TEM & NEM)
solar forcing
volcanic forcing
EXAMPLES OF MODEL OUTPUTS
GDP growth, energy use, policy costs, agriculture and health impacts...
global mean and latitudinal temperature and precipitation, sea level rise, sea-ice cover, greenhouse gas concentrations, air pollution levels...
soil and vegetative carbon, net primary productivity, trace gas emissions from ecosystems, permafrost area...

Figure 4.1 Earth system–human system interactions, as captured in the MIT modeling system [Sokolov et al. 2005].

- Biofuels have the potential to reduce both overall greenhouse gas emissions and dependence on foreign oil. On the other hand, the production of biofuels can increase both water consumption and food prices, with implications for human societies—and, if grown on previously unfarmed land, greenhouse gas emissions [*Sustainable Bioenergy* 2007]. The ultimate cost-effectiveness of biofuels may also be influenced by changes in the productivity of land used to grow biofuels, as a result of climate change, or by changes in energy demand due to changes in temperature patterns.
- Reductions in greenhouse gases may depend on the emergence and adoption of new energy production, distribution, and consumption technologies. Thus, we may ask what factors affect the emergence of new technologies and, for potential climate change solutions, how their success will be affected by technical or other factors, such as availability of capital for these ventures and obstacles to entry for new providers.
- The nature and pace of climate change response may depend on international agreements. In such agreements, there may be winners and losers. We may want to quantify the benefits or costs of alternative responses to different countries, states, and industries and study the interactions that may occur between different parties. Such studies can contribute to the design of negotiation strategies and international emission treaties, including penalties for violators.

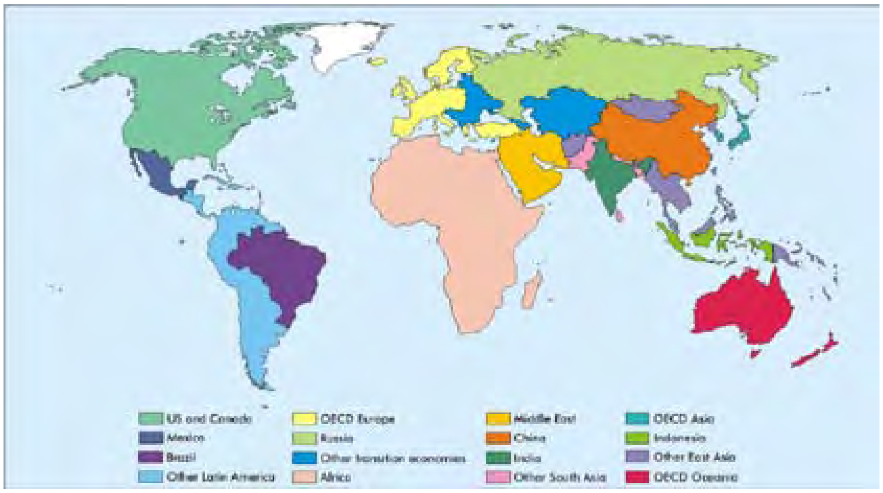


Figure 4.2 Aggregated world regions used in the World Energy Outlook (WEO) 2006 study [IEA 2007].

Answers to such questions may be sensitive to changes in population distribution, employment patterns, income growth and prices, demand for resources, and human preferences. Thus, we need to understand trends in population growth, demographic change, economic growth and development (particularly in Asia), technological innovation, and behavioral shifts such as increasing consumer preferences for particular energy-intensive services and pressure on companies to reduce their environmental footprint.

Socioeconomic modeling seeks to construct quantitative descriptions of these aspects of human systems. Climate models are based on our understanding of the physics, chemistry, ecology, and biology of the Earth system and are often based on well-established laws of nature. Human motivations and behaviors are complex, and thus socioeconomic models are frequently more approximate. In contrast to physical systems, diverse humans, corporate organizations, and other institutions are active, inventive, adaptive, and forward-looking in making decisions. In many areas of economics and the social sciences, however, there exist well-developed theory, substantial amounts of data, and substantial experience with computational methods. Interdisciplinary work is also moving forward, although it has far to go. In particular, socioeconomic models are not yet well integrated with biophysical models.

Because human motivations and behavior are complex, socioeconomic models of climate change effects necessarily will involve approximations.

Proposed responses to climate change range from controlling greenhouse gas emissions to mitigation and adaptation. In order to curb the increase in greenhouse gases, strategies such as carbon taxes, cap and trade, alternative fuels, voluntary agreements to reduce emissions, regulations including building codes and mandatory energy performance standards for equipment, and stricter emission standards for industry and transportation are under consideration. On a larger scale, we may ask how different responses affect energy supply and demand, the overall economy, the environment, demand for individual products and services, public health, and the vulnerability of the U.S. economy and infrastructure. In answering these questions, we must consider feedbacks from human responses to the climate system through, for example, changes in greenhouse gas emissions and surface albedos due to land use changes, which may in turn be affected by the adoption (or not) of low-carbon energy systems—developed, perhaps, as a result of investment in R&D programs designed to position us with options for the future.

Detailed global-scale socioeconomic models capable of capturing some or all of these processes will provide invaluable input to scientific understanding and decision making. In some cases, these models may serve simply to show that a particular response may not work as expected. In other cases, they may provide data that can guide the design of effective responses.

Because the models required to study socioeconomic aspects of climate change must necessarily encompass many aspects of human behavior, they can be expected to have important applications to numerous problems besides climate change, such as the following:

- Managing natural resources: water, forests, biodiversity, ecological systems, land use
- Addressing issues of energy security and exhaustible resources: oil, gas, coal, uranium

- Optimizing the supply and delivery of energy services
- Promoting economic development, income growth, poverty reduction
- Improving industrial productivity and competitiveness
- Investing in human capital: education, health care, job opportunities, social security
- Making tax policy more efficient and equitable
- Promoting common human values such as protecting the earth's environment

2. State of the Art

The use of computation-intensive modeling to study climate system-human system interactions is far from new. For several decades, researchers have developed integrated models for this purpose, on a variety of time scales and geographical scales [Edmonds et al. 1997; IMAGE team 2001; Sokolov et al. 2005]. Such integrated models couple different types of (sub)model, such as climate (atmosphere, ocean, cryosphere, biosphere, etc.); land use, vegetation and ecology (urban, agriculture, forests, biomass, wildlife habitat); environmental and resource changes; and socioeconomic (greenhouse gas emissions, health, political stability, feedbacks on economic activities and production costs).

Concurrently, researchers in economics and the social sciences have been applying computation to issues such as energy system optimization, investment analysis, improving human institutions such as our tax system and conditions for economic and social development, game theory strategies, and hedging uncertainties [Amman, Kendrick, and Rust 1996; Judd 1998]. The time scales involved are typically in the range of 5–40 years, depending on the decision problem, and the unit of analysis may be country, major region, industry sector, or income group.

In principle, integrated models can (and arguably should) incorporate detailed descriptions of all aspects of the social and economic subsystems. However, many of these problems are computationally extremely challenging even in isolation (e.g., when using game theory solution concepts, market equilibrium concepts that require the finding of fixed points, or stochastic dynamic programming formulations) and are not necessarily well understood. This situation, as well as a general lack of access to high-end computers within the socioeconomic modeling community, has led to integrated models being limited in various regards. They may sacrifice sectoral detail (e.g., see Figure 4.2), avoid consideration of certain subsystems, assume myopic behavior, ignore certain feedback effects and the notion of an overall dynamic equilibrium, or fail to adequately address issues of uncertainty. In short, they may be theoretical models focused on understanding a single dimension.

Many socioeconomic models currently are single dimensional, sacrificing detail or ignoring certain feedback effects; linking such models with climate models will involve major computational challenges.

To go from simpler, single-dimensional models to elaborate multidimensional models will require both adding detail and consistently specifying coupling among different systems.

Linking climate models to socioeconomic models is similar to what was done in the 1980s for the Acid Precipitation Assessment, in which a socioeconomic model set was assembled and linked to large-scale atmospheric models and used to address key questions of interest [NAPAP 1990]. The climate assessment would presumably be on a much grander scale.

For socioeconomic systems, the various parts of the economy operate within a global, macro environment. At the same time, the macro economy is made up of the sum of its parts. In order to help understand the behavior of the integrated socioeconomy and to estimate impacts due to changing conditions, large-scale computational modeling frameworks can be used. Over the past 25 years, economists, systems analysts, and quantitative social scientists have developed and expanded their models and computational techniques to solve specifications with many diverse elements, including;

- various household consumer groups;
- industrial sectors and business services;

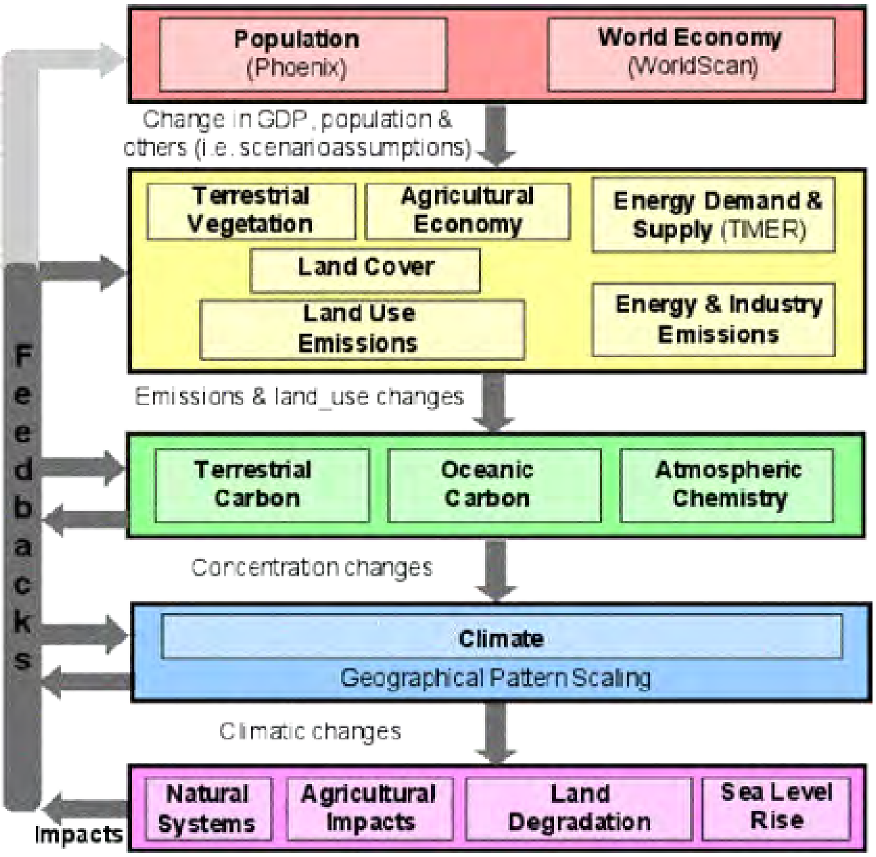


Figure 4.3 Models included in the IMAGE 2.2 integrated modeling system [IMAGE 2007].

- conventional and low-carbon or high-efficiency technologies and production activities;
- electricity, petroleum, gas, biofuels, solar, nuclear, and hydrogen energy forms;
- implications of carbon constraints on transportation and goods distribution;
- environmental emissions and impacts;
- developing supply and demand optimizations for sectoral resources;
- land, water, other resources, and climate feedback;
- regional and country divisions, including policy variations;
- taxation systems and financial markets; and
- public policy instruments to protect private property, mitigate environmental damages, and promote technology and investment.

Computational models are typically structured as hierarchies to be able to aggregate consistently from the micro to the macro and then to disaggregate income formation and macro environmental conditions to households and firms. Computational methods systematically calculate prices of goods and services up the levels of a hierarchy to final demands and allocate various quantities down the hierarchy, taking into account price, income, and other elasticities. The levels of the hierarchy include opportunities to improve economic efficiency through trade and substitution and penetration of advanced technologies. These production hierarchies are created for each sector in each region.

The structure of these socioeconomic models provides "hooks," or connection points, for the uptake of low-carbon resources and technologies such as biofuels and biorefining or energy efficiency. The comprehensiveness of the model allows estimation of the relative market roles that different technologies may play (which may depend on uncertain factors), and their end-to-end evaluation (life cycle analysis). Constraints may be added to the model to represent security considerations.

Applications for such a model include R&D investments under uncertainty (a current DOE-identified need); economy-to-climate-to-economy feedbacks; making strategic investments to position ourselves to maintain future options [Dixit and Pindyck 1992]; analysis of the interconnected issues of world oil security; optimizing energy demand, supply, and infrastructure choices; analyzing the impacts of carbon caps and taxes within the existing tax system structure; and more integration with issues related to climate, such as economic development, international interests, security, transportation, logistics networks, health, and vegetation/ecology.

Figure 4.1 illustrates one model of the climate system/ecological system/socioeconomic system designed for integrated assessments and policy evaluation. Another model, IMAGE, is shown in Figure 4.3 to illustrate the range of issues that may be included in such models. IMAGE, which was developed in the Netherlands, provides its detailed global land use database and its calculation of land use changes over a spatial grid. However, many of its other modules lack detailed representations and supporting data. IMAGE is one of 19 models that participated in Stanford University's Energy Modeling Forum study in climate change policy and assessment (EMF-21) [EMF Study-21 2006]. The AIM model from Japan, which also participated in the EMF-21 study, also has strong integrated assessment capabilities.

Much detailed work is under way, particularly in the United States, on land use and agriculture, including biocrops. The Polysis [English et al. 2006] and FASOM models [Murray et al. 2005] are leading examples, but a number of midwestern universities have major research and modeling efforts, financed by the U.S. Department of Agriculture. Polysis models the production of about 20 commercial crops by county in each state and groups land resources into 6 productivity categories. Its results are being used by the Energy Information Administration (EIA) in preparing Annual Energy Outlooks for biomass production. These results are also used by national laboratories, including Oak Ridge, Argonne, and the National Energy Technology Laboratory. The Climate Change Division at the U.S. Environmental Protection Agency uses the FASOM model for biofuels studies and soil carbon sequestration assessments.

Most climate-socioeconomic models have certain common ingredients. Goods and services are produced to meet human needs. The value of these goods and services sums to gross domestic product. The output of each production sector is based on a production technology employing labor, capital, energy, and, for agriculture sectors, land. Sectors trade materials and semi-finished goods among themselves. For example, car manufacturers purchase tires from the rubber sector. Some of this trade crosses international boundaries. Industrial capital stock is disaggregated in some models, such as the Argonne All Modular Industry Growth Assessment (AMIGA) model [Hanson and Laitner 2006], to better represent the substitution of capital for energy in end-use energy-intensive applications. Labor can also be disaggregated into skill levels and occupation groups. Electricity can be disaggregated into peak, shoulder, and intermediate demands. Other forms of purchased energy include natural gas, petroleum products, and hydrogen.

A popular trend in model integration is called "hybrid modeling," in which physical energy forms, resources, emissions, and technologies are integrated into socioeconomic models.

Cultural values and income, taking into account income distribution over population groups in the countries of the world, affect residential household expenditure patterns. For example, in both developed and developing countries, major energy-intensive purchases include light-duty cars and trucks. Thus, a socioeconomic model must track the stock, new sales, fuel demands, and carbon emissions of different types and sizes of light-duty vehicles. Fortunately, computer code exists to represent these aspects of society and the economy, and this code is easily scalable. Similarly, greenhouse gas emissions and petroleum fuel use by aircraft and heavy trucks and other forms of freight transportation are large and increasing, both absolutely and as a share of the total. Hence, the modeling of transportation and fuel technologies is of high importance [Wang 2001].

This integration of physical forms of energy, resources, emissions, and specific technologies into socioeconomic models has become known as "hybrid modeling" because of its

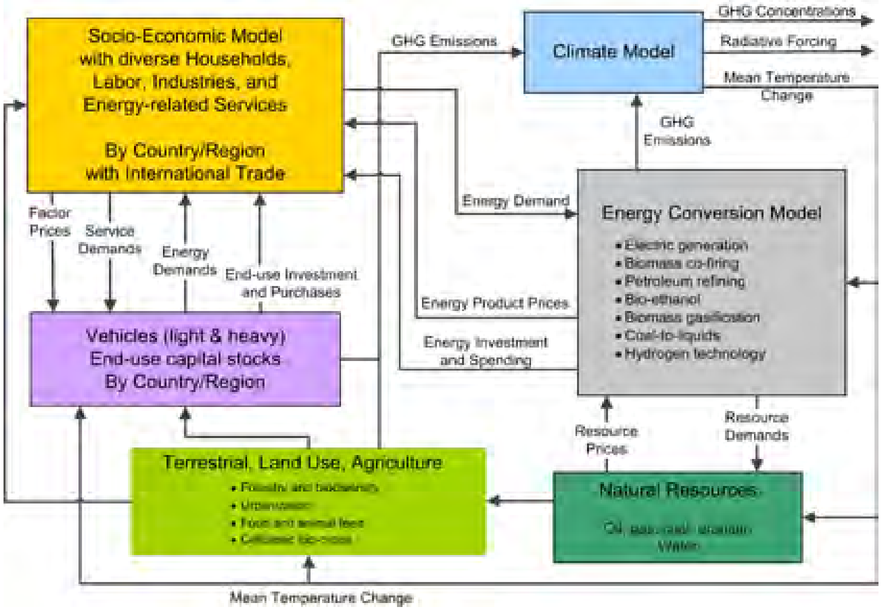


Figure 4.4 Typical configuration of a hybrid energy-economic model.

Exascale computing allows us to be far more ambitious in conceiving issue-oriented Earth system-human system models that can provide insights for policy makers.

combination of physical and observation-based statistical models. Figure 4.4, which illustrates a typical configuration for a hybrid energy-economic model, shows anthropogenic greenhouse gas emissions as an input to a climate model and feedback from climate to certain economic sectors.

Climate and climate-related changes then become inputs to analyzing these problems. As these decisions unfold over many years, there will also be feedback to the climate system, such as economic decisions that influence greenhouse gas emissions. Two examples of problems in this category are planning for a low-carbon energy system (including vehicles), taking into account technological synergies (e.g., biomass and petroleum-based fuels), and planning an R&D program that positions us with options for the future.

A hybrid energy-economic model, with similar features to those shown in Figure 4.4, has been used by the International Energy Agency (IEA) to prepare the World Energy Outlook (WEO) for 2006 [IEA 2007]. Running this model shows a difference between a business-as-usual growth scenario and an alternative scenario that included policies to promote renewable energy, energy efficiency, and nuclear power. The WEO also highlights the human dimension: about a billion people in the world still lack access to electricity. This kind of energy poverty illustrates the interplay between economic development, social goals, and environmental goals.

For WEO 2006, the world was divided into 22 country regions (Figure 4.2). Interfacing with climate models will require much greater geographical and spatial disaggregation. Part of this disaggregation can be accomplished by using modern spatial statistics [Stein 1999].

Socioeconomic models have been used extensively to generate sets of greenhouse gas emission scenarios to drive climate models [IPCC 2007]. A new round of global scenario simulations is starting in preparation for future IPCC Assessment Reports.

Outputs of most interest from the socioeconomic model include prices of energy and fuels, energy consumption and related greenhouse gas emissions, technologies adopted, sector outputs and final demands for goods and services, distribution of income, macroeconomic variables, resource management, and other measures of welfare or quality of life in societies.

3. Need for Exascale Computing

For many decades, researchers investigating climate change impacts have been forced by profound limitations in both computational capacities and data to grossly simplify their models and analyses. As a result, they have mostly ignored issues such as the following:

- integration of diverse complex systems and their relationships;
- incorporation of (often only partially understood) nonlinearities and thresholds;
- full representation of feedback effects; and
- meaningful analysis, reduction, and treatment of uncertainties

Exascale computing allows us to think differently about what is possible. While any useful model must incorporate simplifications, the anticipated availability of exascale computing encourages us to be far more ambitious in conceiving issue-oriented Earth system–human system models that shed insights on the complex realities that policymakers must deal with.

First, significant progress can be achieved in sectoral detail. Many dynamic economic models developed for climate studies are based on the computable general equilibrium (CGE) approach, the dynamic optimization approach, or the overlapping generations (OLG) approach, and most aim to model the dynamic paths of economic growth—but sacrifice sectoral detail. Some existing models describe energy-related sectors at regional and national scales (e.g., the U.S. National Energy Modeling System, NEMS [National Energy Modeling System 2003]) or global scale (e.g., the AMIGA model [Hanson and Laitner 2006; Laitner and Hanson 2006]). However, such models typically do not model economic dynamics in a consistent, modern manner. Most climate change issues require both sectoral detail and careful dynamic modeling.

Each of the three approaches used in modern economics has distinct advantages. Climate change issues challenge us to synthesize a new set of model frameworks suitable for advanced computational architectures. For example, when using the OLG framework, an addition will be needed to address the intergenerational time scales involved in issues related to climate change, and greater sectoral detail will be necessary to produce reliable results. Consistently coupling sophisticated sector models such as NEMS with new-generation dynamic CGE models will greatly improve the quality of climate change research.

Increased sectoral detail is required for coupling with climate models. Existing socioeconomic models are driven with simple climate models with limited geographic and temporal resolution. But understanding of both climate change impacts and feedbacks between socioeconomic systems and the Earth system requires the treatment of finer-grained interactions. In one early example involving a one-way coupling of a global climate model and NEMS, energy demands for increased cooling were found to outweigh savings due to reduced heating in the United States (Figure 4.5) [Hadley et al. 2006].

Increased detail is also required for the global water cycle. As climate changes, so do the distribution, intensity, and usability of precipitation and groundwater [Held and Soden 2006]. Such changes have significant socioeconomic impacts and feedbacks, not only on agriculture but also migration, wages, industrialization, and prices. Simulation of the global and regional hydrologic cycle and associated economic/management decision–making processes over the next 25–100 years represents a major challenge for socioeconomic modeling.

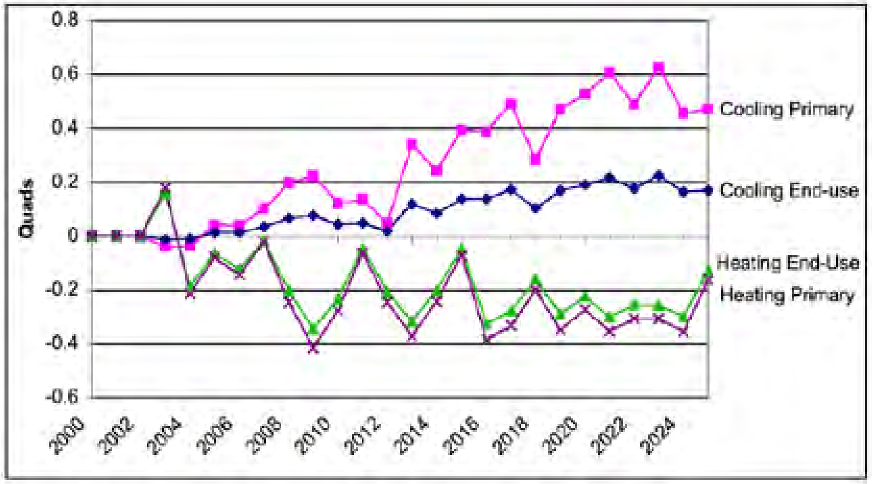


Figure 4.5 Changes in heating and cooling end use and primary energy under climate change [Hadley et al. 2006].

Larger computers will also advance socioeconomic models that describe, for example, consumer preferences, technology, and the rate of adoption of new technologies. The determination of the many parameters in these models presents us with many challenging computational tasks. We must develop suitable estimation methods with new types of more comprehensive data collection and calibration. Because of uncertainty in parameter values, we must also develop efficient methods, such as the use of low discrepancy sets or sparse grids, to explore the variables used to generate these parameters. This uncertainty will be even greater when we couple a socioeconomic model with an ESM that has its own uncertainties. Current methods for addressing large parametric system uncertainties, such as Monte Carlo and more general Bayesian methods, may need to be revisited and adapted for this potentially large-scale problem. This work must be performed at the component, subsystem, set of subsystems, and full system levels.

Progress is also possible in the treatment of social and political inputs. These inputs are modeled as offline forcings in the current generation of integrated models. Although some demographic and population-based studies have generated numerical modeling techniques for predicting future population growth and human migration patterns, much of this work is still ad hoc and in need of greater rigor. It will be in our interest to work toward developing more quantitative approaches, as in economics, and to collaborate with and strengthen institutes such as the International Institute for Applied Systems Analysis. It may also be useful to collabo-

Exascale computing challenges scientists to synthesize a new set of climate change models with greater sectoral detail and intergenerational time scales.

rate on setting up a similar institute dedicated to quantitative sociological modeling in the United States.

A fourth area of opportunity concerns uncertainty and risk. Some climate change data is known with relatively high confidence (e.g., average global warming), while other data is far less certain (e.g., regional climate change, frequency of extreme events). We need to study how uncertainty in the climate system propagates through economic models, both to enable quantification of uncertainty when reporting on outcomes and to enable evaluation of risk. Economic cycles are also subject to shocks. It is important to design mitigation strategies and policies that are robust under economic shocks and avoid costly policy changes. Researchers at the University of Chicago are undertaking computationally intensive research attempting to quantify the long-run risks associated with coupled climate-socioeconomic models. In their model, the researchers let household investors and firms base their expectations about climate-related risks, and who can most efficiently bear these risks, on actual risks calculated by the climate system model.

Mathematical issues such as convergence become important as model complexity increases.

A fifth area is the integrated computation of energy-economy hybrid models, in which detailed technology models are embedded in key areas such as power generation, petroleum refining, combined heat and power, and vehicle stocks.

A sixth area concerns solution techniques. For example, NEMS is solved by using a Gauss-Seidel process that is not guaranteed to converge. Mathematical issues such as convergence become more important as model complexity increases [Judd 1998]. Alternative solution techniques with better numerical properties are frequently more computationally demanding.

A seventh area in which increased computational power can enable new approaches is in the calculation of large ensembles of integrated climate-socioeconomic model runs, for sensitivity studies and to calculate probabilities of extreme events.

The increased power of exascale computing can enable new approaches for calculating the probability of extreme climate events.

Given the availability of detailed global socioeconomic models, a range of other complex computational problems can be tackled that call for exascale computing. The following are examples:

- Possible strategic behavior of countries, or coalitions of countries, to improve their positions relative to other countries
- Calculation of payoff matrices for different countries under different policy approaches, taking into account specific climate-related risk exposure of different countries
- Design of incentive systems to encourage cooperative behavior
- Computation of general equilibrium conditions
- Solution of stochastic dynamic programming problems that recognize that societies can wait for feedback information in order to make better decisions in the future about greenhouse gas abatement measures and other mitigation and adaptation decisions
- Representation of noncompetitive market situations such as Middle East oil production
- Optimal technology deployment with learning rates and stochastic outcomes (these are typically two point boundary value problems and may use optimal control methods with deterministic conditions)
- Other problem formulations involving complex human behavior

4. Major Challenges

Tackling the problems outlined in the preceding section will require significant advances, not only in economics, the social sciences, natural sciences, and computational sciences, but also in cross-discipline interaction methodologies and institutional arrangements. These advances will be essential if we are to make required progress at the interface between technological advances, cultural shifts because of such advances, and the response of these changes to macro-economic sensitivities and climate feedbacks in general.

Many climate change impacts and adaptive responses will apply at the regional or sub-regional scale. Thus, reductions in climate model uncertainty, as targeted by DOE's Earth system modeling program for the next five to seven years, will be crucial for efforts aimed at studying climate system–human system interactions.

Socioeconomic models such as those discussed earlier become large-scale, nonlinear systems when accounting for resource supply functions or constraints, production-possibility frontiers involving joint production, nonconvexities in the development and adoption of new generations of energy demand and supply technologies, and social behavior pattern responses. Progress is required in nonlinear optimization techniques, as suggested by this class of problems.

We have referred already to the need to obtain new types of quantitative and qualitative data as well as to assemble and use the data that currently exists effectively. New methods will be required for acquiring, accessing, evaluating, and integrating these data. We need efficient search and summary technologies; the ability to draw on data for estimation, including data coded and compatible with geographic information system (GIS) technology; and the ability to compare actual and simulated data rapidly. Again, a problem facing modelers of socioeconomic processes is the computational burden of applying the best statistical methods to existing data. This problem has led econometricians to search for "computationally light" methods. The best statistical methods applied to large data sets will require exascale computing power and present novel challenges in the utilization of large–scale computer architectures—but will produce far superior empirical results.

One approach to dealing with uncertainty is to perform multiple ensemble runs (parameter sweeps) with various combinations of the uncertain parameters. Since the space of parameters will be of high dimension, we will have to address the challenges of designing efficient parameter sweep methods for high-dimensional spaces. Recent advances in approximation theory and data mining methods, such as sparse grids, offer new approaches to this problem. Furthermore, recent results in approximation theory can be used to guide us in using exascale computing power to search for efficient methods.

Systems are so complex that mathematical analysis of convergence properties is impossible—we will be coupling Navier-Stokes in ocean and atmosphere with ice, physics, and economics. We need to develop methods for quantifying uncertainty in both data and models.

Progress is also required in various areas of socioeconomic modeling. One major challenge, which has been the focus of intense effort over the past decade, partly in response to climate change concerns, is capturing induced technical change in economic growth models. This is referred to as the endogenous technical change problem. We know that technology advance is not entirely accidental. Instead the need, or demand, for some technological solution increases the likelihood of it occurring. Economists, historians, and others are studying how successful technologies have gone through stages of discovery, innovation, early adoption by particular groups, and ultimately to market transformation. Empirical evidence continues to be collected on these stages of innovation. Innovation results in more available technology options and ultimately shifts sectors' production possibility frontiers in economic models.

Once we have an economic model specification of induced technical change, which may push the boundaries of performance potential, we may need to verify that physical laws of nature are not violated (e.g., mass and energy balance, the second law of thermodynamics). In other words, do engineering configurations exist consistent with the economists' models of technology advance?

Modern finance can also assist with economic models to represent the willingness of owners of buildings and production facilities to retrofit/replace processes with new ultraclean investments. Modern finance has developed an understanding how investment decisions tend to be made.

On the human dimension, there will be a real challenge to understand, model, and manage

Exascale computing power will enable modelers of socioeconomic processes to replace "computationally light" methods of data analysis, producing far superior empirical results.

rapid change that leads not only to major winners but also to extreme losses. Under historically normal rates of growth, markets have managed change reasonably well. Over time, "all ships tend to rise"—though not without increased inequality and persistent poverty on a relative, if not absolute, basis. These adjustments occur in the context of the usual economic forces: comparative advantage, specialization of production, gains-from-trade, and the resulting distribution of income over nations, population groups, sectors, and owners of capital/investors. However, it may be possible (or even likely) in a world experiencing significant climate change to have a shift in the ordering of fundamental economic forces, such as comparative advantages in certain production activities, resulting in rapid and notable losers (e.g., coal mining, island states, hurricane–prone regions, agricultural areas that turn arid, and emergence of certain groups less able to adapt).

5. Accelerating Development

A focused DOE R&D program can achieve significant advances in the state of the art in modeling the human elements of the climate system. Building on DOE expertise in climate and energy system modeling, and bringing to bear the latest methods in economics, quantitative techniques in behavior and decision theory, and modeling and HPC tools, such a program can develop a deepened understanding of the technical, economic, political, and social issues that underpin the climate change challenge.

Achieving this goal will require a sustained, large-scale program aimed at applying computational science methods, with the objective of creating, within a decade, the tools and methodologies needed for the quantitative study of questions introduced earlier. This program should aim to create a highly integrated, time-dependent modeling system that encompasses detailed micro socioeconomic data and models in a comprehensive, integrated global framework. This system will allow for treatment of socio-demographic groups and the marketplace, including producers, consumers, and intermediaries, in unprecedented detail.

A focused R&D program in socioeconomic modeling must include basic research in foundational issues such as spatial statistics, multiscale modeling, and coupling of micro-activity and the biosphere.

We envision the following specific tasks:

- Construct a comprehensive suite of models of unprecedented geospatial and temporal detail, with comprehensive error analysis on the representation, some based on existing models, some entirely new.
- Leverage state-of-the-art climate modeling activities (e.g., SciDAC) to include economic prediction models under alternative climate regimes.
- Perform basic research into such foundational issues as spatial statistics, modeling of social processes, relevant micro-activity and biosphere coupling issues, and relevant mathematical challenges, such as multiscale modeling.
- Assemble and perform quality control of extensive data collections—much from existing sources, but also much from new and unconventional sources.
- Perform comprehensive and detailed validation of both individual models and large model systems.
- Develop novel, robust numerical techniques and high-performance computing approaches to deal with the expected orders-of-magnitude increase in model complexity.
- Conduct a wide range of application studies aimed at both validation and application.
- Develop education programs aimed at training the next generation of computational economists and other social scientists, including not only formal training programs but also web-based modeling and simulation tools that allow widespread access to the new models and their results.

Partnerships with international scientists and organizations will be important to develop the global-scale models and obtain the data needed for realistic assessment.

Figure 4.6 illustrates some of the advances that we expect as we move to future petascale and exascale computers.

Further study is required to determine the best organizational structure for such a program. We outline here one approach that we feel holds promise. In this approach, we define, as one dimension of the organizational matrix, five teams, each focusing on solving problems and issues in a specific area (and also interacting and sharing information and progress):

1. Modern economic computation approaches, involving greatly expanded micro databases to capture diversity in decision-making agents and strategic incentives that different country players may exercise.

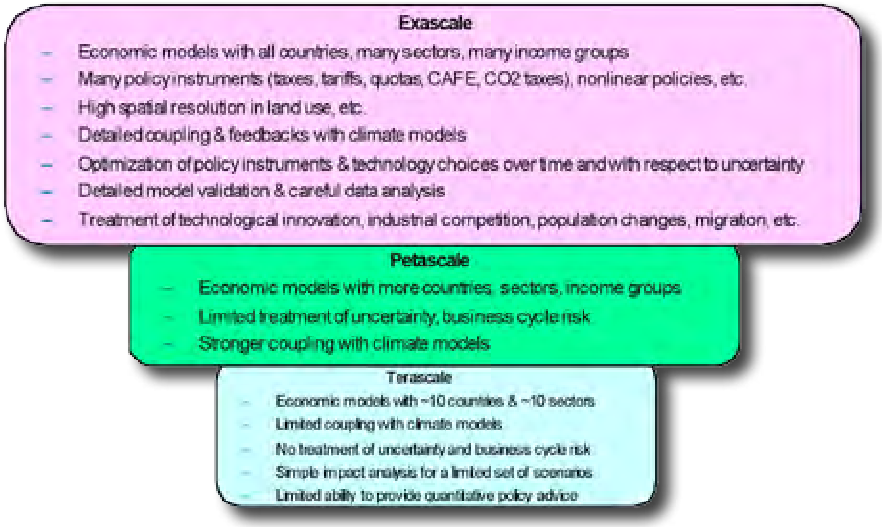
2. Consistent socioeconomic modeling with many regions/sectors and with technology rich specifications that capture technology-system efficiencies, synergies, and diversities and represent least-cost dynamic transition pathways to a low-carbon, adaptive global economy

3. Social science modeling to capture the human dimension of climate-induced changes in greenhouse gas emissions, mitigation, and adaptation

4. Integrated assessment, bringing in climate and natural resource models (water, agriculture, land use, forest changes, biodiversity, health and disease) and their coupling to socioeconomic models

5. Advanced optimization and computation methods needed for integrated assessment, solving large-scale nonlinear systems, and addressing approaches for accounting for uncertainty

As progress is made in all five of these areas, and as the various modeling teams better understand how to consistently couple their modeling areas together, we will move to a common large-scale integrated model system.



Exascale

- Economic models with all countries, many sectors, many income groups
- Many policy instruments (taxes, tariffs, quotas, CAFE, CO2 taxes), nonlinear policies, etc.
- High spatial resolution in land use, etc.
- Detailed coupling & feedbacks with climate models
- Optimization of policy instruments & technology choices over time and with respect to uncertainty
- Detailed model validation & careful data analysis
- Treatment of technological innovation, industrial competition, population changes, migration, etc.

Petascale

- Economic models with more countries, sectors, income groups
- Limited treatment of uncertainty, business cycle risk
- Stronger coupling with climate models

Terascale

- Economic models with ~10 countries & ~10 sectors
- Limited coupling with climate models
- No treatment of uncertainty and business cycle risk
- Simple impact analysis for a limited set of scenarios
- Limited ability to provide quantitative policy advice

Figure 4.6 Anticipated model improvements resulting from a major DOE program in detailed global socioeconomic modeling.

6. Expected Outcomes

A focused DOE program in this area is expected to achieve a significant improvement in understanding important questions relating to socioeconomic aspects of climate change. This improved understanding can improve the effectiveness of U.S. and international policy responses to climate change, the ability of U.S. industry to engage effectively in developing and "productizing" climate change solutions in the areas of low-carbon energy supply and demand, and the ability of the United States to minimize the vulnerability of its economy and energy infrastructure.

DOE efforts in this area will also have a significant impact on the development of computational tools, expertise, and understanding in economics and social sciences as a whole. Other anticipated results include the following:

- Socioeconomic models will become major users of DOE supercomputers, both individually and within integrated models linking socioeconomic models and ESMs.

- DOE efforts in the simulation of energy production, distribution, and use will be accomplished via exascale computational platforms.

A pilot study would help demonstrate the viability of merging socioeconomic theory within Earth system model frameworks and of simulating testable hindcast economic responses to climate variations.

- Various DOE-centric economic climate-energy scenarios will be provided to policy makers.

- DOE will become a place to which other parts of the government go to increase understanding of these issues and relevant policy information.

7. Required Investment

The creation of an integrated global socioeconomic model, the validation of model components, and applications to the understanding of human actors and their coupling to the climate system involve a massive grand challenge, requiring an interdisciplinary team of arguably unprecedented scale and scope. A rough estimate of the required resources would be $100M/year over 10 years for human resources—in addition to the hardware resources needed for storage and simulation.

This work would be organized and divided among several institutions, representing the multiple disciplines and expertise involved. Regular communication, coordination, and publication of interim results would be required.

Fortunately, there already exist years of accumulated research experience and preliminary model development activities upon which to draw.

8. Major Risks

The proposed activity represents a significant departure from current practice in the climate change and socioeconomic modeling communities.

Perhaps the biggest challenge is that of uncertainty. The ESMs used to quantify climate change benefit from large quantities of data and (mostly) well understood physical laws. Yet despite enormous complexity and large amounts of computing power, their results exhibit considerable uncertainty, particularly when it comes to regional effects. Human factors are in many regards less well understood and are heavily parameterized with empirical relationships. In both cases, the only responsible approach is to examine a range of plausible alternative scenarios and search for policies that are robustly successful. Furthermore, any analysis must consider how human institutions and agents will react to those uncertainties.

As with any cutting-edge simulation, there is a need to clearly portray the uncertainties and associated levels of credibility. A pilot phase would help to demonstrate the viability of merging socioeconomic theory within ESM frameworks, simulating testable hindcast economic responses to climate variations, and applying testable hypotheses on the expected outcomes from a new human/climate Earth system approach.

The most compelling reason to create and nurture a program such as that described here is the risk of not acting. The challenge of merging the diverse economic and geophysical models is significant but must be addressed if we are to understand and predict future socioeconomic systems in a warming world.

References

H. M. Amman, D. A. Kendrick and J. Rust, eds. (1996), Handbook of Computational Economics, vol. 1. Elsevier.

A. Dixit and R. Pindyck (1992), Investment Under Uncertainty. MIT Press.

J. Edmonds, M. Wise, H. Pitcher, R. Richels, T. Wigley and C. MacCracken (1997), An integrated assessment of climate change and the accelerated introduction of advanced energy technologies: An application of MiniCAM 1.0, Mitigation and Adaptation Strategies for Global Change 1(4) 311–339.

EMF Study-21 (2006), Special issue on multi-greenhouse gas mitigation and climate policy, EMF Study 21. The Energy Journal.

B. C. English, D. G. De La Toore Ugarte, K. Jenson, C. Hellwinckel, J. Menard, B. Wilson, R. Roberts and M. Welsh (2006), 25% renewable energy for the United States by 2025: Agricultural and economic impacts. Department of Agricultural Economics, University of Tennessee.

S. W. Hadley, D. J. Erickson III, J. L. Hernandez, C. T. Broniak and T. J. Blasing (2006), Responses of energy use to climate change: A climate modeling study. Geophys. Res. Lett., 33 (L17703).

D. A. Hanson and J. A. Laitner (2006), Technology policy and world greenhouse gas emissions in the AMIGA modeling system, The Energy Journal (Special Issue on Multi-Greenhouse Gas Mitigation and Climate Policy).

I. Held and B. Soden (2006), Robust responses of the hydrological cycle to global warming. Journal of Climate 19:5686–5699.

IEA (2007), International Energy Agency. World Energy Outlook, www.worldenergyoutlook.org.

IMAGE 2.2 Model Flow Diagram (2007), www.mnp.nl/image/model_details.

IMAGE team (2001), The IMAGE 2.2 implementation of the SRES scenarios: A comprehensive analysis of emissions, climate change and impacts in the 21st century, RIVM CD-ROM publication 481508018, National Institute for Public Health and the Environment, Bilthoven, the Netherlands.

IPCC (2007), Intergovernmental Panel on Climate Change (IPCC) 4th assessment report, www.ipcc.ch, 2007.

K. Judd (1998), Numerical Methods in Economics, MIT Press.

J. A. Laitner and D. A. Hanson (2006), Modeling detailed energy-efficiency technologies and technology policies within a CGE framework, The Energy Journal (Special Issue on Hybrid Modeling of Energy-Environmental Policies: Reconciling Bottom-up and Top-down).

A. J. McMichael, D. H. Campbell-Lendrum, C. F. Corvalán, K. L. Ebi, A. K. Githeko, J. D., Scheraga and A. Woodward, eds. (2003), Climate change and human health: Risks and Responses. World Health Organization, Geneva.

B. C. Murray, A. J. Sommer, B. Depro, B. L. Sohngen, B. A. McCarl, D. Gillig, B. D. Angelo and K. Andrasko (2005), Greenhouse gas mitigation potential in US forestry and agriculture. EPA Report 430-R-05-006.

NAPAP Integrated Assessment Report. National Acid Precipitation Assessment Program (NAPAP), Office of the Director, 1990.

The National Energy Modeling System: An overview (2003). Energy Information Administration, U.S. Department of Energy.

A. P. Sokolov, C. A. Schlosser, S. Dutkiewicz, S. Paltsev, D. W. Kicklighter, H. D. Jacoby, R. G. Prinn, C. E. Forest, J. Reilly, C. Wang, B. Felzer, M. C. Sarofim, J. Scott, P. H. Stone, J. M. Melillo and J. Cohen (2005), The MIT Integrated Global System Model (IGSM) version 2: Model description and baseline evaluation. MIT.

M. L. Stein (1999), Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Sustainable Bioenergy: A Framework for Decision Makers (2006), UN Energy.

M. Q. Wang (2001), Development and use of GREET 1.6 fuel-cycle model for transportation fuels and vehicle technologies (2001), Center for Transportation Research, Argonne National Laboratory.

5 Astrophysics

To understand our universe and our place in it will require that we understand the universe on all scales, from the evolution of the universe as a whole and the formation of its largest-scale structures to the formation and evolution of compact objects, only one one-hundredth the size of Earth but with mass greater than our own sun. The largest scale structures in the universe and compact objects differ in size by 22 orders of magnitude. Understanding the universe on each scale presents exascale computing challenges; while exascale computing will allow increased connections to be made between scales (e.g., connections between large-scale structure formation and galaxy formation), beyond exascale computing lies the challenge of tying these scales together more broadly in order to develop a seamless description that can take us through the history of the universe, from the Big Bang to the development of conscious life peering back on this unfolding history.

1. State of the Art

Our discussion of the state of the art in simulation of the universe covers all scales: from large-scale structure to galaxies to stars and finally to compact objects.

1.1 Large-Scale Structure

We begin with the challenge to simulate and understand the formation of the largest structures in the universe, the lacelike structures that thread the universe and are composed of clusters and "superclusters" (clusters of clusters) of galaxies (see Figures 5.1 nd 5.2). The formation of these structures in the early universe is inexorably tied to the evolution of the universe as a whole, which in turn depends on the nature of the mysterious dark energy that permeates it [Hu and Dodelson, 2002]. Dark energy is the vacuum energy field responsible for the accelerating rate of cosmic expansion. Understanding the nature of dark energy has been declared the most important and fundamental problem in the physical sciences.

Funding agencies are currently considering several major ventures—the NASA/DOE/NSF Joint Dark Energy Mission (JDEM), the Large Synoptic Survey Telescope (LSST), and the Square Kilometer Array (SKA), each in the range of $0.3 billion to $1 billion—as well as a variety of lower-cost pathfinder missions. The generation of mock catalogs from petascale cosmological simulations will be used to demonstrate the feasibility of these missions. If one or more of the missions are funded, exascale simulations will be needed in the next decade to help pull out the dark energy parameters from the observations.

1.2 Galaxies

From collections of galaxies we move to the scales of individual galaxies and their interactions. One of the most fundamental questions about the universe—What is the nature of its major constituents?—remains shrouded in mystery. Visible matter is only 4% of the content of the universe. This situation poses great challenges for the computations required to constrain dark matter and dark energy properties from the observed distribution of galaxies. Simulations to follow the formation of our galaxy in sufficient detail to compare with observational data from the James Webb Space Telescope (JWST) and LSST will require dynamic ranges of order 10,000 in space and time. A simulation with enough resolution to accurately model the visible properties of individual galaxies and yet contain a fair sample of the universe suitable for compari-

Understanding the universe — from the smallest objects to those 22 orders of magnitude larger — requires exascale computing, and more.

son with large surveys will require exaflop-weeks of computing.

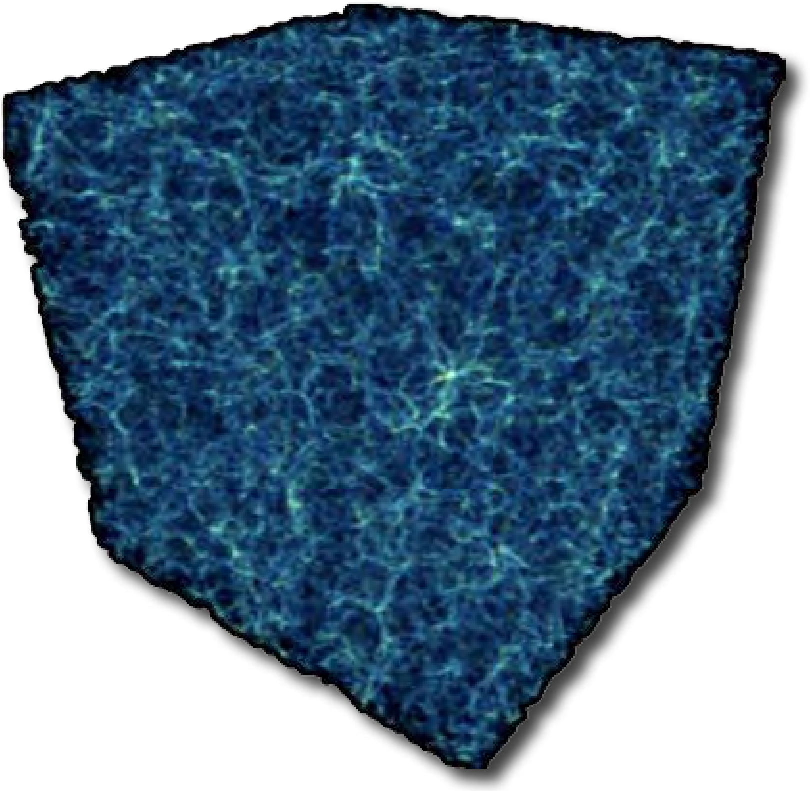


Figure 5.1 Structure formation in the Santa Fe Light Cone [Hallman et al. 2007]; 100 square degrees of sky incorporating 700 megaparsecs were simulated with the ENZO code.

An understanding of individual galaxies cannot be obtained, of course, without an understanding of their cores. About 1% of galaxies in the local universe host powerful central radiation sources known as active galactic nuclei (AGNs). These sources produce emission across a wide range of photon energies (from radio to X-rays) and are usually quite variable in their output. They are often associated with bidirectional outflows of material moving at relativistic velocities. AGNs appear to have been much more common several billion years ago, when galaxy formation was at its peak.

The most widely accepted model for AGNs proposes that they are powered by accreting supermassive (more than 100 million solar masses) black holes surrounded by obscuring tori of dust and gas [Lynden-Bell 1969]. Since spiral galaxy rotation curves suggest that most galaxies harbor such black holes at their centers, whether or not they host an AGN, the fact that AGNs are rare today is interpreted to mean that most black holes are no longer accreting material. This interpretation further suggests that AGN activity may be closely tied to the assembly of galaxies. AGNs within clusters of galaxies provide a particularly revealing look at how AGNs interact with their host environments. Such AGNs create enormous cavities ~30,000 light years in size.

Computational study of the AGN-environment interaction is at present limited to simulations that cover only a small portion of the dynamical range and crudely model the effects that arise outside that range. Questions that we would like to answer include the following: What turns AGNs on and off? How do AGNs affect the physical state of the intracluster medium? What role do AGNs play in galaxy and cluster assembly? How do the central black holes form in the first place?

The convergence of these observational datasets at the same time that exascale computing makes it possible to elaborate our theoretical predictions for this problem promises to significantly advance our understanding of the astrophysical questions posed above.

1.3 Stars

We owe our existence to stars for a number of reasons, not the least of which is that our own sun is a star. The elements necessary for life are made in stars during their lifetimes and during their deaths as supernova explosions. It is through violent phenomena such as dramatic stellar outflows during a star’s life or supernovae that these elements are dispersed into the interstellar medium, peppering the soup from which our solar system formed billions of years ago.

In particular, half of the elements heavier than iron are made in stars through what is called the s-process (slow neutron capture process). These elements are then injected into the interstellar medium through the expulsion of the outer layers of the star, forming planetary nebulae. Both the production and transport of these heavy elements are sensitive to 3D effects, including the turbulent mixing of stably stratified material into the convection zone, the dredge-up of processed material into the envelope, and the global convection in the outer envelope of the star and the expulsion of this envelope.

The cause of the repeated envelope expulsion in stars is thought to be the periodically occurring flashes of the helium-burning shell within the central, Earth-sized core of the star. These can be likened to a global storm, and the full sequence of helium shell flashes can be likened to stellar climate. With petascale computing, high-quality simulations of portions of the space-time domain for one such helium shell flash could be performed. With exascale computing, we may be able to simulate the entire helium shell flash, with its 2-year duration, using validated statistical models of phenomena operating on smaller length and time scales. Exascale computing will allow full-scale simulation with validation quality for the entire helium shell convection zone for time scales of hours.

Aside from understanding the complete phenomenology of thermonuclear supernovae as a goal in itself, the importance of a complete picture of the explosion mechanism extends to a fundamental question in cosmology as well: What are the nature and effect of dark energy? Only when we fully understand the explosion mechanism can we validate the use of Type Ia supernovae as “standard candles” for use in measuring the size and expansion history of the universe. Current Type Ia surveys rely on purely empirical corrections to determine the intrinsic brightness of observed supernovae. As observations extend to higher and higher redshifts, we must understand how evolutionary effects (e.g., metallicity) might contribute to the intrinsic brightness of the events. This level of calibration is required to make use of data to be obtained by JDEM, which is designed to detect thousands of Type Ia supernovae and thereby put real constraints on the equation of state of the mysterious dark energy that appears to make up 70% of our universe.

Massive stars more than ten times the mass of our sun evolve for millions of years and then die in a matter of hours in spectacular stellar explosions known as core-collapse supernovae [Mezzacappa 2005; Janka et al. 2006]. Such supernovae are an important link in our chain of origin from the Big Bang to the present day. Understanding how they occur is key to understanding how we came to be in the universe. They are the dominant source of elements in the periodic table between oxygen and iron, and there is growing evidence that they are indeed responsible for producing half of the elements heavier than iron (the other half coming from the s-process in stars, discussed above). Moreover, they are the most energetic explosions in the universe, and there is now an indisputable connection between “peculiar” hyper-energetic core-collapse supernovae, also known as “hypernovae,” and one of two classes of gamma ray bursts (GRBs) in the universe [Woosley and Bloom 2006]. Both phenomena occur under a common umbrella of massive stellar core-collapse. And core-collapse supernovae are among the events expected to produce gravitational waves—ripples in the fabric of

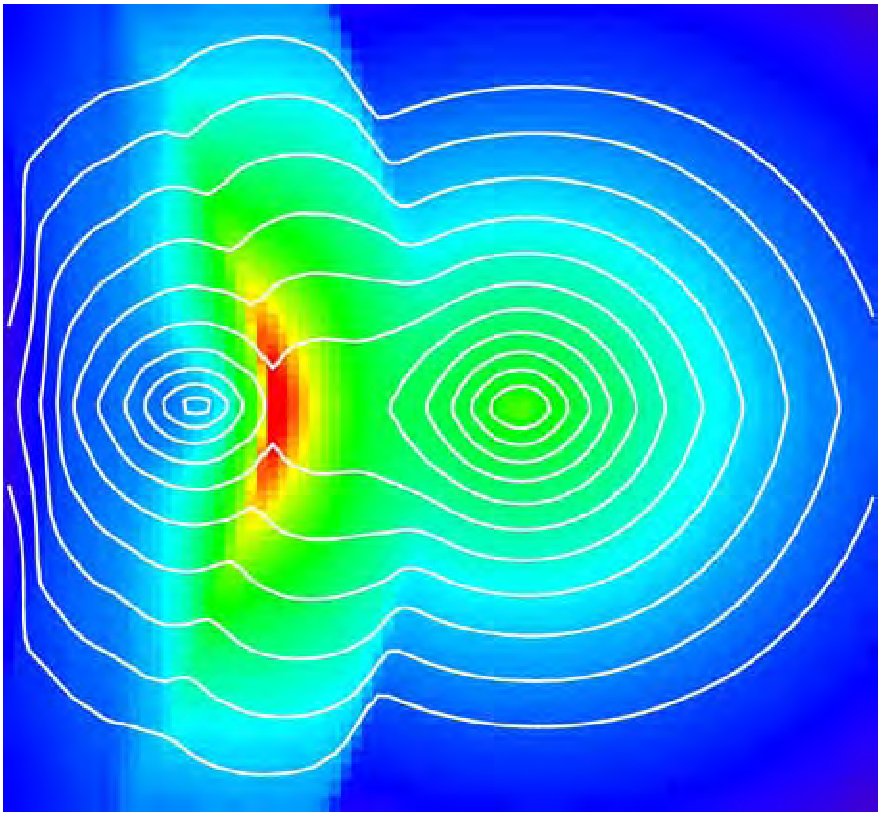


Figure 5.2 Simulated X-ray observation of a simulated merger between clusters having a mass ratio of 1:3. Contours indicate X-ray surface brightness, while the colors indicate X-ray temperature, with blue indicating cooler gas and red indicating hotter gas. (P. Ricker)

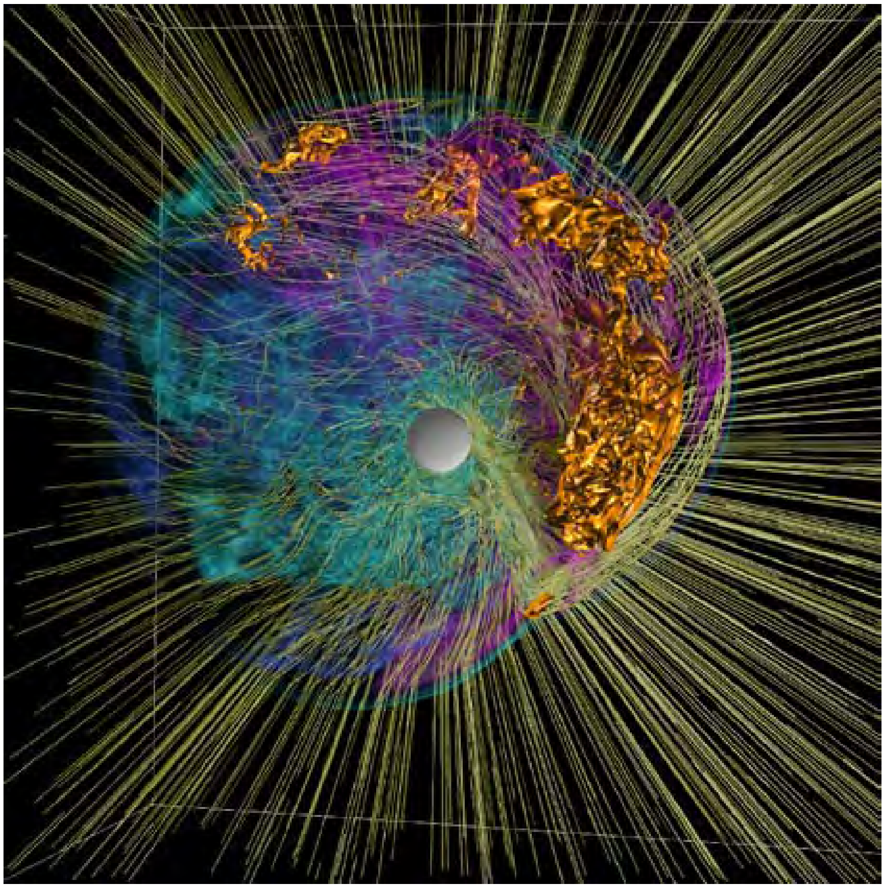


Figure 5.3 Development of the computationally discovered (SciDAC Terascale Supernova Initiative) instability of the core-collapse supernova shock wave, shown in a snapshot from a 3D simulation by John Blondin (NCSU) and Anthony Mezzacappa (ORNL). The visualization was performed by Kwan-Liu Ma (UCD). The instability leads to growing deformations (away from spherical) of the shock wave, which is represented by the surface in this image. The deformations in turn lead to circulating flow below it. Two strong, counter-rotating flows are formed. Streamlines in this image highlight one flow moving clockwise just beneath the shock surface and a second, deeper flow moving counter clockwise just above the proto-neutron star surface. Moreover, the inner flow is capable of spinning up the proto-neutron star, perhaps explaining the origin of pulsars [Blondin and Mezzacappa 2007]. The spins generated in these simulations are consistent with observations of young pulsars.

space—in the galaxy, the detection of which will mark a historic event.

The temperatures to which the core-collapse supernova shock wave (see Figure 5.3) heats the ejecta from such supernovae make ultraviolet observations a very useful way of studying the enriched composition of recent ejecta. Examples include observations performed in the past with the International Ultraviolet Explorer (IUE), more recently with the Hubble Space Telescope (HST), or currently with the Far Ultraviolet Spectroscopic Explorer (FUSE). Indeed one objective of the FUSE mission is “studies of nova and supernova explosions and their remnants, to test theories of heavy element nucleosynthesis and study how supernova shock waves heat the interstellar gas.” With missions such as NASA’s SWIFT GRB Mission, these observations can be extended to include the X-ray region. Furthermore, observations of supernova remnants, such as the Chandra X-Ray Observatory observations of Cassiopeia A and SN1987A, also provide detailed data on the composition and distribution of heavy elements in supernova ejecta, data which must be explained by explosion models.

Observations of gamma-ray lines are an excellent way to determine not just the elemental but also the isotopic production of supernovae. They are therefore of particular interest, both to supernova theory in general and to supernova nucleosynthesis. Observations such as the Compton Gamma-Ray Observatory (CGRO) measurement of the ratio of ^{57}Ni to ^{56}Ni in SN1987A or the detection of ^{44}Ti in the Vela and Cassiopeia A SN remnants probe the deepest layers of the supernova ejecta, just above the mass cut. Only the neutrino and gravitational wave signals, which follow the evolution of the proto-neutron star, provide information from deeper within the explosion. Therefore, gamma-ray line observations, past (CGRO) and future (Integral, the International Gamma-Ray Astrophysics Laboratory), coupled with supernova models that include realistic nucleosynthesis studies, will greatly constrain explosion models.

1.4 Compact Objects

Neutron stars and stellar mass black holes are the remnants of the death throes of massive stars, discussed above. They are among the smallest objects in the universe but are examples of nature at its most extreme. They are important in their own right, as key components of some astrophysical systems, such as binary systems comprising a star and neutron star or two neutron stars, and as laboratories for fundamental physics. Neutron stars are approximately 10–15 km in radius, with a mass approximately one to two times the mass of our sun.

We stand at the threshold of the birth of a new type of astronomy: gravitational wave astronomy. Observatories such as the NSF-funded Laser Interferometric Gravitational Wave Observatory (LIGO) are poring through the sky in search of gravitational waves from astrophysical sources. Three key sources have been identified: the merger of two neutron stars in orbit around one another, the merger of two black holes in orbit around one another, and core-collapse supernova explosions. All three are expected to be seen by LIGO for galactic events.

In the past few years, the community has produced the first stable and accurate simulations of binary black holes (BBHs) [Baker et al. 2006; Campanelli et al. 2006]. However, the complexity of these simulations increases with the addition of matter (i.e., nonvacuum scenarios). Binaries composed of a black hole and a neutron star (BHNS) or two neutron stars (BNS) are also of great interest and the subject of active study (see Figure 5.4). Current matter simulations in numerical relativity are beginning to add microphysics such as realistic equations of state, magnetic fields, photon and neutrino radiation transport, and nuclear networks. Exascale computers will be needed to describe with accuracy systems as physics-rich as BHNS and BNS. These simulations are essential for the understanding of matter at extreme densities.

2. Major Challenges

Given the need to resolve all relevant scales in the astrophysical environments described in section 1 and the ability to do so with the promise of exascale computing, given the increasing use of adaptive mesh refinement in these simulations, and given the generally long run times anticipated to cover the evolution of the above systems in both space and time, two issues emerge:

- An increased need for dynamic load balancing and smart algorithms to provide this capability

- An increased need for fault tolerance, that is, for both efficient parallel algorithms that minimize the time to solution and thereby mitigate this issue and fault-tolerant solution algorithms.

The data anticipated from both observational facilities and computational simulations will be enormous. For example, 100 PB of data are expected from the LSST, and simulations of core-collapse supernovae at the exascale are expected to produce ~100 PB of data per simulation over a period of a few months. The analysis and visualization of these data,

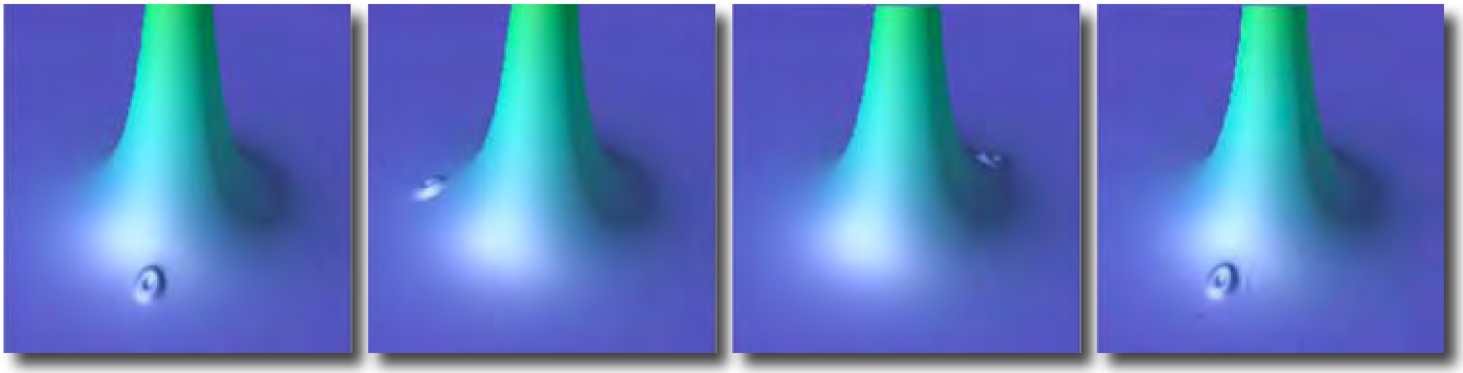


Figure 5.4 Snapshots of the space-time evolution of a neutron star orbiting a massive black hole [Sopuerta, Sperhake, and Laguna 2006].

with an eye toward scientific discovery, will present great challenges. Data analysis algorithms will need to share the same efficiency and scalability of the algorithms used in the original numerical simulations that produced the simulation data, and given the size of the data sets, the analyses themselves will likely require exascale computing resources, or at least dedicated computing resources with significant computational capabilities. The infrastructure surrounding the exascale platforms must provide sufficient capabilities for data archiving, data access, networking, and visualization for both local and wide area networks. This will further accentuate the need for fault-tolerant approaches to data storage and access (e.g., multisource, multistream approaches), latency-tolerant approaches to remote visualization, and the like.

As the data volume from observational facilities expands, the possibilities of integrating "sky truth" into simulations become more and more appealing. Nevertheless, the consensus suggests that the more likely mode of inclusion will be more widespread construction of artificial observations—that is, simulation data convolved with instrument characteristics to produce a representative notion of what the simulated object would "look like" to an observer. This construction has two immediate consequences. First, the data volume from the simulations actually tends to expand markedly when subjected to this type of a posteriori analysis, in contrast to more traditional forms of data analysis (often termed "reduction") that produce a more compact set of data. Second, this wedding of data and experimental uncertainty is the nexus in which validation is ultimately performed for astrophysical simulation. As it is the observations that will tell us whether our simulations solve the proper set of equations (i.e., the equations that describe nature in the settings we investigate), having a complete view of the interplay of observer and observed phenomena is crucial to actually realizing the goal of validated astrophysical simulation.

If we are to understand our accelerating universe, we must first faithfully simulate a good fraction of it computationally.

3. Advances in the Next Decade

The computational astrophysics community has a significant tradition in computing at scale and in using existing architectures to do so. Given this history, the overriding sense of the community is that existing approaches will scale to the exascale and that exciting advances in science will be feasible within the next decade in each of the four areas discussed in Section 1: large-scale structure, galaxies, stars, and compact objects.

3.1 Large-Scale Structure

At present, the only means we have to investigate dark energy is to use the new observational surveys of the universe on a vast scale. The interagency (DOE/NSF/NASA) Dark Energy Task Force (DETF) has identified four promising observational techniques for measuring the so-called dark energy equation of state (the relationship between important quantities that characterize the dark energy, such as pressure and energy density) [Albrecht et al. 2006]. Three of the four will require the guidance and interpretation of cosmological simulations of the large-scale distribution of galaxies and galaxy clusters of high precision (1%), high physics fidelity, and performed on an unprecedented scale (a good fraction of the observable universe). In essence, if we are to understand our accelerating universe, we will first have to faithfully simulate a good fraction of it on a computer. While this sounds preposterous, it is in fact what is needed to complement and to derive the benefit from the observational surveys that will likely be moving forward.

3.2 Galaxies

JWST is the designated successor to the HST as a "great observatory." It was the astronomical community's top priority in the last decadal survey and has an estimated cost of $4.5 billion. It is specifically designed for understanding the formation of the first stars and galaxies, measuring the geometry of the universe and the distribution of dark matter, and investigating the evolution of galaxies. Comparison with theoretical models will require simulations with enough fidelity to match the JWST observations of the earliest galaxies as they form to the objects they become today.

The LSST is another national priority, with an estimated cost of $300 million. It will constrain the nature of dark energy by following the evolution of structure, and the nature of dark matter by mapping strong gravitational lensing in clusters of galaxies.

Both of these observational programs require theoretical predictions about the clustering of dark matter and the distribution of galaxies in order to fully constrain the dark energy/dark matter models.

For the timescale on which exascale computing is expected to become a reality (2015–2020), several exciting observational developments promise to dramatically enhance our knowledge of the structure of AGNs and the properties of their environments. In the X-ray, the Constellation-X telescopes will provide spectra and images of AGNs and cluster cores that will allow us to study the dynamics of the innermost accretion disks of AGNs and the structure of the intracluster medium. In the radio, SKA and the Enhanced Very Large Array (EVLA) will produce sensitive maps of radio emission from the enormous cavities produced in clusters with AGNs, providing detailed information about the relativistic plasmas and magnetic fields that fill them. In the optical, LSST will produce numerous (more than 10 million objects) samples of AGNs, stretching through most of the epoch of galaxy formation and telling us how the AGN-environment interaction has shaped the luminosity and variability of AGNs during this critical period. The coming generations of very large optical telescopes, such as the Giant Magellan Telescope (GMT), will also make it possible to image the central regions of nearby AGNs with a resolution hitherto inconceivable.

3.3 Stars

Stars like our own sun end their lives as white dwarfs: compact remnants held up against their own self-gravity by electron degeneracy pressure. If the white dwarf is in a binary, the companion star can shed mass onto its more evolved neighbor. This increase in mass can lead to an uncontrolled thermonuclear explosion, completely disrupting the white dwarf in a cataclysm visible across most of the observable universe. The details of the explosion mechanism of these thermonuclear supernovae (identified with the spectroscopic Type Ia label) are still an unsolved problem. The interplay of rapid nuclear burning and strong gravity in the interior of the white dwarf results in a combustion problem that requires the resolution of flamelets roughly the width of a finger while simulating [Gamezo et al. 2005; Calder et al. 2007] an object the size of the Earth. Ultimately, such simulations must also include a detailed understanding of the nuclear species formed in the event and their spectroscopic signatures. Exascale computing will enable simulations with resolutions down to the Gibson scale (the length scale where turbulent motion is effectively smoothed by the propagation of the nuclear flame) with definitive prescriptions for nuclear energy release and the associated nucleosynthesis.

Core-collapse supernovae are driven in part, or perhaps largely, by an intense flux of radiation in the form of nearly massless particles known as neutrinos that emerges from the proto-neutron star at the center of the explosion. The enormous energy contained in the neutrino radiation field must be modeled accurately in order to model the much less energetic phenomenon of the supernova explosion itself. This modeling in turn will require the solution of the six-dimensional (6D) neutrino transport equations, the solution of which will give the distribution of neutrinos in angle of propagation (two dimensions) and energy (one dimension) at each 3D spatial location. Petascale platforms are already required for 3D simulations with multiangle, multienergy neutrino transport at moderate resolution. Simulations with the spatial resolution required to properly model other critical aspects of the explosion dynamics—for example, the evolution of the stellar core magnetic fields and their role in generating the supernova—will require much higher resolution, which in turn will require exascale computing, particularly if a number of simulations are to be performed across the range of stellar progenitors and input physics. One such simulation

Three-dimensional simulations of neutrino transport equations already require petascale computing; six-dimensional simulations, needed to accurately model supernova explosions, will require exascale computing.

is expected to take ~8 weeks, assuming 20% efficiency on an exaflops machine.

In the future, the enhanced spectral resolution and throughput of missions like Constellation-X will provide data from more distant (and therefore more numerous) supernovae. Infrared observations of heavy element abundances, like those performed in the past with the Infrared Space Observatory (ISO) and in the present with the Spitzer Observatory, and those possible with future spectroscopic missions like JWST, can report the temporal history of element production from supernovae. These observations, coupled with theoretical study of the ways the earliest supernovae differ from those of the present epoch, are important to understanding how the first stars formed and how they changed over time into the objects recognized in the present universe.

The coincidence of several GRBs with core-collapse supernovae has opened a new chapter in the study of supernovae. Time-sensitive detections, such as the HETE-II observations of GRB030329 (which led to the discovery of Supernova 2003dh), and Swift's detection of GRB060218, as well as those that may be provided in the future by missions like the Energetic X-ray Imaging Survey Telescope (EXIST), will enhance the opportunity for simultaneous detailed multiband observations from missions such as Integral, Chandra, and HST. These will shed light on the GRB/supernova connection by greatly improving the quality and quantity of data available on these events.

3.4 Compact Objects

While simulations of BBHs are progressively becoming more accurate and efficient, some aspects of these calculations are outside the realm of current computational resources. The advent of exascale computing resources will allow a wider coverage of BBH parameter space, determined by the masses, spins, and orbit eccentricities. It will also allow for the simulation of BBHs with extreme mass ratios (smaller than 1/20).

The addition of matter will also improve the realism of simulations involving supermassive holes found at the center of galaxies, where the inspiral dynamics are bound to be affected by the presence of accretion disks and galactic environments. These simulations will (one hopes) shed light on the origin and dynamics of galactic jets and quasar evolution.

Among the driving forces behind simulations of compact object binaries are the new generation of laser-interferometric gravitational wave observatories such as LIGO, VIRGO, and LISA and future missions such as NASA's Black Hole Finder Probe and Imager, and the associated large investment that has been made. One expected outcome of a simulation is the computation of the gravitational waves produced during the binary merger. Without the gravitational wave "templates" from numerical relativity, detection and characterization of gravitational radiation sources will be extremely difficult.

In addition, binaries with at least one neutron star are the most likely engine of short-duration GRBs (the other class of GRBs in the universe; long-duration GRBs are associated with core-collapse supernovae as discussed above). Their modeling is crucial for the interpretation of data from current missions such as Swift and future probes such as NASA's Gamma-ray Large Area Space Telescope (GLAST).

Modeling of long-duration gamma-ray bursts is crucial for interpreting data from NASA missions.

4. Accelerating Development

Exascale simulation of the universe on all scales will require a variety of simulations that will be performed using a variety of codes founded on different numerical methods. The community has produced codes that solve the equations of *N*-body dynamics, hydrodynamics, MHD, radiation transport, radiation hydrodynamics, radiation MHD, and both Newtonian and general relativistic gravity. The underlying methods have thus far proven robust in scaling to present-day architectures. The underlying assumption, and hope, of the computational astrophysics community is that these methods can be extended to the exascale without significant modification. Of course, the community is aware that, in some cases, new algorithms for the underlying PDEs will be needed. Moreover, modifications will have to be made to accommodate the increasing numbers of cores on each socket and the changes in the demand for memory bandwidth associated with this new feature, but there are no obvious bottlenecks at present that suggest that an entirely new set of codes will have to be deployed. The community also recognizes that this hope is predicated upon the hope that the architectures at the exascale will not be dramatically different either from those used now or from those that will be used in the near term at the petascale. Thus, of primary importance to the computational astrophysics community will be collaboration with the applied mathematics community with an eye toward porting existing methods and codes to exascale platforms.

The ability to perform simulations at much higher resolution will be of great benefit to the computational astrophysics community. Nonetheless, given the multiscale nature of the astrophysical systems studied by this community, adaptive mesh refinement (AMR), which is becoming more prevalent now, will no doubt see increased use in the future. AMR can be performed in different ways (e.g., grid-based or cell-by-cell refinement), each having its pros and cons. Thus, we will have to consider both the scaling of our numerical solution methods, as discussed above, and the scaling of our approach to AMR. The primary issue with regard to AMR and parallel computing is, of course, load balancing. Thus, the development of methods for dynamic load balancing across large numbers of processors will be an increasingly important need of the computational astrophysics community, regardless of the approach to AMR.

Exascale computing will also provide the ability to simulate astrophysical systems for significantly longer physical evolution time scales. In short, we can expect significantly longer run times at the exascale. The development of fault-tolerant solution algorithms and efficient parallel solution algorithms that minimize the time to solution will be required to mitigate the need for fault tolerance.

Area	Science
Large-Scale Structure Formation	Simulations of the large-scale distribution of galaxies and galaxy clusters over a large fraction of the observable universe with 1% precision, required of the observational program proposed by the DOE/NASA/NSF-sponsored Dark Energy Task Force.
Galaxy Formation	Simulations of galaxy formation with sufficient resolution to predict the observed properties of individual galaxies in a volume containing a sufficient fraction of the observable universe to compare with large-scale surveys. Simulations of the formation of the Milky Way galaxy with sufficient precision to compare with data from JWST and LSST.
Stellar Evolution	Simulations of the entire stellar envelope in AGB stars, responsible for the supply of half of the heavy elements (elements above iron) in nature.
Supernovae	Definitive 3D multiphysics simulations of core-collapse supernovae, the dominant source of elements between oxygen and iron and the half of the heavy elements not produced in AGB stars.
Compact Objects	Definitive simulations of binaries involving two neutron stars, or one black hole and one neutron star, which are among the leading candidates for the production of gravitational waves in our galaxy.

Table 5.1 Examples of exascale-computing-enabled science

Longer run times will also present a very different challenge. In many instances, explicit (in time) methods are used in the codes mentioned above to follow the evolution of astrophysical systems. The confluence of higher grid resolution, or a larger number of particles, and longer physical simulation times will push explicit methods to their limits. In some instances, explicit methods will no longer be viable. Thus, the development of implicit (in time) methods for the solution of some of the underlying PDEs listed above will be required.

The computational astrophysics community is already using implicit (in time) methods to solve the systems of equations governing radiation transport in astrophysical systems. Such methods lead to a set of underlying large, sparse linear systems of equations,

Proposed Observatory	Estimated Cost	"Launch" Date
JWST	$4.6B	2013
LISA	$1.7B	2015
Constellation-X	$1.7B	2017
LSST	$300M	2017 (ground-based)
JDEM	>$0.6B	2015

Table 5.2 Scheduled investments in observatories during the next decade

which at present are solved with both matrix-based and matrix-free implementations of Newton-Krylov (NK) methods. These linear systems can be ill conditioned, which has motivated a need for higher-precision arithmetic and for the development of the associated solution algorithms.

Throughout the scales of the universe, gravity is (of course) prevalent in astrophysical systems. The gravitational field, in conjunction with the energy (mass) content that defines it, is among the major components of most astrophysical systems/models. In the Newtonian case, the gravitational field is determined by solving the Poisson equation. In the general relativistic case, the complete system of Einstein's equations (or a good approximation thereof) needs to be solved. In both instances, scalable methods for the solution of elliptic systems of equations on exascale platforms will be needed.

5. Expected Outcomes

Table 5.1 lists examples of astrophysical science that will be enabled by the advent of exascale computing.

Computational astrophysics has a long and storied tradition of driving advances both in computing, per se, and in algorithmic development. For these reasons and others, computational astrophysics has also proven to be a fertile training ground for computational scientists who ultimately spend most of their professional lives in other disciplines. The need for broadly trained computational scientists will only increase as the ambitious aims outlined here are undertaken.

Exascale computing will enable study of the entire sky, rather than merely a slice of the sky at a time.

What will exascale computing mean to the computational astrophysics community in the context of the science delineated above? First and foremost, exascale computing will enable 3D modeling across the phenomena we have discussed. At present, the ability to perform complete simulations in three spatial dimensions is not a given. For example, no 3D multiphysics simulations of core-collapse supernovae have been performed to date. The current state of the art remains at two spatial dimensions. Second, exascale computing will mean higher grid resolution or an increased number of particles. Third, exascale computing will enable the study of larger physical volumes — for example, the entire sky rather than a slice of the sky in simulations of large-scale structure formation or the entire convective stellar layer in AGB stars vs only a portion of it. Finally, exascale computing will enable complete multiphysics simulations. Almost without exception, simulations performed to date across the suite of areas discussed above have left out a significant physical component.

Our ultimate goal is to understand the formation and evolution of the major constituents of our universe, from the largest to the smallest of scales, from the universe as a whole to the clusters of galaxies making up its large-scale structure to individual galaxies and stars to compact objects such as neutron stars and black holes, and, when possible, to understand the connection between phenomena at different scales.

6. Required Investment

As shown by a few representative examples in Table 5.2, over the next decade significant investments will be made to develop and deploy the next-generation ground- and space-based observatories that will provide more complete and more precise observations of the universe across the electromagnetic spectrum, and in other forms of radiation such as gravitational waves and neutrinos, in regions currently observed. Equally important, theses instruments will allow us to peer farther into space and farther back in time to regions of the universe and its history heretofore unexplored. The success of these new observatories, and the benefit of the significant investments that will be made to make these observatories a reality, will depend on the development of more complete and more precise simulations of phenomena in the cosmos. Without a significant investment in computational astrophysics and the development of the computational platforms, applied mathematics, and computer science on which simulations by the computational astrophysics community will depend, the benefit of investments in new observatories will not be harvested.

Without the development of exascale platforms and without the associated development of applied mathematics and computer science, the benefits of investment in new observations will not be realized.

7. Major Risks

The computational astrophysical community has identified the following key risks in embracing exascale computing:

- The simulations proposed above will put severe constraints on the memory per processor, memory bandwidth, and total memory of the new exascale machines. This situation is largely due to the multidimensional (even beyond three spatial dimensions, in order to include radiation angles and energies) and multiphysics nature of the phenomena being simulated.

- An astrophysical system often has an inherent physical parallelism. Ideally, architectures would be designed to exploit such parallelism by performing the associated work on a single socket rather than across sockets, but this would require sufficient socket memory.

- The increased number of cores per socket envisioned for the exascale will likely overwhelm the memory bandwidth to the shared socket memory available to these cores.

- Some proposed simulations will be memory bound. For example, core-collapse supernova simulations at the exascale will require approximately 1 exabyte per flop. A more complete response to the issue of risk will evolve as detailed knowledge of exascale architectures comes to light.

References

Andreas Albrecht, Gary Bernstein, Robert Cahn, Wendy L. Freedman, Jacqueline Hewitt, Wayne Hu, John Huth, Marc Kamionkowski, Edward W. Kolb, Lloyd Knox, John C. Mather, Suzanne Staggs and Nicholas B. Suntzeff (2006), Report of the Dark Energy Task Force, arXiv.org:astro-ph/0609591.

John G. Baker, Joan Centrella, Dae-Il Choi, Michael Koppitz and James van Meter (2006), Gravitational wave extraction from an inspiraling configuration of merging black holes. *Phys. Rev. Letters* 96:111102.

J. M. Blondin and A. Mezzacappa (2007), Pulsar spins from an instability in the accretion shock of supernovae, *Nature* 445:58–60.

A. C. Calder, D. M. Townsley, I. R. Seitenzahl, F. Peng, O. E. B. Messer, N. Vladimirova, E. F. Brown, J. W. Truran and D. Q. Lamb (2007), Capturing the fire: flame energetics and neutronization for Type Ia supernova simulations, *ApJ* 656:313–332.

M. Campanelli, C. O. Lousto, P. Marronetti and Y. Zlochower (2006), Accurate evolutions of orbiting black-hole binaries without excision, *Phys. Rev. Letters* 96: 111101.

V. N. Gamezo, A. M. Khokhlov and E. S. Oran (2005), Three-dimensional delayed-detonation model of Type Ia supernovae, *ApJ* 623:337–346.

E. J. Hallman, B. W. O'Shea, J. O. Burns, M. L. Norman, R. Harkness and R. Wagner (2007), The Santa Fe Light Cone Simulation Project, I: Confusion and the WHIM in upcoming Sunyaev-Zel'dovich effect surveys, ArXiv e-prints 704:0704.2607.

Wayne Hu and Scott Dodelson (2002), Cosmic microwave background anisotropies, *Ann. Rev. Astronomy and Astrophysics* 40:171.

Hans-Thomas Janka, K. Langanke, A. Marek, G. Martinez-Pinedo and B. Mueller (2007), Theory of core-collapse supernovae, *Phys. Rept.* 442:38–74.

D. Lynden-Bell (1969), Galactic nuclei as collapsed old quasars, *Nature* 223:690.

Anthony Mezzacappa (2005), Ascertaining the core-collapse supernova mechanism: The state of the art and the road ahead, *Ann. Rev. Nucl. Part. Sci.* 55, 467–515 (2005).

Carlos F. Sopuerta, Ulrich Sperhake and Pablo Laguna (2006), Hydro-without-hydro framework for simulations of black hole-neutron star binaries, *Classical and Quantum Gravity* 23:S579.

S. E. Woosley and J. S. Bloom (2006), The supernova – gamma-ray burst connection, *Ann. Rev. Astronomy and Astrophysics* 44:507.

6 Math and Algorithms

Advanced and improved simulations for climate, renewable energy, nuclear energy, astrophysics, biotechnology, and nanoscience (to name a few areas in the DOE portfolio) demand significant advances in mathematical methods, scalable algorithms, and their implementations. The required advances are driven by the increased complexity of the problems, involving multiple and coupled physics models, high dimensionality described by large numbers of equations, and huge time and spatial scales—from nano to global and even to astronomical scales.

These advances will represent a fundamental, exciting, and powerful shift in the computational science modus operandi, enabling a move beyond simply solving larger problems at higher resolutions to providing new capabilities for optimizing and quantifying uncertain systems and unknowns and for assessing risk.

The results of these simulations and their interpretation will help guide high-impact policy and scientific decisions with broad social, engineering, and ecological consequences.

1. Advances in the Next Decade

As depicted in Figure 6.1, the needs of exascale applications will be addressed by advances in four major and interlinked areas: coupled models, uncertainty, optimization, and data.

- *Coupled Models:* Most exascale applications will involve multiple models (PDE or data-based) for different phenomena or at different scales of the same phenomenon. In some cases, new models and corresponding scalable implementations will be developed for the coupled systems. In many cases, existing models and codes will be extended in terms of scalability, and new mathematical approaches (general and domain-specific) to model coupling will be developed. New implicit approaches for dealing with long time-scale coupled simulations will also be pursued.
- *Uncertainty:* To add rigor to exascale simulation results, we must develop a systematic approach for quantifying, estimating and controlling the uncertainty caused, for example, by reduced models, uncertain parameters, or discretization

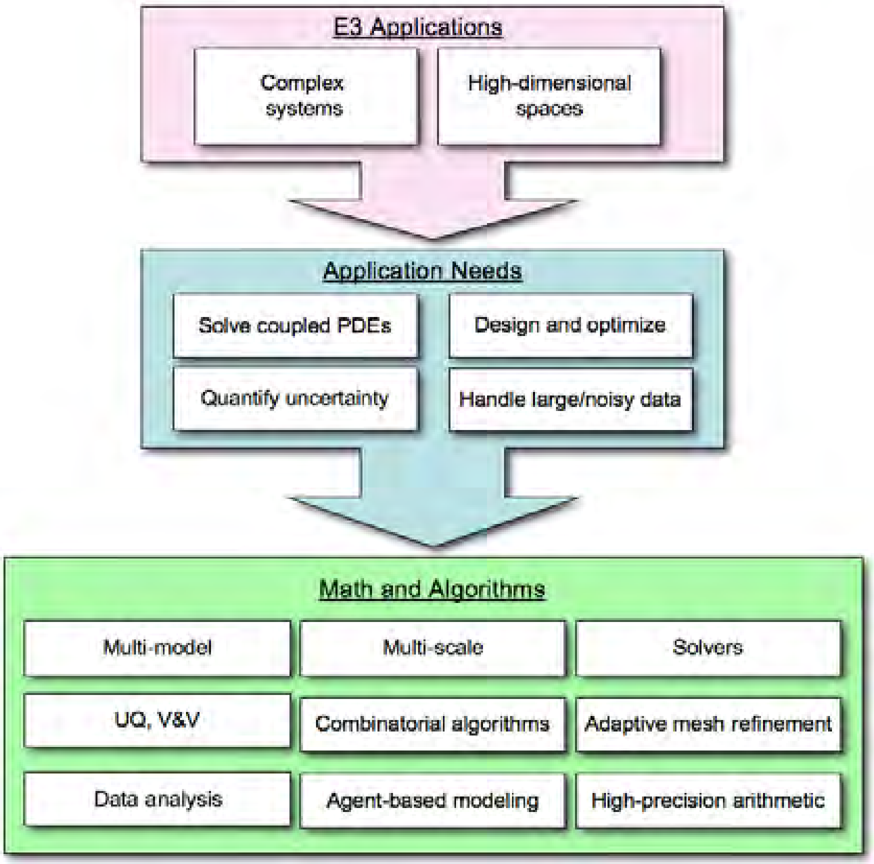


Figure 6.1 E3 applications characteristics and math and algorithms needs.

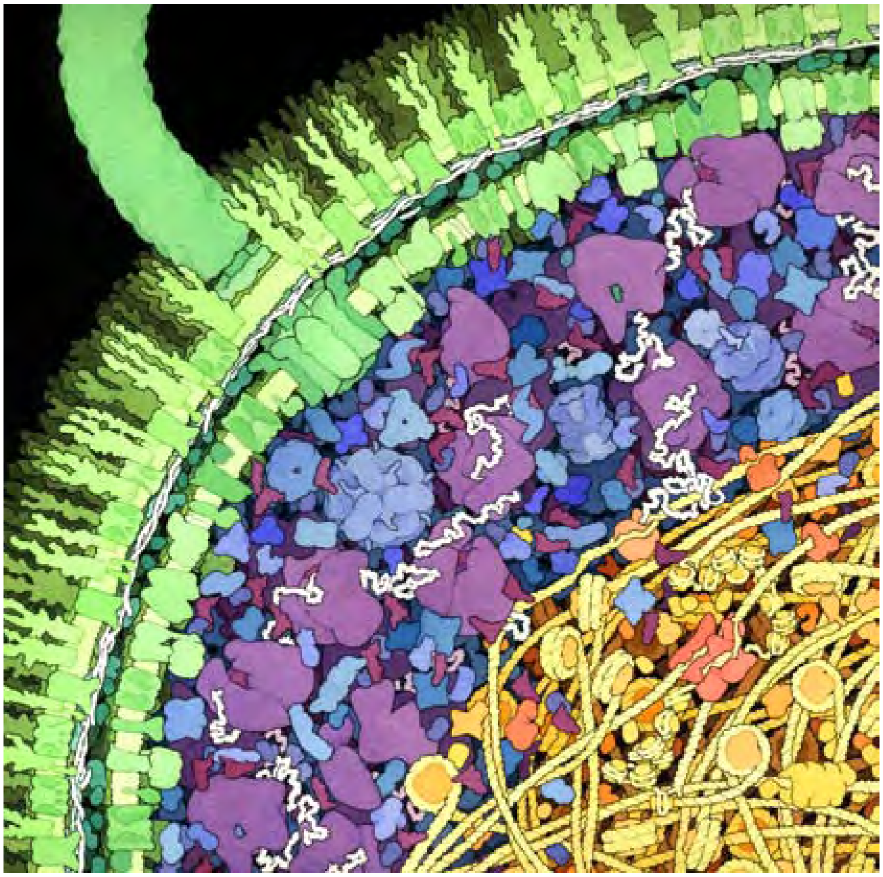


Figure 6.2 A cross section of a small portion of an *Escherichia coli* cell, an example of a complex system that might one day be the subject of an ab initio simulation. The cell wall, with two concentric membranes studded with transmembrane proteins, is shown in green. A large flagellar motor crosses the entire wall, turning the flagellum that extends upwards from the surface. The cytoplasmic area is colored blue and purple. The large purple molecules are ribosomes; the small, L-shaped maroon molecules are tRNA; and the white strands are mRNA. Enzymes are shown in blue. The nucleoid region is shown in yellow and orange, with the long DNA circle shown in yellow, wrapped around HU protein (bacterial nucleosomes). In the center of the nucleoid region shown here, one might find a replication fork, with DNA polymerase (in red-orange) replicating new DNA. Image courtesy of David S. Goodsell.

error. Complex, fully coupled multiphysics and multiscale applications require intrusive tools that automatically construct representations of uncertainty, handle the uncertainty propagation and coupling effects, and provide sharp estimates of the uncertainty of key merit criteria.

- *Optimization:* Large-scale design problems demand scalable algorithms for continuous nonlinear optimization. Also needed are parallel branch-and-cut methods for linear and nonlinear optimization problems with discrete variables, as well as more sophisticated parallel methods for solving stochastic optimization problems.

- *Large, Messy, and Noisy Datasets:* Advances in technology have enabled the production of massive volumes of data through observations and simulations in many applications such as biology, high-energy physics, and astrophysics. To effectively utilize this flood, we will develop new data representations, data-handling algorithms, efficient implementations of data analysis algorithms on high-performance computing platforms, and representations of analysis results.

Exascale computing presents an opportunity not just for quantitative but also for qualitative changes in the role of computation in scientific discovery and in high-consequence decision support. Applications for exascale systems will demand rigorous V&V methodologies with quantified uncertainties that are presented in a manner that enables simulation-based critical decisions and optimization. Exascale computing opens many opportunities that will help drive mathematical and algorithmic research, making possible the design of safe, reliable, economical, and socially acceptable solutions for energy, climate, and the environment. The E3 initiative must ensure that critical applied mathematics, algorithms, and software challenges are addressed.

For computational biology, sustainability, and global security applications, HPC systems must place greater emphasis on scalable shared-memory performance, memory latency and bandwidth, and integer operations, in contrast to the current focus that rests almost exclusively on floating-point performance. Bioinformatics and computational biology, genomics, and medical applications may differ significantly from the current HPC workloads in that the data structures are often irregular (based on strings, trees, graphs, and networks), without the high degree of spatial and temporal locality seen in physics-based simulations using regular matrices. We also see a growing need for database research in areas of probabilistic queries and queries by structure, such as DNA sequence data, protein structure, and phylogenetic graphs or networks. We envision that new exascale algorithms will require tight integration of computation with database operations and queries, as well as the ability to handle new types of queries such as combined structural, ethno-botanical, socio-geographic, phylo-geographical queries.

2. Major Challenges

Exascale computing has the transformational power to allow scientists to move beyond simulation of complex systems and to consider a paradigm shift that will enable them to address more challenging questions such as design and the quantification of uncertainty. This paradigm shift, together with the massive increase in computing power offered by exascale computing, will create a set of computational and mathematical hurdles that must be overcome. Four broad areas have been identified that encapsulate the application needs: uncertainty quantification, optimization, PDEs, and the management of massive sets of data.

Uncertainty Quantification. The quantification of uncertainty answers such fundamental questions as simulation code V&V against reality. It is important to establish mathematically firm foundations for these fundamental questions and to develop computational tools within an integrated computational environment that reduces time to simulation, provides compatible geometry representations, allows for a hierarchy of model fidelities running on a range of architectures from workstations to state-of-the-art parallel computers, and includes a rich suite of postprocessing capabilities. Typically, scientists have assumed independence between various sources of input uncertainty. While this assumption simplifies the analysis, it does not represent reality, and there exists a need to support higher-dimensional input densities to model true model uncertainty. New methodology is needed to systematically explore the input uncertainty space for an optimal characterization of the output uncertainty space. Typically, researchers have developed techniques principally for optimizing physical experiments; these techniques must be adapted for computational experiments, leading to uncertainty quantification for thousands or millions of parameters. The development of these mathematical foundations and tools is critical to the understanding of the "teleconnections" in model output that depend on the dynamical system under study; such dependence information is critical to studying extremes and extremes of uncertainty.

Optimization. The design of complex systems and physical processes that maximize some performance measure can be expressed as optimization problems [Nocedal 2006]. Optimization holds the promise of better designs and a more efficient use of natural resources. The key challenge is the development of robust optimization techniques that are reliable and exploit the evolving new computer architectures. These techniques must be easy to use, in order to ensure their widespread acceptance within the scientific community. An additional challenge is the determination of appropriate algorithms for novel optimization paradigms that could not have been approached on a system scale before the advent of exascale architectures. Such novel paradigms include hierarchical optimization problems over multiple time stages and problems with chance constraints that better reflect licensing and operational requirements. A third challenge that will arise with the use of exascale architectures is the handling of problems with hundreds of thousands of discrete parameters. Moreover, as scientists move increasingly from simulation to design, they will need to address the solution of complex, nonlinear optimization problems—possibly involving trade-offs between multiple objectives and discrete choices. A portfolio of techniques will be needed, ranging from more sophisticated PDE-constrained optimization methods to new, rigorous techniques that exploit a hierarchy of physical models.

PDEs. Solving linear and nonlinear systems of PDEs is at the heart of many DOE applications, such as accelerator modeling, astrophysics, nanoscience, and combustion. As the focus of scientists extends from simulation to optimization and uncertainty quantification, the solution of linear and nonlinear systems increases in importance. The efficient solution of PDEs requires AMR to provide mesh-quality-preserving automatic adaptation for different types of complex geometries, while allowing the addition of domain-specific in-

In enabling scientists to move from simulation to design, exascale computing will also demand novel paradigms, such as hierarchical optimization over multiple time stages and problems with chance constraints that better reflect licensing and operational requirements.

terpolation strategies. Dynamic load balancing is especially challenging at the exascale level, leading to hard combinatorial problems [Streensland 2002]. Many complex physical systems, such as astrophysics and climate modeling, face bottlenecks in time step restrictions because of higher spatial resolutions, which must be overcome to ensure convergence for reasonable time steps.

Massive Datasets. The handling and management of increasingly large and heterogeneous sets of data will create new challenges for scientists and computer architectures. Data can be the result of an exascale simulation that must be postprocessed for human interpretation, or it can form the input to complex problems via data assimilation. Browsing or looking at data is no longer possible as we near a petabyte. To visualize 1% of 1 petabyte at 10 MB/s takes 35 workdays. There is an enormous need for methods to dynamically analyze, organize, and present data by variability of interest. For example, how do we organize ocean eddies in a climate simulation so that by viewing a very small set of them we have a good idea about all the types of eddies present in the entire petabyte dataset? What about particles in a fusion simulation? Solutions to these problems will likely come from dynamically considering high-dimensional probability distributions of quantities of interest. This approach requires new contributions from mathematics, probability, and statistics. Besides being large, data is often noisy, or inaccurate, thus creating additional challenges. For example, we may be interested in estimating the response of a nonlinear dynamical system that is polluted by noise. How can we detect the real signal? Answers to such questions require new mathematical theory and computational tools to classify shapes in terms of stochastic models and to deal with both local and long-range correlations between features.

3. State of the Art

In this section, we examine the current state of the art in six areas that will require significant mathematical and algorithmic advances in order to meet the emerging needs of exascale applications.

Dynamic load balancing is especially challenging at the exascale level, leading to hard combinatorial problems.

3.1 Solvers

The dominant computational solution strategy over the past 30 years has been the use of first-order-accurate operator-splitting, semi-implicit and explicit time integration methods, and decoupled nonlinear solution strategies. Such methods have not provided the stability properties needed to perform accurate simulations over the dynamical time scales of interest. In most cases, numerical errors and means for controlling such errors are understood heuristically at best. In addition, the impact of these choices is difficult to assess and control. For this reason, solutions for these complex systems can be fragile and exhibit nonintuitive instabilities, or they may simply be stable in a crude sense but contain significant long-time integration error.

Direct methods for solving sparse linear equations are used in many applications, such as the inversion operator for the shift-and-invert algorithms for high-accuracy eigencomputations, solution of coarse-grid problems as part of a multigrid solver, and subdomain solutions in domain decomposition methods, as well as in cases where the linear systems are extremely ill-conditioned [Li 2006]. Iterative methods are required for most 3D and coupled physics problems, however, since direct methods become computationally infeasible because of memory and time requirements [Eijkhout 1998].

The numerical optimization community has focused principally on the development of serial algorithms for solving linear, quadratic, and nonlinear optimization problems with continuous variables or a mixture of continuous and discrete variables. Techniques for solving linear and quadratic programs can be split into active-set and interior-point methods, while algorithms for solving nonlinear programming problems are generally based on solving a sequence of linear or quadratic approximations or interior-point methods [Nocedal 2006]. Parallel implementations have been written for some problem classes by parallelizing the linear algebra [Gondzio 2003] or exploiting the problem structure by using a decomposition method, for example. Preconditioned Krylov methods and approximate solutions to the systems of equations are not well understood from a theoretical perspective, especially for problems with inequality constraints, but they have been used successfully in some contexts. Algorithms for optimization problems with discrete variables solve a sequence of linear and nonlinear relaxations in a branch-and-cut framework [Nemhauser 1988]. These methods are augmented by a number of heuristics for finding an initial feasible point, selecting the branching variables, and determining the cuts to add. Parallelization of these methods is achieved by solving multiple relaxations in parallel. Optimization problems with stochastic variables have also been studied, in which one typically samples the random variables and solves the resulting deterministic problem. The problems for the realizations can be solved in parallel. For large parameter spaces, the sampling method employed becomes very important for obtaining reasonable results.

3.2 Uncertainty Quantification

A complete answer to the uncertainty question in computational experiments involves three distinct steps: (1) representation of input uncertainty [Klir 1994], (2) propagation of uncertainty through a simulation model [Christianson 2005], and (3) representation and calculation of output uncertainty. The overall approach is guided by the initial representation of the uncertainty. For pure stochastic representations, the state of the art for step 1 includes Gaussian randomness of inputs and parameters, used in conjunction with perturbative techniques for step 2 and with Monte Carlo for step 3. Such representations have been carried to dimensions of the random vector as high as millions. Nonetheless, further progress in this direction is hindered by the severe nonlinearities exhibited by most models of interest or by the failure of the data to support efficient Gaussian approximation. A relatively recent development is the use of stochastic finite elements to represent uncertainty in inputs and parameters as an element in a linear space at step 1. This has the promise of very compact and efficient representations of the uncertainty. Moreover, the approach has been able to accommodate quite severe nonlinearities. Nonetheless, to date we have no efficient way of achieving step 2 when several models with such uncertainty representation are coupled. This is a major bottleneck in expanding the reach of such methods and is responsible for the fact that parametric uncertainty quantification for PDE applications has been limited to tens of parameters for problems whose state space is on the order of millions in magnitude.

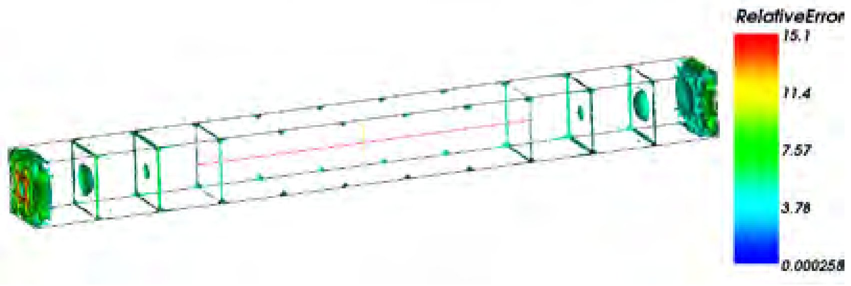


Figure 6.3 Error of a multiscale model reduction approach for electronic structure density functional theory (DFT) energy minimization for a string of hydrogen atoms.

3.3. Adaptive Mesh Refinement

AMR is not widespread in scientific simulation today but is slowly gaining acceptance. It is used within both structured and unstructured mesh contexts. *Block-structured* AMR combines the efficiency of having resolution only where needed with the simple array-based storage and mathematical operations of structured meshes. Most AMR computations are performed using packages such as Chombo, GrACE, or Paramesh, since such computations require sophisticated numerical, algorithmic, and software constructs to run efficiently on parallel machines. Most discretization techniques (finite differences and finite volumes) on block-structured AMR are second-order accurate, although there have been recent extensions to fourth-order numerical schemes. The use of AMR within *unstructured* meshes, and especially finite-element calculations, is far more advanced. Simulations with both automatic resolution and discretization order refinement (*hp*-refinement) are conducted today.

AMR necessitates load balancing on parallel computers. Several models for partitioners and load balancers have been applied successfully to unstructured problems [Hendrickson 1995; Pilkington 1994; Simon 1991]. The best-known model is *graph partitioning*, where data/work is represented by graph ver-

Stochastic finite-element methods can accommodate severe nonlinearities, but propagating uncertainties through coupled models presents a major bottleneck to expanding the range of such methods to exascale architectures.

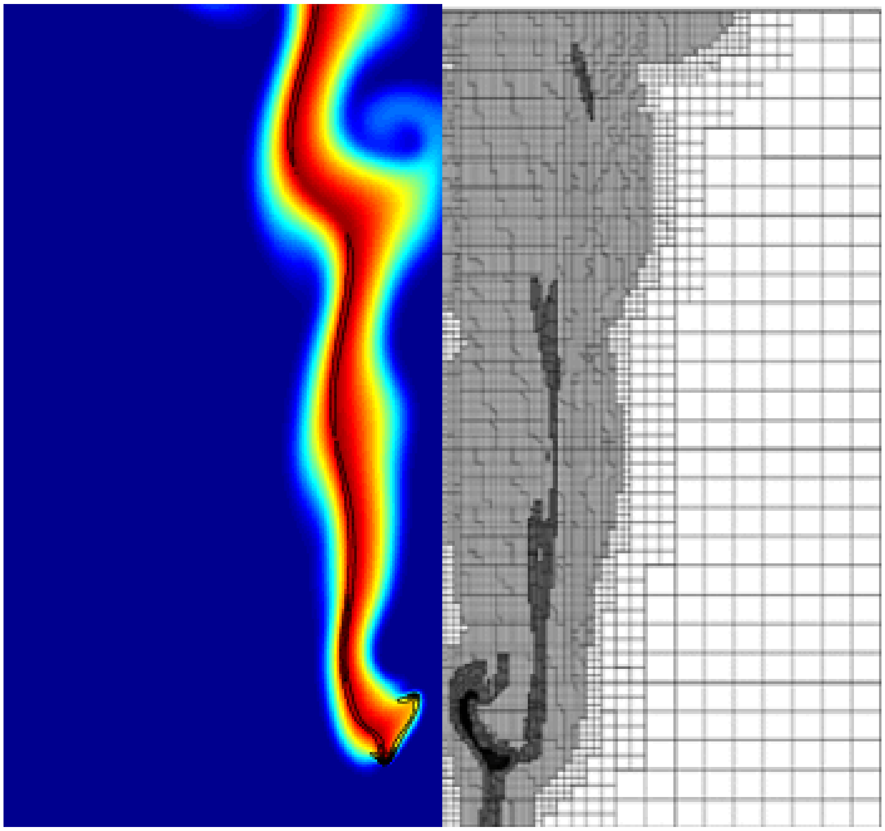


Figure 6.4 Left: Temperature (color map) and heat release contours for a partially premixed unsteady methane jet flame. Right: The corresponding AMR mesh.

tices. Hypergraph partitioning approaches improve on the graph model by representing data dependences within sets of vertices, thus allowing nonsymmetric and nonsquare data dependences to be modeled. In general, hypergraph algorithms produce decompositions with lower communication volume (vis-à-vis graph partitioners) while supporting a larger range of applications.

Scientific simulations that use block-structured adaptive meshes use mainly *geometric models* [Patra 1995], which implicitly assume that objects that are physically near each other depend on each other. Geometric methods are very fast and scalable compared to graph and hypergraph algorithms. Current techniques perform domain decomposition in such a way that the subdomain per processor can be represented as a collection of nonoverlapping rectangular boxes. Two extreme objectives can be adopted in partitioning such meshes: (1) reduce load imbalance, even at higher communication costs (patch-based partitioning), and (2) reduce communication costs, at the price of increased load imbalance and synchronization costs. Successful algorithms (e.g., bi-level, knapsack) take a middle-ground approach. Furthermore, many algorithms have tunable parameters to trade off load balance with communication costs, given the character of a particular machine. Determining optimal values for these parameters is difficult, however, and doubly so for time-evolving simulations on adaptive meshes.

3.4 High-Dimensional Spaces

In genomics applications involving mutation analysis, parameter spaces may exceed 5,000 dimensions and may take only discrete values. Yet state-of-the-art algorithms, such as entropy-based methods, have not been successful for problems exceeding 100 dimensions.

In the simulation of a fission reactor, the situation is even more challenging. Some parameters are outside the user's control, such as scattering, absorption, and fission cross sections; these parameters suffer from a larger degree of uncertainty than do others. For a moderately high number of energy groups, the number of such parameters can easily surpass 10,000. Yet the state of the art currently is limited to parameter spaces of perhaps 20 to 30 dimensions.

A complicating factor in the topic of high-dimensional parameter spaces is the fact that such problems can also have a huge state space. In a fission reactor, for example, the state space is composed of an ensemble of the neutron flux, temperature, and coolant velocity distributions in a 3D configuration with sizes on the order of meters in any direction and physics that must be represented at scales far below the millimeter range.

3.5 Data Analysis

The current state of data analysis lags far behind our ability to produce simulation data or record observational data. A particular gap exists in the mathematics needed to bring analysis and estimation methodology into a data-parallel environment. Parallel linear algebra methods go a long way toward enabling data-parallel analysis, but they do not solve it, just as they would not solve a climate simulation problem. For example, the standard principal component analysis computation does not become data-parallel with a parallel singular value decomposition (SVD) solver, even though the SVD is the core computation in that analysis. Data-parallel solutions for applications on exascale resources will require new mathematics that considers an entire estimation problem for developing scalable data-parallel algorithms in data analysis.

3.6 High-Precision Arithmetic

No HPC vendor currently offers hardware support for 128-bit floating-point arithmetic, and there is little prospect that this situation will change within the next few years. Some Fortran compilers support the REAL*16 datatype in software. Unfortunately, such facilities are not provided in all compilers—not in the GNU compilers, for instance—and even when present, they are usually very slow, often 50 to 100 times slower than conventional 64-bit floating-point arithmetic. Few scientists are willing to experiment with such a facility when the performance penalty is so great.

The alternative is to use independently written software libraries, such as those mentioned by Bailey [2005]. With such software, one specifies which variables and arrays are to be treated as high precision by means of special type statements. Then when one of these variables or arrays appears in an arithmetic statement or argument, the proper library routines are automatically called via operator overloading. Such facilities have several drawbacks: (1) at present they are available only for Fortran and C++; (2) it is necessary to make alterations (mostly minor) to one's source code; (3) certain subexpressions may be performed only to conventional precision; and (4) slowdown is typically a factor of 5 for double-double arithmetic, 25 for quad-double, and even higher if transcendental function references are involved.

4. Accelerating Development

The movement to multiscale, multiphysics simulations necessitates advances in several fundamental classes of numerical algorithms, including linear solvers, nonlinear solvers, preconditioners, eigensolvers, algorithms for mesh generation and adaptation, and ordinary differential equation (ODE) / differential algebraic equation (DAE) integrators.

4.1 Solvers

In order to achieve accurate, stable, efficient and scalable predictive simulations for multiple-time-scale systems with *implicit methods*, many advances in numerical methods and computational science are required.

- Development, demonstration, and comprehensive evaluation of stable, accurate, efficient, and scalable fully implicit methods coupled with uncertainty quantification techniques (deterministic and probabilistic) and with estimation and control of long-time integration error for large-scale, complex multiple-time-scale applications. Verification and well-characterized prototype problems for multiple-time-scale multiphysics systems must be developed.
- Deterministic uncertainty quantification tools based on sensitivity and adjoint-based techniques for data, integration, and model error estimation and control must be further developed and demonstrated. Adjoint methods have shown promise for estimating and controlling data, model, and long-time integration error. In adjoint-based methods, advances are required to limit solution storage requirements, memory usage, parallel communication, and cost of the adjoint solve. There is a significant need for the development of public-domain software to enable adjoint methods for time-dependent problems. In addition, adjoint techniques for hyperbolic systems are critically required.
- Probabilistic approaches based on sampling methods (e.g., Monte Carlo) and direct methods (e.g., polynomial chaos) require advances in algorithms and computationally efficient implementations for transient simulations. Hybrid deterministic/probabilistic approaches must be developed and studied in the context of complex systems.

The current state of data analysis lags far behind our ability to produce simulation data or to record observational data.

Few compilers or software libraries can handle high-precision arithmetic, and those that do so are often extremely slow.

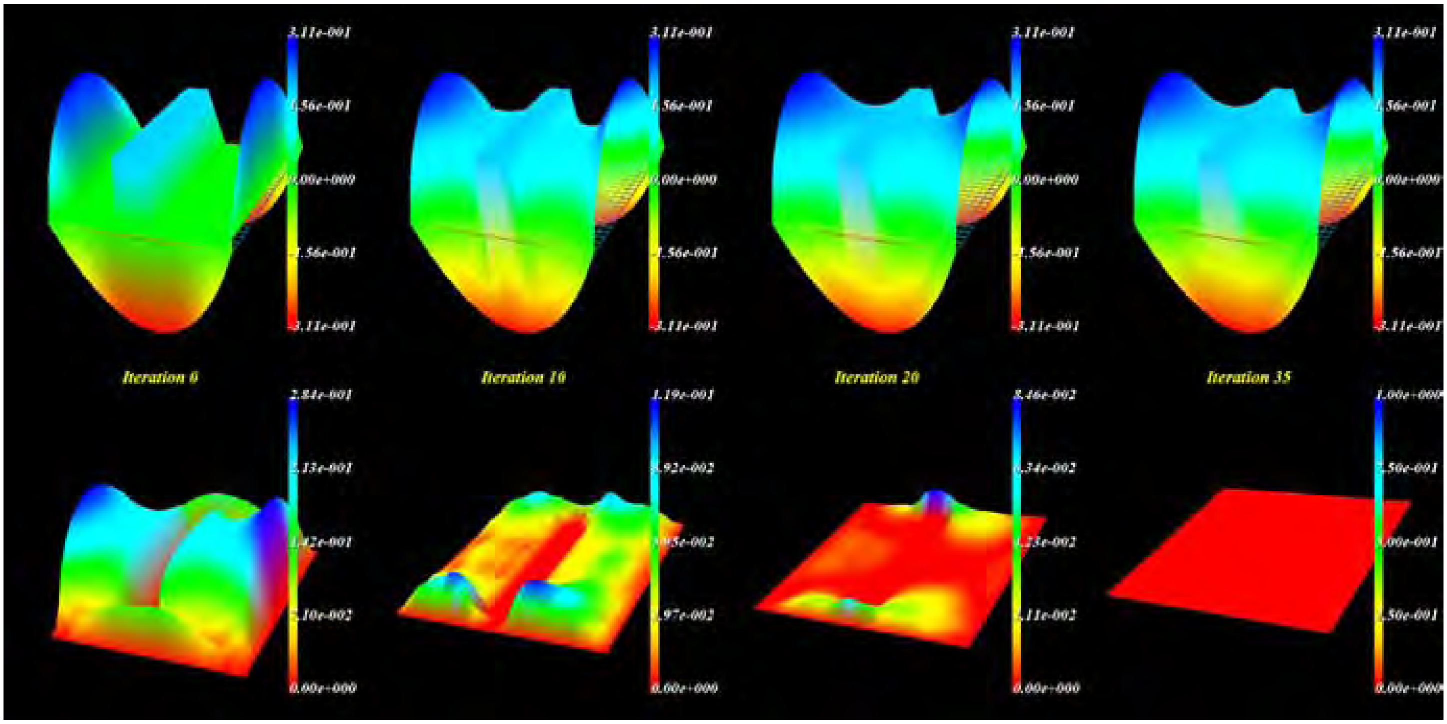


Figure 6.5 A sequence of solutions and their deviation from the optimal solution for a bound constrained minimization problem. The objective is the surface with minimal area that satisfies Dirichlet boundary conditions and is constrained to lie above a solid plate.

- NK and Jacobian free NK (JFNK) techniques in complex large-scale applications must be significantly extended. Needed extensions include automated generation of adjoint models and automatic differentiation technologies to support NK and JFNK methods in complex applications.

- NK methods require algorithmic and software developments in physics-based and approximate block factorization preconditioners for coupled elliptic, parabolic, and hyperbolic systems. Critical subcomponent physics include incompressible and compressible flow, transport-reaction systems, coupled porous media flow, nonlinear elasticity, fluid-structure interaction, electromagnetics, and ideal and resistive MDH.

- Efficient and scalable physics-based preconditioners require subblock solvers based on multilevel (multigrid and algebraic multigrid) methods for scalar and vector systems with strong anisotropic effects and large-scale variations in coefficients. Advances in algebraic multigrid methods for compatible physics-based discretizations are also required.

- The System of Systems (SOS) approach offers promise in building alternative physical models based on the integration of a smaller, well-understood set of known physical models into a larger model. This approach may prove useful, for instance, in exploring change in global temperature based on smaller local models (e.g., climate models for small regions or series of physics-based PDE models).

Research in *linear and nonlinear solvers* remains a critical focus area because the solvers provide the foundation for more advanced solution methods. In fact, as modeling becomes more sophisticated and increasingly includes optimization, uncertainty quantification, perturbation analysis, and more, the speed and robustness of the linear and nonlinear solvers will directly determine the scope of feasible problems to be solved.

- The performance of direct methods relies heavily on the ordering of rows and columns of the matrix, and research in effective orderings is still important. Furthermore, for extremely large systems, out-of-core implementations will be needed to overcome the memory bottleneck. The efficiency of such out-of-core solvers will need to be revisited. Efficient and scalable algorithms for sparse indefinite systems, particularly those from optimization, must be investigated. Moreover, further work is needed for fully distributed direct sparse solvers, where all data objects are distributed across the parallel machine and efficient parallel algorithms are used for all phases of the solver.

- Block iterative methods, which can effectively solve multiple simultaneous right-hand sides, represent an area of growing importance in order to exploit the growing number of such problems. Furthermore, block iterative methods have attractive machine performance characteristics that will become even more important on future architectures.

- Convergence of iterative methods is strongly affected by the quality of preconditioners. Advanced efforts in preconditioning are focused on exploiting problem characteristics via segregation of variables for coupled systems, multilevel methods for keeping complexity costs low, and nesting of iterative methods via inner-outer techniques. However, even these techniques still rely on basic preconditioners (incomplete factorizations, Jacobi, Gauss-Seidel, etc.) and iterative methods as building blocks. Therefore, any improvements in basic preconditioners, such as better reorderings, reduced error in incomplete factors, or emerging approaches such as support graph techniques, can have a broad impact.

- Progress in nonlinear solvers continues to be important, especially for tightly coupled and highly nonlinear systems where continuation methods can be essential to getting a converged solution. Work in graph coloring is an important related problem for efficient Hessian and Jacobian computations, as is automatic differentiation as a means of obtaining accurate and flexible operator formulations.

Optimization research should focus on the following four areas; a complementary effort should involve educating scientists to select the best modeling and solution techniques for optimization problems.

- Modeling systems that allow application scientists to express their models in a natural, domain-specific format and that couple applications to solvers. This effort will involve the development of new tools that support PDE constraints, chance constraints, and simulation-based applications.

- Scalable parallel solvers for optimization problems with PDE constraints. One approach is to extend current nonlinear optimization techniques to allow for inexact subsystem solves.

- New methods for hierarchical and simulation-based design problems that exploit exascale architectures. Such methods should include constraints, allow for inexact simulations, and provide support for multiscale optimization.

- Efficient parallel branch-and-cut solvers for nonlinear discrete optimization problems that enable the solution of applications with hundred of thousands of discrete parameters.

Block iterative methods have attractive machine performance characteristics that will become important on exascale architectures.

4.2 Uncertainty Quantification, Validation, and Verification

Effective, scalable algorithms for uncertainty quantification and optimization will likely be intrusive, requiring extensive changes to existing application codes, the underlying algorithms, and the solvers required for efficient solution that will drive the need for improved iterative solvers and preconditioners. While the advantages of intrusive methods have been well documented, their use has been limited by the substantial changes and related solver advances required. E3 should promote next-generation applications that support fully intrusive uncertainty quantification and optimization algorithms. To achieve this objective, advances in numerical methods and computational science are required:

- *Verification*. Address the mathematical challenges of error estimation for

Advances in preconditioners, such as better reorderings, can have a broad impact on the performance of iterative solvers.

complex, coupled, multiscale, and multiphysics calculations with possibly noisy and discontinuous data and engage the community to improve the overall quality of computational testing, including the design, documentation, and repository of useful and influential benchmark problems.

- *Validation*. Focusing on the selection and design of appropriate experimental benchmarks as well as the mathematical and computational challenges of validating systems of coupled simulations that may be used well outside their validation regime.

- *Uncertainty quantification.* Formulate mathematical foundations, scalable algorithms, and high-quality software implementations for stochastic PDEs, sampling methods, polynomial chaos expansions, uncertainty propagation, adjoint-based sensitivity methods, Bayesian inference strategies, and possibilistic and fuzzy inference strategies. Also needed are effective models to deal with information presented with imprecise probability, nonmodel uncertainty mitigation, and experimental characterization of input uncertainties.

- *Decision making*. Explore the role of uncertainty quantification for effective communication of computational results in complex decision environments, including within optimization. A major remaining mathematical challenge is how to properly formalize confidence in computational simulations to best support decision-making under uncertainty and imprecise information. Uncertainty quantification must be tightly coupled with development and execution of sophisticated mathematical methods for V&V, advanced analysis, and visualization methods for uncertain data.

- *Community awareness***.** Develop a research training regime for computational science V&V and uncertainty quantification. Such a regime will certainly be highly interdisciplinary, involving mathematics, computation, scientific application domain experts, probability and statistics, and decision theory.

Adaptive mesh refinement at the exascale requires development of efficient and rigorous error metrics.

4.3 Adaptive Mesh Refinement

AMR at the exascale requires significant advances in the following areas, some of which overlap with the combinatorial algorithms research discussed in Section 4.5.

- Exascale simulations will exhibit large length-scale heterogeneities within complex geometries [Chandra 2007]. Maintaining resolution and mesh quality will require new algorithms as well as *hp*-refinement techniques [Patra 1995].

- AMR meshes change with time and must be rebalanced. The availability of a portfolio of partitioning strategies will enable the selection of the optimal rebalancing strategy for a given problem or computation stage [Streensland 2002].

- High-order methods can reduce the degree of refinement required, leading to smaller meshes and more efficient solutions. While AMR finite-element calculations have successfully incorporated high-order elements, block-structured methods are limited to second order. The primary challenge is creation of tractable stencils at coarse-fine interfaces for implementing conservation laws.

- The development of efficient and rigorous error metrics for refinement is an important challenge. Indirect techniques that estimate the spectral content of the solution locally (and thus recommend a level of resolution) provide a trade-off between efficiency and mathematical rigor.

- Block-structured AMR meshes are usually restricted to simple geometries. Mapping such meshes to smooth geometries automatically is an important challenge that will enable new applications to benefit from AMR [Schwartz 2006].

- Efficient algorithms must be developed to unravel grid hierarchy into a sparse matrix form (and vice versa) to allow new solvers such as Krylov solvers to be used instead of traditional geometric multigrid methods.

- AMR will need adaptive partitioning to scale to exascale levels. Rather than configuring a partitioner at the beginning of the run, a "control system" approach to partitioning may have to be adopted.

- AMR in turn will benefit from advances made on discrete algorithms to address new multicore architectures, data layout, and processor interconnectivity, leading to multiple competing objectives.

4.4 High-Dimensional Spaces

Calculations have been done for resolutions that are far coarser than what will be required by exascale applications, but current and expected advances in scalable algorithms for PDE software make it likely that the computation of state variables will be achievable on exascale computers. This is not the case for increasing the parameter space, where the curse of dimensionality applies to most obvious algorithms, including possibly random sampling (since the variance itself may be unbounded with the increase in the size of the parameter space, even if the error decrease is dependent on the number of samples and not on the dimension of the space). Therefore, significant research must be undertaken to develop new algorithms that make use of problem structure for breaking the curse of dimensionality.

Random sampling tends to be the most commonly used technique to approach this problem; however, the points need not be chosen by a stochastic algorithm. Indeed, recent approaches that seem promising choose deterministic points. Progress may come from the refinement of ideas currently at the forefront of this research, such as higher-order sampling methods (e.g., randomized Monte Carlo sampling), structured stochastic finite-element approaches, sparse grids, and hierarchical models for importance sampling and preconditioning; from the identification of new assessment and design paradigms that may be better suited for large-scale applications, such as robust optimization and the use of chance constraints; or from radically new ideas brought about by increased focus in this crucial direction.

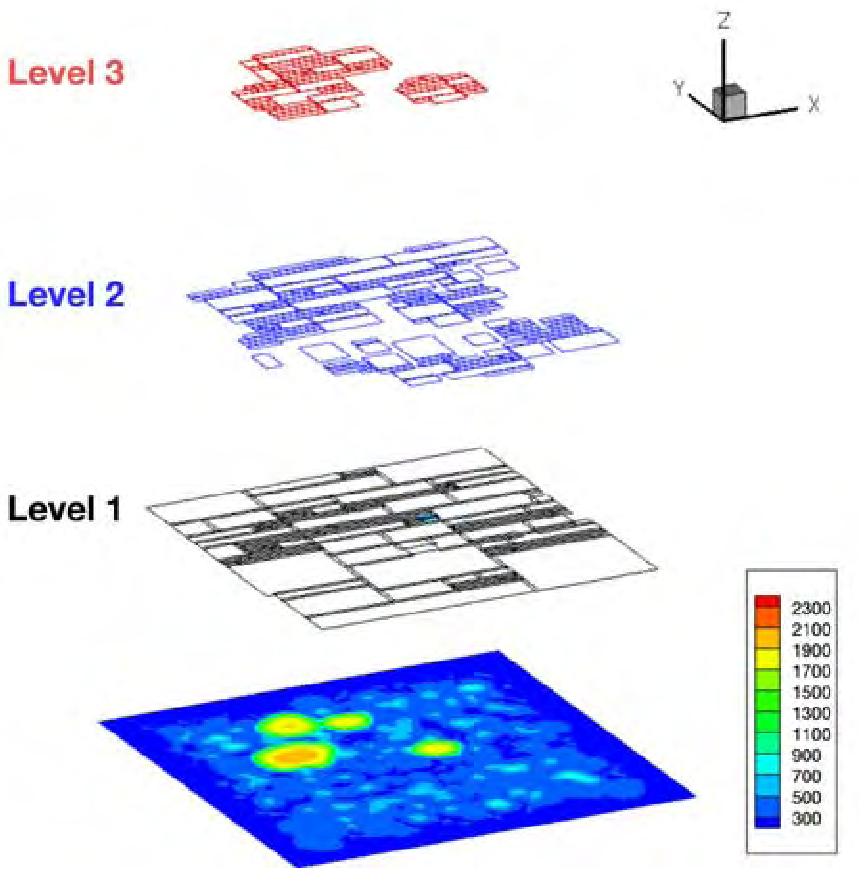


Figure 6.6 Temperature map and mesh refinement from a H_2-air mixture ignited by random high-temperature kernels, after ~90 μs of simulation time.

4.5 Combinatorial and Discrete Algorithms

Combinatorial and discrete algorithms have long played an important role in many applications of scientific computing, such as sparse matrix computations and parallel computing, to enhance the performance of numerical algorithms. Emerging applications such as computational biology, scientific data mining, and complex systems bring combinatorial algorithms to the fore as an integral part of computational sciences. We must develop highly scalable kernels for discrete mathematics to support such applications. These kernels should be developed independently from the application use and should be leveraged by a wide array of code bases on the petascale, exascale, and beyond. Systems that

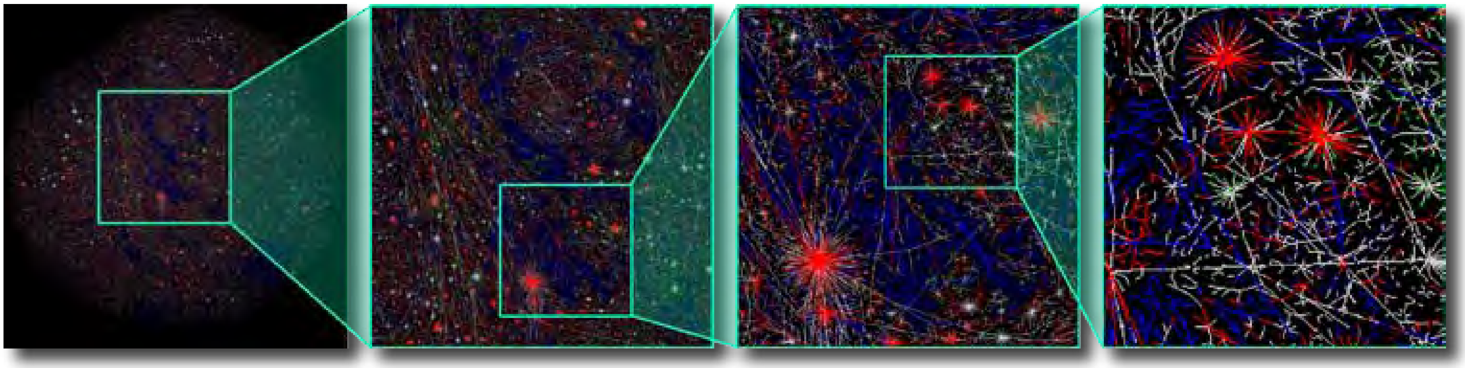


Figure 6.7 A graph layout generated by LGL [Adai et al. 2004] showing a portion of the minimum spanning protein homology tree with over 300,000 proteins. An edge is colored blue if it connects 2 proteins from the same species, and red if it connects 2 proteins from 2 different species. If that information is not available, the edges are colored based on layout hierarchy. They remain white if there is no species information available. Image courtesy of Alex Adai and Edward Marcotte.

will benefit from improved combinatorial algorithms include the following:

- The unit commitment problem for power systems. Optimal power flow is a challenging combinatorial problem, and power system analysis for dynamic security assessment is vital. The solution of realistic-sized models involves discrete parameters on an unprecedented scale.

- Bioinformatics models of DNA, RNA, and proteins as sequences over small alphabets. Searches for specific structures, gene regulatory networks, protein interaction networks, and metabolic networks all give rise to combinatorial problems.

- Efficient utilization of underlying computational infrastructure and interconnection networks. The increasing gap between CPU and memory performance argues for the design of new algorithms and data structures and for data reorganization to improve locality at memory, cache, and register levels.

- Utilization of observational devices such as telescopes. Projects now share telescope time with interleaved schedules. Yet experiments are designed for longer terms (a complete supernova observation, for example, requires observations over a 60-day period), making scheduling a crucial challenge.

Interval arithmetic may help researchers determine where a code is having numerical difficulty and to certify final results.

4.6 High-Precision Arithmetic

Some, and possibly many, exascale applications will require high-precision arithmetic facility, yet there is little prospect for vendor support. While some software packages are available, they have shortcomings. What is needed is a simple-to-use facility that infallibly converts large application programs for high precision, yet results in only a modest inflation in run time. Such a facility is possible but not yet available.

Beyond the immediate needs of usable high-precision software, better understanding is needed of situations that can lead to numerical difficulties in large computations. Also needed are tools that can quickly detect whether and where an application program is experiencing difficulties in this area. Along this line, current research in numerical analysis, particularly in the area of interval arithmetic, may be of use in the HPC world, in that it may provide a means to determine whether and where a code is experiencing numerical difficulties, and to certify that the final results are within a certain specified tolerance of their correct values [Hayes 2003].

4.7 Data Analysis

Solutions for estimation problems based on a mixture of integer (categorical) and real values are needed for biological data in particular. Many solutions exist (most based on the likelihood principle), but they are complex, and their implementations are serial, so they typically do not scale to even terabyte datasets. The mathematics of many estimation problems must be reorganized so that the estimation can be performed in a data-parallel environment of the future petascale and exascale environments.

4.8 Agent-Based Modeling

Agent-based modeling [Woolridge 2002; North and Macal 2007] must be advanced along several directions before it can present a viable approach for addressing exascale application needs:

- Distributed query resolution to allow agents to flexibly and repeatedly find other agents and recognize affordances for interaction in a dynamic environment with a continually and endogenously evolving structure (e.g., nonreified networks)
- Situational activation of agents based on contextual factors and associated reallocation to workings sets of processors with appropriate interprocessor locality
- Efficient implementation of periodic fine-grained interactions between agents where the payoffs from the interplay are endogenously defined as a function of the ongoing interactions themselves, such that players are free to enter and leave the interactions at idiosyncratic times
- Distributed time scheduling at a level of parallelism beyond the current approaches
- Extremely high volume data warehousing to allow efficient exploration of huge numbers of large model runs
- Efficient directed sweeps across huge model parameter spaces with appropriate adaptation as results are discovered
- Domain decomposition techniques for parallel agent-based simulations where the computational load per agent is variable, in time for the same agent, as well as from agent to agent, and where the geographical locality has no relation to the nature and volume of communication between agents

Domain decomposition techniques for parallel agent-based models must be formulated to address exascale application needs.

5. Expected Outcomes

A focused program to address the qualitatively and quantitatively different challenges of next-generation scientific applications, tailored to encourage and support close collaborations between applied mathematicians, computer scientists, application scientists, and hardware vendors, will result in the ability to tackle emerging complex, coupled problems in climate, biology, energy, the environment, and other global problems.

6. Required Investment

Sufficient and timely investment in new mathematical approaches and algorithms is crucial for transforming computation into a powerful tool for scientific discovery and high-consequence decision support. The effort will require multidisciplinary teams to advance the state of the art in each of the key areas laid out in Section 4. Significant energy must be expended on designing, developing, testing, and deploying the integrated frameworks and tools required to ensure effective use of the software in the major algorithm classes. Some effort must also be dedicated to maintaining collaborations that will support cross-cutting aspects of the development in mathematics and algorithms with the development in the application areas and in the software, hardware, and cyberinfrastructure development teams. We anticipate that this will require between $100M and $200M over the next decade. More finely tuned projections will become possible as the envisioned efforts advance through planning and ramp-up stages.

7. Major Risks

The principal risks and downside consequences linked to R&D in mathematics and algorithms are as follows:

- If advances in theoretical performance of new hardware architectures are not matched with significant R&D in mathematics and algorithms, then real performance gains are unlikely to be realized. The investment in hardware advances will be rendered ineffective for lack of scalable algorithms.

- If significant advances in methods for managing complexity in models and computational simulations fail to materialize, then the opportunity to expand our simulation capability into a number of critical application areas will be lost. We will be left with a dwindling cadre of applications that can benefit from advances on the hardware front.

- If we fail to develop new mathematical methods, data structures, and algorithms for integrating the growing volume of data coming from our experiments into our simulations, we will not be able to extract sufficient benefit from such development.

- If we inadequately connect the development described in this report to the application developers' needs, we will generate tools and methods that are not widely accepted and therefore miss the mark on broad benefits.

- If R&D is not supported with long-term investment, short-term gains will fail to mature and evolve with the changing needs of the applications and with the changing details of the hardware architectures. The benefit of initial investments will be lost.

- If we fail to build robust abstraction layers insulating the application developer from the details of the hardware architecture and the algorithm, the anticipated complexities of expressing and managing computations at the exascale will stifle development of applications.

- If we do not develop invasive implementations for some of the more complex and onerous aspects described above, particularly optimization and uncertainty, they will generally go unused, thereby limiting the kinds of applications that will be developed, as well as casting doubt on the utility of results from large and complex simulations.

- If we do not design our algorithms and interface layers in collaboration with the hardware and software developers, it will be extremely difficult to realize the goal of developing scalable implementations. This situation will be increasingly relevant as raw performance results from combinations of parallelism through core count and inclusion of special-purpose acceleration hardware.

- If we do not increase the use of good software engineering practices, the complexities of developing and maintaining our software will overtake the energies expended in mathematics and algorithms R&D.

- If we fail to monitor and maintain an appropriate balance between precision calculation and robust solutions as guiding metrics of our algorithm design, we run the risk of becoming needlessly and overwhelmingly burdened with the demand for precision, which will result in delayed deployment, incomplete risk assessment, or outright failure.

Aggressive R&D in mathematical methods and scalable algorithms is key to successful development and deployment of future applications relying on computational simulation, data assimilation, and analytics. It will be a critical component of an effort that will result in quantitatively and qualitatively new and powerful scientific methods.

The complexities of managing computation at the exascale will require robust abstraction layers that insulate the application developer from the details of both the hardware and the algorithms.

References

A. T. Adai, S. V. Date, S. Wieland and E. M. Marcotte (2004), LGL: Creating a map of protein function with an algorithm for visualizing very large biological networks, *J. Mol. Biol.* 340(1):179-190.

D. A. Bader (2004), Computational biology and high-performance computing, special issue on bioinformatics, *Communications of the ACM* 47(11):34–41.

D. A. Bader, A. Snavely and G. Jacobs (2006), NSF workshop report on petascale computing in the biological sciences, August 29–30, Arlington, VA.

D. H. Bailey (2005), High-precision arithmetic in scientific computation, *Computing in Science and Engrg*., May–June, 54–61.

M. Barad and P. Colella (2005), A fourth-order accurate local refinement method for Poisson's equation, *J. Comp. Physics* 209(1) 1–18.

M. J. Berger and S. H. Bokhari (1987), A partitioning strategy for nonuniform problems on multiprocessors, *IEEE Trans. Computers* C-36(5): 570–580.

J. Birge and F. Louveaux (1997), Introduction to Stochastic Programming, Springer.

U. Catalyurek and C. Aykanat (1996), Decomposing irregularly sparse matrices for parallel matrix-vector multiplications, *Lecture Notes in Computer Science* 1117: 75–86.

U. Catalyurek and C. Aykanat (1999), Hypergraph-partitioning-based decomposition for parallel sparse-matrix vector multiplication, *IEEE Trans. Parallel Dist. Systems* 10(7): 673–693.

S. Chandra, M. Parashar and J. Ray (2007), Analyzing the impact of computational heterogeneity on runtime performance of parallel scientific components, in Proc. 15th High Performance Computing Symposium (HPC-07), SCS Spring Simulation Multiconference, Norfolk, VA.

B. Christianson and M. Cox (2005), Automatic propagation of uncertainties, pp. 47–58 in M. Buecker, G. Corliss, P. Hovland, U. Naumann, and B. Norris, eds., Automatic Differentiation: Applications, Theory, and Implementations. *Lecture Notes in Computational Science and Engineering* vol. 50, Springer.

K. Devine, E. Boman, R. Heaphy, R. Bisseling and U. Catalyurek (2006), Parallel hypergraph partitioning for scientific computing, in *Proc. IPDPS 2006*.

V. Eijkhout (1998), Overview of iterative linear system solver packages, *NHSE Review* 3(1).

J. Gondzio and R. Sarkissian (2003), "Parallel Interior-Point Solver for Structured Linear Programs." *Mathematical Programming* 96: 561-584.

B. Hayes (2003), A lucid interval, *American Scientist* 91(6) 484–488.

B. Hendrickson and R. Leland (1995), A multilevel algorithm for partitioning graphs. in *Proc. Supercomputing '95*.

H. Johansson and J. Steensland (2006), A performance characterization of load balancing alogrithms for parallel SAMR applications, Technical Report 2006-047, Uppsala University, Dept. of Information Technology.

G. Karypis and V. Kumar (1998), A fast and high quality multilevel scheme for partitioning irregular graphs, *SIAM J. Sci. Computing* 20(1) 359–392.

G. Karypis and V. Kumar (1997), A coarse-grain parallel multilevel *k*-way partitioning algorithm, in *Proc. 8th SIAM Conf. Parallel Processing for Scientific Computing*.

G. J. Klir (1994), The many faces of uncertainty, pp. 3–19 in B. M Ayyub and M. M. Gupta, eds., Uncertainty Modeling and Analysis: Theory and Applications, *Elsevier Science*.

X. Li (2006), Direct solvers for sparse matrices, http://crd.lbl.gov/~xiaoye/SuperLU/SparseDirectSurvey.pdf, September.

G. Nemhauser and L. Wolsey (1988), Integer and Combinatorial Optimization, John Wiley & Sons.

J. Nocedal and S. Wright (2006), Numerical Optimization, 2nd ed., Springer.

M. J. North and C. M. Macal (2007), Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation, Oxford.

A. Patra and J. T. Oden (1995), Problem decomposition strategies for adaptive hp finite element methods, *Computing Systems in Eng*. 6(2) 97–109.

J. R. Pilkington and S. B. Baden (1994), Partitioning with spacefilling curves, CSE Technical Report CS94-349 Dept. Computer Science and Engineering, University of California – San Diego.

J. Ray, C. A. Kennedy, S. Lefantzi and H. N. Najm (2007), Using high-order methods on adaptively refined block-structured meshes derivatives, interpolations, and filters, *SIAM J. Sci. Computing* 29(1):139–181.

P. Schwartz, M. Barad, P. Colella and T. Ligocki (2006), A Cartesian grid embedded boundary method for the heat equation and Poisson's equation in three dimensions, *J. Comput. Phys.* 211(2) 531–550.

H. D. Simon (1991), Partitioning of unstructured problems for parallel processing, *Computing Systems in Engrg*. 2: 135–148.

J. Steensland (2002), Efficient partitioning of dynamic structured grid hierarchies, Uppsala University Library, Uppsala, Sweden.

J. Steensland and J. Ray (2003), A heuristic remapping algorithm reducing inter-level communication in SAMR applications, in *Proc. 15th IASTED International Conference on Parallel and Distributed Computing and Systems 2003* (PDCS03).

R. W. Walters and L. Huyse (2002), Uncertainty analysis for fluid mechanics with applications. Technical report, NASA, Feb.

M. Woolridge (2002), An Introduction to Multi-Agent Systems. Wiley & Sons.

7 Software

Solutions to the major software, algorithm, and data challenges in newly emerging programming environments are needed to make it possible for applications to scale up to the required levels of parallelism and integrate technologies into complex coupled systems for real-world multidisciplinary modeling and simulation. Developing these solutions will likely involve a shift from current static approaches for application development and execution to a combination of new software tools, algorithms, and dynamically adaptive methods. Additionally, we must bring together new developments in system software, data management, analysis, and visualization to allow disparate data sources (both simulation and real-world) to be managed in order to guide research and to directly advance science.

Computer vendors have little incentive to tackle these challenges because the commercial market at this scale is relatively small. For this reason, government support for the necessary R&D in exascale software is essential.

In particular, significant new efforts are needed to

- fundamentally change how applications are built and maintained;
- improve scientists' and administrators' productivity when working with exascale systems;
- improve the robustness and reliability of systems and applications through fault tolerance, validation, and verification of software components;
- integrate knowledge discovery into the entire software life-cycle, from discovering bugs to monitoring and steering simulations to discovering new scientific results hidden in huge volumes of data; and
- develop new approaches to handling the entire data life-cycle of exascale simulations, from inputs that may be dynamic, live feeds, to distributed data analysis, and long-term archival.

1. Advances in the Next Decade

The infrastructure must be designed in a way that allows it to adapt to future trends and remain relevant, correct, and high performance as systems evolve to exascale and beyond. The characteristics of these exascale systems with millions of CPUs require the development of new ideas and approaches.

It is both feasible and necessary to build on newly emerging programming environment (interpreted broadly) technologies to develop the infrastructure that will allow applications to scale up to the required levels of parallelism and integrate into complex coupled systems for real-world multidisciplinary modeling and simulation. Considerable investment has been made in the past few years in new languages designed to improve productivity for programmers. Examples are the Partitioned Global Address Space Languages and the High-Productivity Computing System (HPCS) languages. In 5–10 years, these efforts should have a chance to coalesce into one or two standard, widely supported languages.

Exascale systems with millions of CPUs demand new approaches to address the challenges of the data tsunami, system reliability, and memory hierarchies.

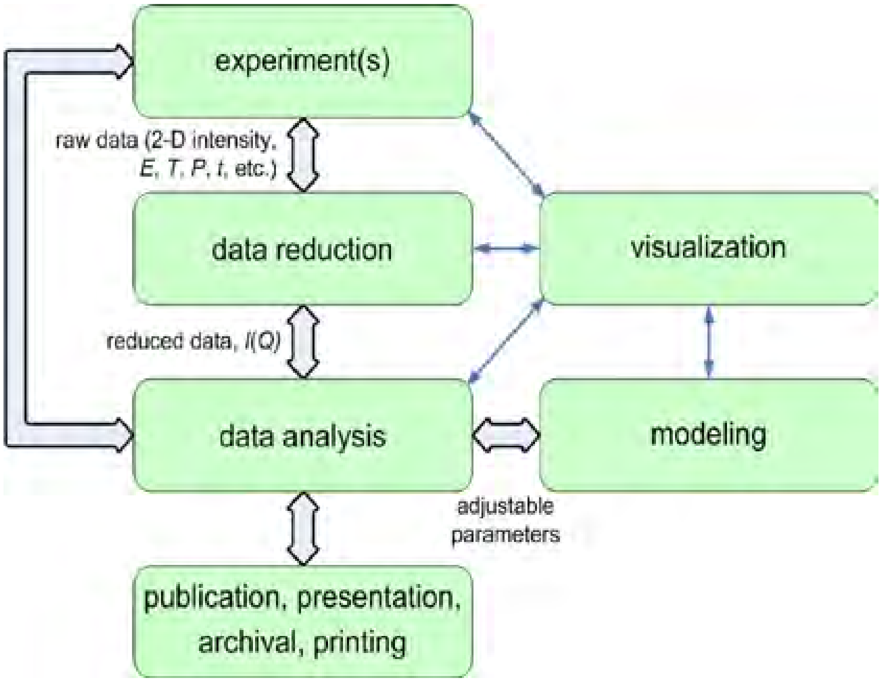


Figure 7.1 Scientific workflow diagram

It is plausible to bring together new developments in system software, data management, analysis, and visualization to allow disparate and voluminous data sources (both simulation and real-world) to be managed in order to guide research and to directly advance science.

For these developments to be useful at full scale, however, a more flexible and dynamic resource management capability is needed throughout the computing environment to allow simultaneous integration of computing, analysis, visualization, and live data during a simulation.

The vision for the next decade is to have a totally integrated approach to how applications are built, modified, updated, and used in other applications. In this development environment, the tools will interoperate and assist the scientists in writing, debugging, tuning, and maintaining their codes. Such an environment will support fundamentally new approaches:

- *New ways of specifying computations.* Scientists must be freed from the details of managing data movement among memory systems and synchronizing access to shared memory among threads of control. They will need languages and libraries, in some cases discipline- or even application-specific, which specify results to be obtained with less attention to the details of the computation than is currently necessary. Implementation of such libraries and languages will require lower-level programming models and tools that permit execution on a wide range of hardware and exploit the capabilities of exascale architectures.

- *Rapid, modular construction of new applications from existing suites of interoperable components.* Scientific software components with well-defined interfaces, recently being prototyped but not yet widely adopted, have the potential to greatly increase code reuse, thus shortening development times and increasing software reliability.

- *Coupling of multiple applications into ever-larger applications through automated workflows.* Single large runs remain an important class of large-scale computations, but many applications need parameter studies consisting of large numbers of coordinated sets of runs, each perhaps consisting of a pipeline of computation and analyses. High-level, standard languages for coordinating such families of executions will enable scientists to focus on science rather than “run management.” (See Figure 7.1.)

- *Dynamic data storage and management.* In order to support exascale data generation, data storage will fundamentally change. Much of the data will be archived. For archival storage to be energy efficient, yet available on demand, tools will be needed to manage the movement of data automatically across a storage hierarchy. Frequently used data will be moved to highly parallel dynamic storage, while archived data will reside in powered-down storage or passive storage devices. Furthermore, algorithms for automatically tracking and removing unused data from dynamic storage will be essential to minimize storage costs. Collections of datasets will be organized as directories. Such abstraction will fundamentally change the way the

Software components, recently prototyped but not yet widely adopted, will shorten development time and increase software reliability.

I/O is expressed by applications and will involve a storage management layer that maps datasets into physical devices without affecting the applications.

Achieving this vision will require fostering long-term, sustained, community-wide activity in evolving code suites. Large-scale applications, like large-scale computers themselves, require the support of multiple specialists within a single community. Indeed, the community of computer vendors, application scientists, and computer scientists, together with the hardware and software that they both develop and use, forms an integrated, interdependent ecosystem.

2. Major Challenges

Five major challenges must be addressed in order to realize efficient and effective exascale computing.

2.1 Improving Programmability

Exascale computer architectures will require radical changes to the software used to operate them and the applications that run on them. The shift from faster processors to multicore processors is as disruptive to software as the shift from vector to distributed memory supercomputers 15 years ago. That change required complete restructuring of scientific application codes, which took years of effort. The shift to multicore exascale systems will require applications to exploit million-way parallelism and significant reductions in the bandwidth and amount of memory available to millions of CPUs. This "scalability challenge" affects all aspects of the use of HPC. It is critical that work begin today if the software ecosystem is to be ready for the arrival of exascale systems in the coming decade.

The shift to multicore exascale systems will require applications to exploit million-way parallelism.

2.2 Building New Applications

The transition from frequency-based scaling to core-based scaling also means that the community will face a memory crisis. This is not just a bandwidth and latency issue that has been faced as memory gets farther away from CPU operations (at least in terms of clocks), but also an overall memory capacity problem. Memory will increase in relative cost, and it will be harder and harder to maintain the desired byte-to-flop ratio—in absolute capacity (flops/s per byte) and bandwidth terms (flops per byte).

Hence, applications will have to be redesigned to make better user of limited memory. This may mean more out-of-core solutions (pushing on the I/O bottleneck) or algorithms with better storage characteristics. Additionally, applications will have to deal with increasing hierarchies of memory (and indeed storage). There are now often five levels of direct-access memory common (register sets, three levels of cache, and main memory). In the future there may be more levels or more (or less) sharing of these levels within an SMP.

2.3 Increasing Parallelism

In today's environment, code development usually takes place assuming a homogenous run-time environment, with parallelization done manually by each code developer. At the scale where applications need to make use of millions of heterogeneous processes, discovering the opportunities for parallelization becomes much more difficult and requires a set of tools that can automate the parallelization of the trivially parallelizable segments of code, and aid the application developer in finding less obvious opportunities. This task is even more daunting when considering future multicore architectures, since the parallelization algorithms have to take into account in-core parallelism vs between-cores parallelism. These tools are also needed for developing simulation code that will run well on heterogeneous hardware, with the compiler automating as much of this as possible and providing code-restructuring assistance where automation is not possible.

2.4 Handling the Data Tsunami

The data tsunami includes dealing with the volume, different formats, transfer rates, analysis, and visualization of massive (potentially distributed) data sets.

Exascale applications running on as many as a million processors are likely to generate

Exascale applications will generate several terabytes of data per second, presenting enormous challenges in storage and management.

data at a rate of several terabytes per second (even assuming only a few megabytes per processor). Because it is not practical to store raw data generated at such a rate, dynamic reduction of the data by summarization, subset selection, and more sophisticated dynamic pattern identification methods will be necessary. The reduced data volume must be stored at the same rate that data are generated, in order for the exascale computation to progress without interruption.

This requirement presents new challenges in orchestrating data movement from the computation machines to local and remote storage systems. It will no longer be possible to store all the data locally and then distribute them as secondary tasks. Data distribution will have to be integrated into the data generation phase. Here again, managing the dataflow using well-coordinated workflow engines will be required as part of the software infrastructure that runs the simulations.

The issue of large-scale data movement will become more acute as very large datasets or subsets are shared by large scientific communities. This will require the replication or movement of large volumes of data between production and analysis machines, often across the wide area. While networking technology is greatly improving with the introduction of optical connectivity, the transmission of large volumes of data will inevitably encounter transient failures, and automatic recovery tools will be necessary.

Another fundamental requirement is the automatic allocation, use, and release of storage space. Replicated data cannot be left in storage devices unchecked, or storage systems will fill and become clogged. A new paradigm of attaching a lifetime to replicated datasets, and automatic management of data whose lifetime expires, will be essential.

2.5 Accelerating Knowledge Discovery

One of the principal bottlenecks in contemporary science is the process of discovering knowledge and testing hypotheses in the presence of a growing deluge of data. A recurring theme in this report—that existing methods will not scale to meet the challenges of exascale systems and data—holds true in the area of knowledge discovery. Existing approaches for knowledge discovery will not evolve to the exascale. Failure to address the issues of knowledge discovery in the exascale ecosystem will have a profound adverse impact on all science programs.

Knowledge discovery applications tend to be more complex than simulation codes in many respects: they often have graphical user interfaces; they interact with distributed computing and data storage resources; they rely on a comparatively deep “software stack”; they are used interactively and on parallel resources; they place great demands on system resources—consuming vast amounts of memory, I/O bandwidth and processor cycles; and they generate many different forms of output (movies, images, derived quantities, new datasets, etc.). The same arguments for software engineering, performance analysis, and optimization tools that apply to simulation codes also apply to knowledge discovery applications.

As with other technology areas in this report, algorithms and approaches for knowledge discovery in use today are not expected to scale into the exascale regime. Existing approaches will fail because of the sheer volume of data, the complexity of data to be processed, and the growing impedance mismatch between size/complexity and human ability to understand and interact with knowledge discovery infrastructure. A number of different, yet complementary, approaches to address these problems will require exploration: (1) the ability to visualize and analyze results at coarse and fine resolutions to support the natural investigatory process that relies on context/focus interaction; (2) better visual data analysis algorithms for characterizing and presenting uncertainty; (3) integration of visual data presentation and data analysis techniques (e.g., clustering, classification, statistical analysis, and representation) to aid in accelerating knowledge discovery; (4) greater emphasis on the human-computer interface to increase the efficacy of visual presentation motifs and interactive knowledge discovery interaction models; (5) context-centric interfaces to simplify use of complex software infrastructure; and (6) rethinking the design and implementation of fundamental knowledge discovery algorithms and software infrastructure to effectively leverage exascale platforms.

3. State of the Art

Devising ways to address the challenges identified in Section 2 requires a thorough understanding of the state of the art. In this section we review the state of the art in several key areas.

Knowledge Discovery. Algorithms exist for analyzing terabytes of static data stored in a single location, but very few analysis algorithms can handle a dynamic dataset distributed across sites or streamed in live. Feature detection is primitive or nonexistent in many science domains. Human interaction through visualization is today's norm. In addition, differences in the rate of increase of performance in computer technology (such as memory and processors) and in storage technology (such as magnetic disk and tape) will result in the need to achieve unprecedented levels of parallelism to enable the storage systems to keep up with the computers.

Application Building. As computational capabilities have grown, so have the resolution and complexity of the simulation models. Today's large simulation codes incorporate multidiscipline, multiphysics, multiple time scales, and multiple solution methods. They represent years of development by teams of programmers and scientists and can include millions of lines of code. As we make the leap to exascale computation, the cost to update, recode, and incorporate more advanced models into the simulations is an order of magnitude higher than the cost of the supercomputer hardware. In order to contain these costs, the exascale software ecosystem must support more efficient program development.

Programming Models. Current large-scale applications have made the transition from shared-memory vector architectures to 100-TF distributed-memory machines by using the message-passing programming model (MPI) together with traditional sequential languages (C, Fortran, C++). New architectures with many cores per chip are expected to prevent this approach from exploiting exascale hardware. Thus new approaches are needed. For example, today's software is largely MPI-based, with some global view techniques such as Unified Parallel C (UPC) and Co-Array Fortran (CAF). In order to facilitate the use of extreme-scale resources, new "hybrid" programming models, or more radical approaches as represented by the project initiated by the DARPA HPCS program, must be explored.

Data Management. Keeping track of data is already a daunting task. The meaning of the data, referred to as metadata, requires precise capture of how the data was generated and the scientific interpretation of each data item. Furthermore, many scientific datasets are generated from other datasets, or perhaps a combination of datasets. This requires the capability of tracking the history, or provenance, of the data. Today, such tools are provided in ad hoc manner; some metadata is collected in various forms of notebooks, some in databases, and some embedded as headers of files. In the exascale regime, the automation of this task is essential because of the sheer volume of the data and the accelerated rate of their production. Standard metadata models and tools will have to be developed, as well as tools to automatically capture the metadata as the datasets are generated. Furthermore, the data models need to support standard ontology for each scientific domain and allow for dynamic evolution of such standards.

Data Organization. Datasets are organized and stored today as collections of files. The sizes of files are often dictated by the storage systems. For example, current mass storage systems that use tape storage prefer certain fairly large size files, a fact that forces scientists to either aggregate smaller files into larger files or partition very large datasets into multiple files. This approach does not scale and is irrelevant to the scientists. Dealing with the volume of data generated by exascale machines will require support for datasets regardless of their size.

Visualization methods will be critical for handling the growing impedance mismatch between the size/complexity of data and a person's ability to understand and interact with that data.

Debugging Tools. An integral part of application development includes verifying that code runs as expected. Current tools, limited mostly to debuggers, have not yet caught up with the needs of terascale computing, with many application developers for today's large systems falling back to a very inefficient method of debugging—dumping user-inserted debug code to output files. With the vast increase of process count going to exascale systems, searching manually for a single anomalous process among the millions of running processes and threads is not tenable. Moreover, today's debuggers are not scalable to a thousand processors. Scientists will need to have their applications run on millions of processors and require the tools to meet the overall simulation requirements.

Fault Recovery. Modern PCs may run for weeks without needing rebooting. Today's supercomputers often run for only a few days before rebooting, because of their complexity and their thousands of processors. Exascale systems will be even more complex and have millions of processors. The scale of the systems means that component failure is the norm, not the exception. This requires a major shift from today's software infrastructure. Every part of the exascale software ecosystem must be able to cope with frequent faults; otherwise applications will not be able to run to completion. The system software must be designed to detect and adapt to frequent failure of hardware and software components. On today's supercomputers every failure, even those that are reconfigured around, kills the application running on the affected resources. These applications have to be restarted from the beginning or from their last checkpoint. The checkpoint/restart technique will not be an effective way to utilize exascale systems, because checkpointing stresses the I/O system and restarting kills 999,999 running tasks because one fails in a million-task application. With the potential for exascale systems to undergo constant failures somewhere across the system, application software will not be able to rely on checkpoint/restart to cope with faults since a new fault is likely to occur before the application can be restarted. For exascale systems, new fault tolerance paradigms will need to be developed and integrated into both existing and new applications.

Searching manually for a single anomalous process among the millions of running process threads on exascale computers in simply not tenable: new tools are essential.

4. Accelerating Development

Focused investments in the major software, algorithm, and data challenges are needed to scale and tune applications to the exascale. Several flagship projects of a truly interdisciplinary nature that address key national problems should be selected and supported to motivate, guide, and validate the computer science research results. Possible examples are modeling the social impact of climate change, modeling the global nuclear materials cycle, or modeling regional social and economic futures at a new level of detail and reliability.

Software development requires a different usage model from the "production" running of applications and is often delayed by lack of access to scalable resources dedicated to the development process.

Creating an exascale software ecosystem entails more than just coming up with novel solutions. It includes educating scientists on how to use the solutions, both new tools and new approaches, and demonstrating why using these solutions is to their advantage. It includes making sure that the solutions are hardened to production quality so that they can be integrated into the software suites of the nation's supercomputer centers. It includes making pieces available as they are completed, rather than waiting until everything is done. And it includes helping users integrate these pieces into existing codes so that science teams can benefit in the near term and build up trust in the solutions being provided for the exascale software ecosystem.

To help applications make effective use of exascale systems, we need to go well beyond the current state of affairs in the performance analysis process. Today's tools are limited in scope, capability, and scalability, and feedback to the application is manual. The overhead associated with current measurement techniques is too intrusive at this size and may skew analysis so much as to render any analysis ineffective. Therefore, we need to develop scalable and less intrusive methods of collecting performance data, develop knowledge discovery methods for extracting key performance features, and provide assistance in feeding the results of these analyses back to the code transformation.

Probability says that exascale systems will spontaneously change answers to wrong values occasionally. Formal verification methods are needed to ensure that calculations are done correctly.

For simulation codes to be able to run correctly as well as effectively, in a reasonable time frame, an investment in developing formal verification methods is essential. The scale and complexity of the science problems enabled by exascale systems require new techniques for making sure that the calculations are done correctly. Cosmic rays have been shown to change the memory values in a supercomputer, causing wrong answers to be calculated. In another supercomputer it was discovered that in one in a billion data transfers, correct values sent to another processor would emerge from the wire as wrong values. Probability says that exascale systems with memories bigger than any computer on the planet and the ability to calculate and send trillions of bits of data per second will spontaneously change answers to the wrong values occasionally. One approach is to incorporate algorithms into the system software and have the application software catch and correct any spontaneous mistakes caused by phenomena such as cosmic rays.

Investment in the reliability and robustness of exascale systems for running large simulations is critical to the effective use of these systems. New paradigms must be developed for handling faults within both the system software and user applications. Equally important are new approaches for integrating detection algorithms in both the hardware and software and new techniques to help simulations adapt to faults.

Knowledge discovery R&D programs are crucial to scientific discovery at the exascale Many potential approaches merit exploration. One is to include some knowledge discovery software in the simulation itself. This approach could have the benefit of reducing the I/O load for the simulation. Another approach entails a closer coupling between knowledge discovery and related technologies, such as data management (I/O, movement, index/search) and analysis. No "one size fits all" formula will meet the needs of all of science. Because of the diversity and complexity of knowledge discovery technologies, a broad technical portfolio of projects will help to maximize the likelihood of success.

5. Expected Outcomes

Successful development of the flagship applications identified earlier would be of real benefit to the government and society. Simulation could inform policy development in a number of diverse areas, including emergency planning, health system improvement, and management of climate change.

The development of predictive power in validated models connected to real-time data would also provide economic value to businesses that must respond to rapid changes in the operating environment, such as airlines, power companies, and emergency management operations.

But without doubt the major impact of this initiative is a software infrastructure that enables new applications beyond the initial flagship applications. This software infrastructure will enable a phase transition in large-scale scientific simulation and modeling.

6. Required Investment

As noted in Section 4, focused investments in the major software, algorithm, and date challenges are needed. An investment of $100M over five years would make a real impact on these challenges. The exact details of its distribution would need to be worked out to respond to realistic proposals.

7. Major Risks

One major risk is a disruptive shift in hardware technology in the next 5–10 years that facilitates a complete change in the approach to data analysis, programmability, and interactive computing. Another risk is the sheer complexity of the software systems eliciting unexpected problems at the scale of the computers and volume of data produced by simulations and experiment in this time frame.

The risk in allowing current trends to continue slowly is that while scientists will envision breakthrough computations and the requisite hardware will be available, the software infrastructure for programming, executing, and understanding the results of these exascale computations will be inadequate to the task.

References

Ray Bair, Lori Diachin, Stephen Kent, George Michaels, Anthony Mezzacappa, Richard Mount, Ruth Pordes, Larry Rahn, Arie Shoshani, Rick Stevens, and Dean Williams (2003), Planning ASCR/Office of Science data-management strategy. http://www-conf.slac.stanford.edu/dmw2004/docs/DM-strategy-final.doc.

J. Duell, P. Hargrove and E. Roman (2002), The Design and Implementation of Berkeley Lab's Linux Checkpoint/Restart, Berkeley Lab Technical Report LBNL-54941.

R. L. Graham, S.-E. Choi, D. J. Daniel, N. N. Desai, R. G. Minnich, C. E. Rasmussen, L. D. Risinger and M. W. Sukalksi (2003), A network-failure-tolerant message-passing system for terascale clusters, *Int. J. Parallel Programming* 31(4) 285-303.

W. Gropp and Ewing Lusk (2004), Fault tolerance in Message Passing Interface programs, *Int. J. High Perform. Comput. Appl.* 18(3):363-372.

J. Li, W. Liao, A. Choudhary, R. Ross, R. Thakur, W. Gropp, R. Latham, A. Siegel, B. Gallagher, and M. Zingale (2003), Parallel netCDF: A high-performance scientific I/O interface, in *Proc. SC2003*, Phoenix, AZ.

E. Lusk and Kathy Yelick (2007), Languages for high-productivity computing: The DARPA HPCS language project, *Parallel Processing Letters* 17(1), March.

J. Parker, T. Engelsiepen, R. Ross, R. Thakur, R. Latham, and W. Gropp (2006), High-performance file I/O for the BlueGene/L supercomputer, in *Proc. 12th International Symposium on High-Performance Computer Architecture* (HPCA-12).

R. Ross, Jose Moreira, Kim Cupps, and Wayne Pfeiffer (2006), Parallel I/O on the IBM Blue Gene /L system, BlueGene Consortium Quarterly Newsletter, February.

Rob Ross, Evan Felix, Bill Loewe, Lee Ward, Gary Grider and Rob Hill (2005), HPC file systems and scalable I/O: Suggested research and development topics for the fiscal 2005–2009 time frame, ftp://ftp.lanl.gov/public/ggrider/HEC-IWG-FS-IO-Workshop-08-15-2005/FileSystems-DTS-SIO-FY05-FY09-R&D-topics-final.pdf.

Douglas Thain, Todd Tannenbaum and Miron Livny (2005), Distributed computing in practice: The Condor experience, *Concurrency – Practice and Experience* 17(2–4) 323-356.

8 Hardware

Exascale computing requires innovation at the frontiers of computer architecture and information technology. Major challenges exist in providing sufficient compute performance per watt, memory performance, interconnect performance, and I/O and persistent storage performance and reliability. In addition, evidence suggests that both software and algorithms will need to be developed *in concert* with hardware to ensure that both legacy and new exascale applications will be able to make effective use of the advanced systems. In addition, since such systems are likely to have 10 million to 100 million separate processing elements and 10 to 100 petabytes (PB) of memory, fault detection and handling by hardware, software, and algorithms will be essential. With sufficient investment, particularly in ongoing point-design studies and hardware-assisted simulation, exascale systems can become an effective resource for attacking complex science and engineering problems.

1. Advances in the Next Decade

To understand that challenges of constructing an exascale system, one must first understand where current technology trends are leading the community. While there is always risk in extrapolation, the 14-year history of the Top 500 data (Figure 8.1) suggests that without further technology acceleration the first system exceeding 1 exaflop LINPACK performance can be expected by 2019. Based on past programs, it is therefore believable that a concerted, sufficiently funded, multiyear research program should be able to accelerate the availability of exaflop technology by four years, to the 2015 timeframe. Concurrency in existing systems continues to be a primary factor in these systems' increasing capabilities. This trend is sure to continue in the near future. The current rate of growth points toward a concurrency level for the high end of 10^7 cores by 2015 and 10^8 cores by 2020. An additional switch from multicore to manycore technology during this time frame could easily increase these levels by an order of magnitude, from 10^8 to 10^9 cores.

Many of these system trends (and their consequences) can be explored further by examining the growth trends for critical performance factors for the commodity component technologies. Based on historical trends, in 2015, one potential exascale system could require 1.3M PEs, where each PE generates 768 gigaflops, derived from 64 cores, where each core produces 4 flops per cycle and operates at 3 GHz. Assuming that the PEs would require 100 W each, the resulting exascale system alone would require 130 MW of power. Assuming 1 GB per core, the resulting exascale system would have 83 PB of memory, composed of about 5 million memory parts. The bisection bandwidth and scale of the interconnect network will depend on the topology and the switch/router configurations, but they will also be limited by the signaling and bandwidth rates. Also, the scale of the entire system will be limited by the physical length of cables in the entire system. Storage systems will continue dramatic increases in capacity (approximately 50% compound annual growth rate), while disk bandwidth and seek latencies will improve only slightly, because of their mechanical characteristics. An exascale system capable of checkpointing 50% of its physical memory in 10 minutes would require approximately 250,000 disks.

Exascale systems are likely to have up to 100 petabytes of memory, enabling new applications but also requiring major innovations in software and algorithms.

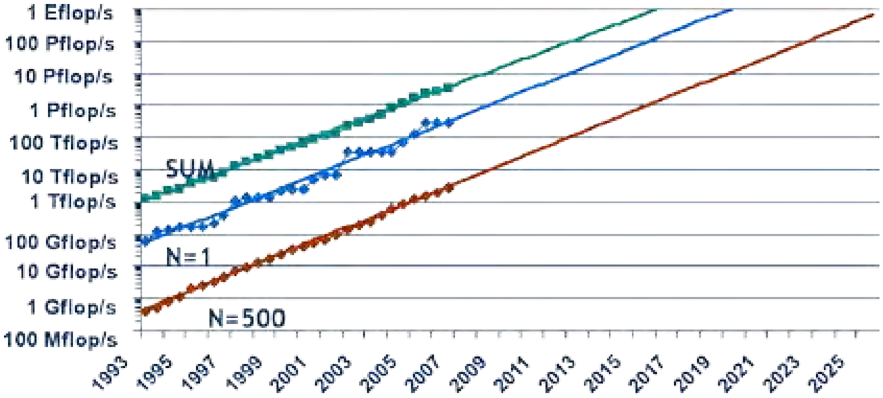


Figure 8.1 Actual and projected Top 500 performance.

The slowing pace of commodity microprocessor performance improvements, combined with ever-increasing chip power demands, has become of utmost concern to computational scientists. As a result, the HPC community is examining other approaches that address the limitations of conventional large-scale computing systems in the coming decade. Already, commodity processors for graphics and computer games achieve far higher computational peaks per socket and per watt than conventional processors (this includes both GPGPUs and the Sony/Toshiba/IBM Cell processor). These processors have many drawbacks for use in scientific computing, including single-precision arithmetic, but could easily evolve over the next decade to provide suitable computing platforms. Embedded processors also provide very high efficiencies in terms of operations per watt; in fact, the processor in IBM's Blue Gene machines (the world's fastest computer in 2007) is a System on Chip (SOC) design employing conventional PPC440 embedded processor cores. Similarly, the SiCortex system employs SOCs with MIPS/embedded processor cores to achieve much lower power utilization for sustained performance [Reilly et al. 2006]. Like BG/L, however, this approach accelerates the march toward massive concurrency and motivates fundamental advances in applied mathematics to develop scalable algorithms. Other recent examples of specialized systems include the MD-Grape system in Japan, developed for less than $10M. A semi-customized architecture for molecular dynamics under development at D.E. Shaw (Shaw et al. 2007) promises to speed high-fidelity protein-folding simulations by several years. A different semi-customized architecture for climate computing [Wehner et al. 2007] shows promise of enabling kilometer-scale climate modeling decades before it would be possible through the conventional Moore's law technology improvements.

Three designs, summarized in Table 8.1, illustrate some of the challenges for exascale computing. New processor and system architectures that are optimized for better power efficiency will be necessary if total system power is to be kept under 30 MW. Since much of the performance will come from orders of magnitude more processor cores than in the current systems, innovations in hardware, software, and algorithms will be needed to make effective use of these systems.

While most of the above discussion has focused on the processing elements, a capability platform must also provide access to a sufficiently large memory hierarchy, an effective interconnect between the processors, reasonable I/O capabilities, and networking to the outside. For example, in order to achieve sustained performance in the 10–20% range on applications such as computational fluid dynamics (direct numerical simulation), high-energy physics (quantum chromodynamics) and computational biology (molecular dynamics), our in-depth understanding of today's algorithmic and subsequent communication requirements drive a system balance requiring injection bandwidth on the order of 150 GB/s, order 1.5 PB/s global bandwidth, and order 500 nanoseconds latency. These drive the technology requirements in the areas of processor, interconnect, I/O, and storage design. The applications will also require 0.5–2 bytes/flop of memory bandwidth, and we should assume at least 2–4 GB of memory per CPU core.

2. Major Challenges

Architecture developers face a number of challenges. These involve two broad questions: How should one engineer the major system components, and what are the implications of these designs on software and application development? Several issues must be addressed:

Example system	Ops/ cycle	Freq [GHz]	Cores/ socket	Peak/ socket [TF/s]	Sockets	Total cores	Peak/ system [EF/s]	Power [MW]
A	4	3.0	64	0.768	1300k	85M	1.0	130
B	8	16.0	128	16.0	120k	15M	2.0	60 - 80
C	8	1.5	512	6.1	200k	100M	1.8	20 - 40

Table 8.1 Some key parameters of exascale systems based on rough projections from current technologies.

- Sufficient effective interconnect performance, particularly latency, bandwidth, and cross-section bandwidth
- I/O (persistent store and file systems), particularly data density, transfer bandwidth, and fault management
- Memory (main system memory), particularly cost and size, power efficiency, access latency, bandwidth, and number of hierarchy levels
- Power consumption everywhere
- Processor architecture and implementation, particularly latency and bandwidth management, concurrency levels, heterogeneous designs, and fault frequencies
- Software and algorithms, particularly as they provide workarounds and alternative approaches to deal with latency, bandwidth, memory hierarchy, faults, heterogeneity, and highly increased concurrency levels

In addition, a number of crosscutting issues need to be addressed.

- Better characterization of algorithm requirements with respect to system ratios
- New algorithms to match system ratios, particularly between disk and main memory and the bandwidth of the interconnect to the flops rate of the processor
- New algorithms and software to detect and handle faults
- New approaches to algorithms and software for specialized/disruptive processor architectures; that is, better methods for moving applications to GPGPUs, PIMs, FPGAs, heterogeneous systems, or other less-conventional architectures (these would help open up the space of hardware approaches)
- Acceleration of applications and algorithms (especially new ones) to petaflops now to prepare for exaflops
- Programming languages and environments
  - PGAS, domain-specific, auto-tuner, hierarchical programming models (built on current models)
  - Interaction with hardware (e.g., user-managed caches, remote atomic updates)
  - Performance modeling and debugging Productivity
  - System software, operating system (e.g., memory management)

Because of cost and power limitation, an exascale system by the year 2015 will likely *not* be a general-purpose system in the current sense. The design of such a system will require some compromises with respect to size of system resources such as memory or disk space and with respect to system ratios such as cross-section bandwidth to link bandwidth to memory bandwidth. A sufficient understanding of the requirements of exascale applications is needed to guide the choices necessary for these compromises, so that the final design supports a reasonably large fraction of all potential applications.

Future storage architectures will have to be designed to be an integral part of the overall system. Currently I/O is an afterthought, which is then “bolted” onto a system. The bandwidth requirements for systems of this size will prohibit this behavior in the future for balanced systems. Also needed are the development of new storage semantics that enable high performance and scalability, the integration of database concepts (relational and object models) into the notions of file systems and storage models, and the creation of effective associative access methods for integrated memory/data systems.

3. State of the Art

Today's state of the art is reflected in several approaches for petascale systems:

- ***Strongly increased concurrency levels*** together with reduced clock-rate to manage power consumption; several specialized system networks (including collective networks); commodity processing-core architectures, but integrated in custom chips (IBM BG/C, L, P, Q; SiCortex (www.sicortex.com); DoD-funded processors). These changes in system design require a heavy investment in adapting the software infrastructure. The ability of rapid integration of commodity cores in custom processor chips is opening up a large potential for semi-specialized scientific computing systems with high performance and power efficiency.

- ***Modified instruction sets and memory architectures*** that attempt to better match current memory hierarchies and in some cases to increase power efficiency (IBM/Sony/Toshiba Cell; stream processors; PIM; Cray Black Widow and Cray Eldorado/SUN Niagara; GPGPU). These approaches require even heavier investment in the entire software development chain.

- ***Commodity processors plus specialized networks*** [Cray XT/Cascade, IBM PERCS [Elnozahy 2006] (HPCS)] follow established engineering models but currently face increasing power efficiency problems.

Balanced exascale systems will require a new approach to I/O — which currently is simply "bolted" onto a system.

- ***Reconfigurable processors and FPGAs*** are still plagued with a very limited software development environment. They do show potential for rapid large-scale system simulation and as platforms for early application development and performance evaluation (e.g., Berkeley RAMP [Krasnov et al. 2007] and emulators for commercial FPGA-ASIC design flows from companies such as Altera [Blyler 2005]).

- ***Quantum computing technology*** solves a set of problems that are orthogonal to the approach taken by existing computational technology. Therefore, quantum computing offers an opportunity to expand the scope of problems that can be addressed with computing as a complementary approach to computing rather than supplanting current computational methods. There has been quite a bit of progress in the basic and underlying technologies, but to date nothing has been realized beyond the basic hardware layer and very small (<16 qbit) systems. Maintaining quantum entanglement for larger-scale hardware has proven difficult, but start-up companies such as D-Wave [Maxcer 2007] have taken the first steps toward commercialization of quantum computing technology. A breakthrough in quantum computing would imply an enormous software investment for anything but critical handwritten or hand-tuned applications.

Today's software for high-end machines is largely based on message passing, sometimes using a thread model for each MPI process on a shared-memory node. As machines grow in scale and complexity, techniques to make the most effective use of network, memory, and processor resources will also become increasingly important. Programming models that rely on one-sided communication or global address space support have demonstrated advantages for productivity and performance, but they are most effective when used with proper hardware and operating system support. Global view languages, such as UPC and CAF, are seeing some use and are likely to be important in the next 5–10 years (see, for example, Coarfa et al. 2006).

Many feel that new programming models and languages are needed for machines at this scale. The same view was held for petascale machines, yet clever use of hierarchical approaches and good algorithms have allowed the simple message-passing approach to succeed with the current near-petascale systems. The DARPA HPCS program [Johnson 2007] has fostered the development of three new languages: Chapel, Fortress, and X10. Exploration of these languages is important, particularly as they promise enhanced productivity and could help increase the number of application areas that made effective use of high-end computing. We note, however, that it typically takes 10 years for a language to reach the level of maturity required by most applications.

Many promising technical developments can address these issues. A sampling is shown in Table 8.2.

Technology	Issue Addressed
Optimizing the use of die space for CPU	Power efficiency, sustained performance
Optical network	Interconnect performance, system scalability
Optical in/out of processor	Interconnect performance, power efficiency
3D chips; integrated memory/processor	Memory performance, packaging density, scalable system hierarchy
Faster development of customized processors	Technology acceleration, memory performance, system scalability
Hardware accelerated system verification (e.g., RAMP)	Software and algorithm development, performance modeling and predictions, accelerated system design
NAND Flash, MRAM, other non-volatile memory	Scalable system hierarchy, power efficiency, I/O performance
Myriad approaches to power efficiency	Power efficiency

Table 8.2 Promising technologies for exascale computing.

4. Accelerating Development

Accelerating the availability of exascale computing requires concerted efforts in several areas. Early evaluations of different potential technologies are key for technology acceleration. These include evaluations of current and new hardware, software, and application designs. The results of such evaluations must be presented and shared openly. Only by iteratively improving on one another's research can the high-performance computing community as a whole make the most rapid progress toward a common goal. In particular, the following steps can be taken to accelerate development.

- Develop technology assessments (5 years out) and "point design studies" (10 years out) that investigate in more detail the specific challenges in different design approaches, including all-commodity and partial custom designs. These design studies should represent a continuous effort to understand future technologies and their impact and should follow a common evaluation methodology, which needs to be established. They should evaluate the impact of specific hardware components on operating systems, algorithm designs, and application performance as well as their impact on the application development process in general. These design studies are essential to provide more detailed analysis and guidance about program directions at all stages of this process.

- Establish ongoing research for early simulation and modeling of future hardware and the impact of promising or disruptive technologies. This will require modeling the elements of a system as well as modeling entire systems. Existing simulation efforts need further support to extend their capabilities for the required scale and to be available in the future whenever new technologies are proposed. Promising approaches for different levels of simulation include FPGA-based systems (e.g., Berkeley RAMP) and highly parallel simulation software on conventional systems.

- Develop new algorithms and supporting software. The rapidly increasing concurrency levels and the increasing complexity of system hierarchies pose tremendous challenges to application and basic software developers. These challenges need to be addressed as soon as possible to give these communities time to have their software ready as soon as exascale hardware becomes available.

- Conduct point design studies and early application development. These areas each require advanced system and performance modeling and evaluation methodologies. Only new evaluation methodologies based on application requirement can provide accurate feedback about utility of new technologies. Advanced performance modeling and prediction methodologies are necessary to aid evaluation of future systems and to guide application development. These evaluation and modeling methodologies have to be combined with simulators, both cycle accurate and system accurate, for hardware, software, and application development and assessments.

- Invest in truly innovative hardware (with a 5- to 10-year development horizon). Such research can be facilitated through research partnerships between laboratories, universities, and end users, along with the vendors who would be the primary developers of full systems. Research finding should be committed to a broad set of paths with checkpoints and incentives for acceleration of technology into production systems. Currently such efforts remain unfunded from the government perspective in various technologies. Without funding, necessary technologies will not be available in time.

- Invest in hardware emulation technologies that make research into innovative hardware design more cost-effective for academic researchers Such environments also accelerate the feedback loop between hardware and software development.

- Commit to building and experimenting with small-scale systems for node architecture changes and massively parallel (but lower-performance) systems for scaling studies. Such early prototypes should complement modeling and simulation efforts. They allow more precise technology assessments and will be essential for facilitating early software and application developments.

- Provide significant incentives for the early participation of applications in design and evaluation efforts. Commit to a deployment timetable and roadmap that enable vendors to shift some development risk to early systems revenue.

- Find the optimal point between fully general (one solution that does nothing well) and fully specialized (solves one problem but does not help with any others) in hardware, software, and algorithms. This technique has proven itself already in many cases. Additional funding to enhance an already planned product line allows the government to optimize its funding strategy and fits well with the proposed solution technologies. In some sense, algorithms are already here, and the hardware is moving in this direction (e.g., Eldorado); software, particularly languages, needs to take this approach with domain-specific approaches.

Multiple prototypes should be built and evaluated to guide development of exascale systems.

Efforts should embrace open source for all aspects (not just software) to allow the maximum leverage from each advance. Multiple prototypes should be built and evaluated; costs can be managed by building slices of potential systems. A framework that follows efforts from the development of a point design to a full system without multiple points of evaluation will be essential in guiding the development of an exascale system.

5. Expected Outcomes

Increased investment could have a positive impact—substantial investments would steer innovation in hardware, software, systems design, and the necessary infrastructures, while smaller investments will not achieve the critical mass needed to effect the necessary changes.

Accelerating the development of the building blocks for exascale systems would have dramatic commercial spin-offs outside the scope of the initiative, into areas such as medical image processing, intelligent sensors, and more capable robotics, including the prospect of fully automated, fully capable vehicles and aircraft. The improvement in the ability to broadly develop applications for parallel systems will have an overall boosting effect on the U.S. software development enterprise, which is currently the most productive by far. It would also help to provide a high-value trajectory for U.S. software development, as the lower-value elements are more easily migrated offshore. This would reaffirm the lead once held by the United States in many technological areas where we are no longer considered a competitor.

6. Required Investment

At a minimum the program of R&D on the hardware-related and associated systems-related technologies would need to ramp to approximately $100M–$150M per year. Early progress could be made for less than that in launching the innovation efforts. The deployment of large-scale systems from each generation of development would require approximately another $150M per year (assuming three rotating deployments of $50M per year each).

7. Major Risks

We have identified four major areas of potential risk in establishing an exascale hardware initiative.

Failing to establish vendor buy-in. Industry is likely to be a major participant in development; in addition, industry will need to develop and bring these systems to market. Industry must believe there is a long-term federal commitment to deployment to be able to provide substantial cost-match during the development. The hardware development program should include both large and small companies and also a range of technologies from somewhat risky to bleeding edge. Some of these are already being discussed, have vendor acknowledgment, and are looking for funding agents. They could be brought on line in less than 12 months.

Not aiming high enough. Research must be aimed at the long term to achieve transformational impact, and it needs to be stably supported to achieve the innovative breakthroughs we are looking for. Focus on near-term advances will stifle innovation. We also must not underestimate the scale of the potential impact of the combined exponentials of improvements in processor performance, storage capacity, networking, and user interfaces. The rate of improvement of these four factors will mean that things will be dramatically different 10–15 years from now. We must also have a component of the program looking at dramatically alternative technologies (quantum, biological, nano, combination, etc.)

Failing to couple hardware, software, and applications. It is critical that the coupling of the critical elements for success happen early and often in this program. The hardware architects, the operating and tool developers, and the scientific users must all be engaged from the beginning to provide the critical feedback loops that are needed for true innovation.

True innovation demands that hardware architects, software developers, and users all be engaged from the start.

References

J. Blyler (2005), Navigating the Silicon Jungle: FPGA or ASIS, Chip Design, June/July, http://www.chipdesignmag.com/display.php?articleId=115&issueId=11

C. Coarfa, Y. Dotsenko, J. Mellor-Crummey, F. Cantonnet, T. El-Ghazawi, A. Mohanty and Y. Yao (2006), An evaluation of global address space languages: Co-array Fortran and Unified Parallel C, in *Proc. Principles and Practice of Parallel Programming (PPoPP),* New York.

M. Elnozahy (2006), IBM has its PERCS, HPCWire 15(14), April. http://www.hpcwire.com/hpc/614724.html.

F. Johnson (2006), DARPA HPCS Program, *SciDAC Review 2*, Fall, http://www.scidacreview.org/0602/html/news2.html

A. Krasnov, Andrew Schultz, John Wawrzynek, Greg Gibeling and Pierre-Yves Droz *(2007),* RAMP Blue: A message-passing manycore system in FPGAs, in *Proc. FPL 2007 - International Conference on Field Programmable Logic and Applications*, Amsterdam.

C. Maxcer (2007), D-Wave claims quantum computing breakthrough, *TechNewsWorld*, www.technewsworld.com/story/55801.html.

M. Reilly, L. C. Stewart, J. Leonard and D. Gingold (2006), SiCortex technical summary, www.sicortex.com/whitepapers/sicortex-tech_summary.pdf.

D. E. Shaw et al. (2007), Anton, a special-purpose machine for molecular dynamics simulations, in *Proc. 34th Annual International Symposium on Computer Architecture*, San Diego, pp. 1-12.

M. Wehner, L. Oklier, and J. Shalf (2007), Towards ultra-high performance resolution models of Climate and Weather, *Intl. J. High-Performance Comp. Apps*., to appear.

9 Cyberinfrastructure

Today's cyberinfrastructure is a loose coupling of libraries, tools, and policies that supports researchers by enabling the application of multiple, distributed resources to a scientific problem. Even for the most basic workflow, such as collecting data, performing a high-performance computation, and analyzing the result, it is rare that all of the necessary people, computing, storage, and analysis systems are within the same location. Beyond the relatively simple problem of moving data across networks is the much more challenging issue of managing secure access across multiple organizational boundaries. Today's cyberinfrastructure is already showing signs of weakness from the standpoint of technical capabilities (data movement and analysis, for example) in the transition from terascale to petascale science, suggesting that either our approach must be radically adjusted or we must accelerate the development of solutions to these challenges, or both. Beyond technical challenges associated with speed and size of resources, we see distributed science teams growing larger. This situation, combined with the distributed nature of the resources, presents a fundamental difficulty with respect to authorization and access. Traditional HPC has been a client-server game, with a set of one-to-one trust relationships between independent computing centers and individual users. As we move toward exascale science, these one-to-one relationships transform into many-to-many matrices, with distributed multidisciplinary teams working together to harness resources from multiple service providers.

Cyberinfrastructure provides the foundation for a wide class of users, from scientists to system administrators to end users. This foundation must be well defined, secure, persistent, robust, and increasingly transparent to the researchers building on top of it. Ensuring such a cyberinfrastructure, and addressing the related cyber security issues, will require considerable investment to support future research. Areas to be investigated include workflow management; collaboration frameworks and techniques; data management and movement of exascale datasets and data collections; authorization and authentication for flexible interdisciplinary computational science teams ("virtual organizations"); and cyberinfrastructure management tools, techniques, and methodologies to understand the performance of the infrastructure and to protect it from attack, disruption, and data loss.

1. Advances in the Next Decade

The challenges associated with cyberinfrastructure are driven by the increased size and complexity of applications and datasets; the need to combine computational, experimental, information, and analysis resources to support the scientific workflow; and the increasingly distributed and multidisciplinary nature of the teams that will be tackling these problems.

Five critical cyberinfrastructure areas have been identified as necessary and feasible to support exascale science: workflow management, collaboration frameworks and techniques, data management and movement, cyberinfrastructure management tools, and cyber security. Success of these teams in working together and harnessing an inherently distributed set of resources will have major impact on the world we live in and America's role in such a world.

Unlike traditional high-performance computing, exascale science will involve many-to-many trust relationships, with distributed multidisciplinary teams harnessing resources from multiple service providers.

Workflow management. Exascale applications will generate vast amounts of data. The process of analyzing that data will require correspondingly vast computational power. Equally important, the systems, toolkits, and user interfaces through which scientists will explore and analyze the data must be engineered for automated end-to-end data transport, resource management, and security and integrity. Workflows comprise the computations and data analysis tasks that are composed of a (potentially vast and complex) sequence of related but distinct jobs, including one or more data preprocessing steps (e.g., data assimilation and cleaning), followed by a series of computational steps related by complex dependences, followed by postanalysis. Because the engineering of such workflow systems at the exascale is a daunting problem in both architecture and software development and quality assurance, it will be vital for workflow components such as job management, dataset or data collection management, and account and access control management to be shared and used across many diverse science communities (for recent work on community accounts, see [Welch et al. 2006]).

Workflow management at the exascale is a daunting problem in both architecture and software development.

Collaboration frameworks and techniques. Scientific pursuits have become increasingly compute-intensive, collaborative, and geographically dispersed; exascale science will continue this trend with unprecedented distributed collaborations involving scientists, their data, and the compute resources. Tools currently available for collaborative science center largely on interactive presentations using video- and web-conferencing for synchronous collaboration and email, wikis, and blogs for asynchronous collaboration. The need for collaborative technologies is evident in the recent explosion of tools for web collaboration, integration of collaboration into standard business tools, and the success of commercial web-conferencing offerings. Scientists should have a current image of the state of their collaboration available to them in their personal working environment (e.g., laptop, desktop) at all times. To the extent possible, they should be able to interact with their local state when they are offline, and synchronize this state with the appropriate group state when they return online (e.g., as Google Gears does for web applications). Such a process should happen in as natural an environment as possible, so that scientists perceive practically no change in their personal working environment, yet interact with and benefit from the current shared state of the collaboration (see, e.g., [Stevens, Papka, and Disz 2003]).

Data management and movement. Exascale science will generate data at rapidly increasing rates, causing both short- and long-term challenges that data management and movement will need to address. Researchers today are already creating terascale and petascale datasets and discovering that they are spending a significant portion of their time managing data rather than performing scientific investigation. Data management and movement tools for exascale datasets and data collections must provide capabilities for data virtualization [Nefedova et al. 2006], management of properties and attributes independent of the underlying storage system, access control, provenance metadata, data integrity, and data security.

Cyberinfrastructure management tools. The computational systems that will enable exascale science will be orders of magnitude more complex than today's resources. Performance tuning, troubleshooting, and systems management capabilities on today's terascale systems are barely keeping up with adequately monitoring, reporting usage, and providing troubleshooting capabilities for high-performance networks, storage area networks, HPC systems made up of hundreds or thousands of nodes, and archival systems made up of multiple petabytes of storage resources. Tools that support local and remote job monitoring, system monitoring, configuration management, and troubleshooting are necessary for the well-being of these expensive resources. Many of the systems at DOE supercomputer centers and data centers are unique resources that expand far beyond the development environments provided by the vendors for their system management and performance tool developers. Often, system owners develop the tools for these specialized systems either on their own or in cooperation with grid projects. These tools, whether developed by the system owners or the vendors, are not capable of scaling to exascale resources. R&D is required to develop new methodologies, techniques, and tools capable of managing these complex systems. In addition, coupling this effort with the data management and movement tools developed for exascale datasets and data collections will prove useful for dealing with the huge quantities of environmental, systems events, and job management data that will be generated by these systems.

Cyber security. Currently every site performs some level of security monitoring using a combination of network and host intrusion detection systems (IDSs), intrusion prevention systems (IPSs), network flow monitoring, vulnerability scanning, and so on. It is essential that the network and host monitoring mechanisms, such as IDS, IPS, and firewalls, scale to support the dynamic and high-speed networking environment that will be required for exascale computing. For large facilities, 100 Gbps networks will be common in 5 to 10 years. Today's commercial security offerings show no signs of being able to keep up.

2. Major Challenges

The scientific and technological efforts proposed as part of this initiative will pose major challenges for cyberinfrastructure and cyber security. R&D timelines will require 5 to 10 years in order to ensure effective scaling and efficient use of systems at the exascale.

Workflow management. Significant improvement in workflow management at the exascale will require research into the identification of common workflow requirements and problems in the different science disciplines and communities of users. Research will be required in order to understand fundamental workflow issues across science disciplines and to identify effective solutions (for recent efforts, see von Laszewski et al. 2007; Zhao, Wilde, and Foster 2007]. The development of tools and methods that follow recognized standards and find widespread adoption will be a challenge, if we are to prevent waste of scarce resources and avoid reinvention of solutions to common problems that may already be solved.

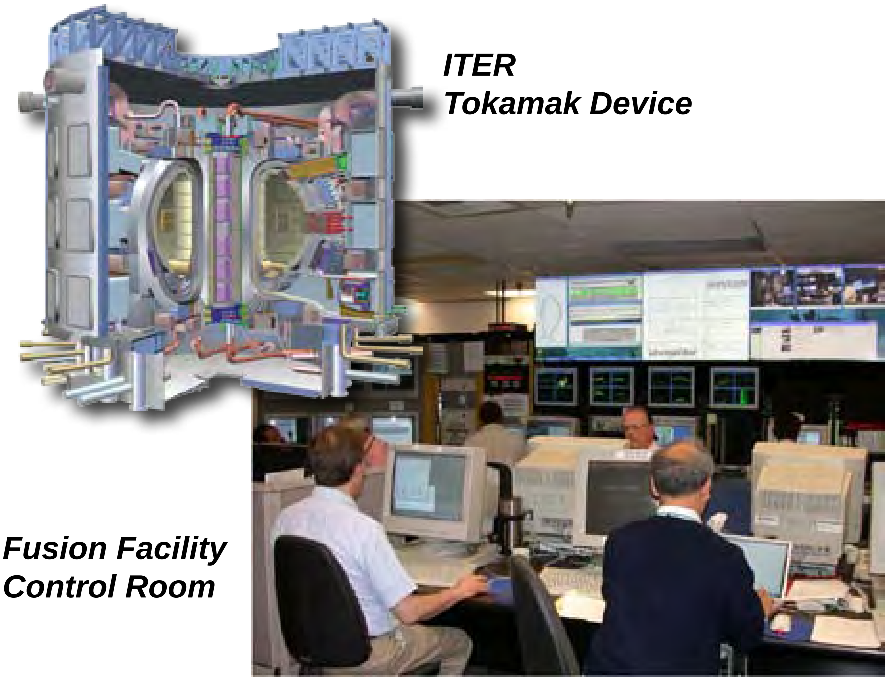


Figure 9.1 Large distributed teams collaborate to run experiments at fusion facilities. Remote participation and collaboration will become even more critical to the U.S. fusion science community, because the next-generation fusion device, the international ITER project, will not be located in the United States.

Collaboration frameworks and techniques. Scientific collaborations increasingly involve more people, more computers, and larger datasets collected across greater distances than ever before. The challenge with the greatest user visibility is representing this information and resource overload in a natural, usable, collaborative manner so that it is understandable and accessible to the researchers involved.

Collaboration-based authentication and authorization. The cross-site and international nature of DOE Office of Science collaborations demands a well-managed, scalable, flexible, and federated approach to authentication and authorization and to the creation and management of the virtual organizations that manage collaboration resources. Current approaches are disparate and impose a high overhead on scientists and security professionals, without producing a high assurance that proper authentication of user identity has been achieved. There is inadequate support for the specialized needs of distributed collaborations, such as dynamic assignment and management of attributes such as roles and group membership to individuals and resources in a virtual organization. Interorganizational trust

We envision a cybersecurity situational awareness capability, "SitAware," that provides accurate summaries of attack traffic and analyses that help in deploying countermeasures.

is fundamental to the operation of virtual organizations.

Data management and movement. Management of large datasets and data collections related to scientific research and related applications have presented significant challenges at the terascale and petascale [Allcock et al. 2001]. This situation is expected to be even more unwieldy at the exascale. Data management and movement tools and techniques must be developed for data centers and archives, portals, and intersite and intrasite file transfers.

Cyberinfrastructure management. Exascale systems are expected to be increasingly complex, comprising thousands or even millions of components. The configuration, verification, troubleshooting, and management of such complex systems, as well as the development of the tools necessary to perform these functions on the often one-of-a-kind resources, will be a significant challenge.

Cyber security. IDSs, firewalls, vulnerability scanners, and other components that make up the cyber security infrastructure of the expected state-of-the-art exascale resources will not have commercial products available that scale to line rates or capacities for months or even years after the exascale resources are deployed. The challenge will be to provide adequate security functionality at the exascale with open source or locally developed tools. Investment in cyber security tools that can be shared by the laboratories seems a plausible option.

Situational awareness, anomaly detection, and response. A key challenge to cyber security methodologies and tools of the future will be the creation of a framework and semantics for integrating information in the individual cyber security component systems for situational awareness, anomaly detection, and intrusion response. Current technologies are segregated and unaware of the related information available in their audit trails. Automated, intelligent tools are needed to detect anomalies in the behavior of users, jobs and processes, applications, and services that scale from system, department, and enterprise to multiple sites. We envision a cybersecurity situational awareness (SitAware) capability to provide analysts with accurate, terse summaries of attack traffic, organized to highlight the most prominent facets. These summaries would include the essential elements that define the pattern of the attack traffic so they can be easily shared with and understood by other sites without disclosing private information. SitAware should also supplement these reports with drill-down analysis to facilitate countermeasure deployment and forensic study. A critical feature of a SitAware capability would be an overall view across multiple sites of a collaboration or virtual organization. What is needed is a distributed cooperative security monitoring framework that can combine security-related data from many sites. This will allow independent sites to extend their sense of the Internet's activity beyond their local viewpoints. Even more important, advanced capabilities in this area will facilitate collaborative threat response, including cross-site notification of and response to security events anywhere in the collaboration environment.

Since exascale compute resources are extremely valuable and a tempting target of attacks, we must develop advanced security data analysis capabilities to facilitate collaborative threat response. These capabilities are essential for such crucial tasks as cross-site notification of security events and subsequent response to these events. Related research topics include audit frameworks, data-mining algorithms for security logs, data visualization techniques for exploration of log data, anomaly detection techniques to help identify suspicious activity, data anonymization techniques to allow cross-site sharing of log data, incident profiling techniques, and ontology-based forensic analysis of security data and logs.

"Sandbox" Technologies: Many DOE Office of Science HPC systems have changed from "low" security categorization to "moderate" security categorization as the resource has become unique and desirable by industry. This trend is expected to continue. The current DOE policy is to certify systems to the highest level of risk for the system and implement the corresponding security control level, forcing all who desire to use the resource to conform, regardless of the risk level of their individual work. These types of tight controls present real barriers to the science community interested in participation in open science. The concept of a "sandbox" is introduced to provide a potential solution. Sandboxes are defined, bounded virtual environments in which virtual organizations (VOs) collaborate and share information. Each VO, regardless of the heterogeneity and physical locations of its members, has its own sandbox and is protected based on its unique information sensitivity and specific constraints (e.g., confidentiality, integrity, availability, risk). Application of a "sandbox" model allows independent certification, such that a "moderate" system can provide both "moderate" and "low" sandboxes. Policy issues related to sandbox technology must be included in any investigation to address the potential DOE policy change from certification of systems to the sandbox level. Within a "low" security sandbox, the level of rigor of the controls would be relaxed as compared to higher-security sandboxes. Basically, the sandbox model moves the perimeter from the outside (machine level) to the inside (sandbox level). The model makes the security problem scalable.

Dedicated network channels. In the exascale environment, dataflow will grow significantly, and effective use and analysis of the cyberinfrastructure are crucial. Cyber security analysis requirements do not necessarily scale up from the terascale environment to the exascale environment. Protection of interactive/control sessions at exascale is similar to what it was at terascale. Yet, the sheer increase in the bulk data transfer for an exascale environment presents significant challenges for cyber security analysis. Currently few protocols exist for bulk data transfers. While some do use specific ports, these are not widely deployed. Tools for data transfer must be developed that use dedicated channels to separate data from control communication and facilitate the application of graded levels of control for exascale. Control sessions represent the primary threat. Segregation of the data flow is one strategy to allow application of appropriate controls commensurate to security risks. System-to-system data channels may require less stringent security, allowing cyber security analysts to focus on the control channels where risks are greatest. Conversely, point-to-point networking may allow bulk data transfers from trusted system to trusted system, regardless of the size of the transaction. The ease of information exchange is important to the open science community. Investment in network monitoring for performance and anomaly detection also has the potential for significant reward with regard to identification of potential attacks, by facilitating cyber security analysis in order to better characterize behaviors to detect anomalies for the exascale environment.

3. State of the Art

In this section, we review the state of the art in each of the five areas discussed above—workflow management, collaboration frameworks and techniques, data management and movement, cyberinfrastructure management tools, and cyber security — to provide an understanding of where the cyberinfrastructure is today and what is needed to meet the challenges raised by exascale applications.

Workflow management. Workflow systems today are evolving at a promising rate but show no signs yet of coalescing around a compact set of solutions. Progress is being made in industry on workflow models and mechanisms for service-oriented architectures (such as BPEL) and architectures for specific environments such as the Microsoft Windows workflow framework. In the realm of cyberinfrastructure, good progress is being made in terascale distributed environments such as TeraGrid [Catlett et al. 2007] and Open Science Grid. Similar progress is evident in application-domain (job-oriented) systems that are resource cognizant and are starting to make transparent to scientists the steps involved in transporting, cataloging, and replicating data between stages of a workflow spread across distributed parallel computing sites. Progress in these systems is also being made in the integration of support for data provenance tracking into the workflow system itself, so that scientists who use such systems to process vast datasets can gain the added benefit of transparently recording details of the data derivation or analysis process for later auditing and discovery.

The "sandbox" model — in which each virtual organization is protected based on its specific constraints — makes the security manageable at the exascale.

Figure 9.2 The Access Grid enables collaborations between multiple groups of geographically distributed researchers. In addition to multiple audio and video streams from each location, participants can share presentations, data, visualizations, and other applications through persistent virtual venues.

Collaboration frameworks and techniques. Collaboration is already a fundamental component of science, with teams often spread across the country or extending around the world. These collaborations require both synchronous and asynchronous tools and communication. Asynchronous collaboration infrastructure today includes the use of wikis, blogs, and other emerging social networking tools, whereas synchronous collaboration infrastructure includes context- and location-aware, persistent visualization and collaboration environments (see, e.g., Stevens 2003]. These environments can seamlessly display a multitude of information from both local and remote sources, but often at the cost of lowering the experience to the lowest common denominator. This means that the collaboration capabilities of today do not create a strong sense of presence, with the remote participants not feeling they are full participants in the experience. In addition, tool developers have no straightforward method for constructing tools and applications that can support collaborative work.

Data management and movement. As noted earlier, many researchers who are creating terascale and petascale datasets find that they spend a significant portion of their time managing data rather than scientific investigation. Exascale science will generate data at ever-increasing rates. Data management and movement tools will become critical for data virtualization, management of properties and attributes independent of the underlying storage system, access control, provenance metadata, data integrity, and security.

Environments that create a strong sense of presence with remote participants will be critical for effective collaboration at the exascale.

Cyberinfrastructure management tools. Currently, cyberinfrastructure management tools for the largest HPC systems are mostly collections of scripts and tools not well suited to managing these systems. Nagios, a commonly used monitoring tool, unfortunately does not scale well. Most sites that use Nagios end up running multiple instances of the tool, tenuously fitted together with local scripts to provide a monitoring system that provides minimal capability. Many additional scripts and tools are needed to fill in the missing elements, resulting in ad hoc local management environments. To some extent, many sites cover the basics and ignore the rest, with performance frequently being neglected. Yet much could be done to provide robust systems functioning at peak performance—if the tools were available. The IBM Blue Gene/L RAS management system is currently one of the best tools for gathering system and job events. It has many shortcomings, some of which will be addressed with the next-generation Blue Gene system. Unfortunately, some of the largest issues—for example, the inability to correlate events between the compute components (i.e., Blue Gene/L hardware) and no–compute components (i.e., external data storage systems and networks), and mechanisms for dealing with the large volumes of events—remain unsolved. Work is under way at the Center for the Improvement of Fault Tolerance in Systems (CIFTS) to build a fault-tolerant backplane to alleviate the difficulties of fault prediction, notification, management, and recovery. This project, funded by DOE, has the potential to move the field closer to what is needed for petascale architectures, but much will remain to be done to provide a production solution at the exascale. Neither the Blue Gene RAS system nor the CIFTS work addresses performance tuning, reporting usage, or configuration management. Tools such as bcfg2 [Desai et al. 2003] or cfengine provide reasonable solutions for systems consisting of as many as 10,000 nodes, and these tools will remain adequate for the configuration management of support systems such as file servers. They do not, however, address the configuration of machines with 100,000 nodes or other, nontraditional environments such as virtual machines. The IBM Blue Gene and other systems have, for the most part, avoided the configuration management problem by running a small, lightweight operating system rather than a full system (such as Linux), loading it over an out-of-band network as needed. This situation has prevented many applications from utilizing these systems to the fullest capability. As we move to providing a full operating system, the problem of configuration management for the petascale and beyond will become a serious problem.

Cyber security. Currently no single technology provides complete cyber security. Most organizations protect their information technology resources through a defense-in-depth approach that covers network, host, and application security technologies; provides cyber security awareness, skills development, and training to staff and users; and details cyber security policy and standards. Tools employed by many organizations to implement a defense-in-depth approach include commercial firewall products and virtual private networks, open source network intrusion detection tools such as Snort or Bro, syslog and tripwire host intrusion detection, Nessus vulnerability scanning, Websense web filtering, cfengine or SMS configuration management, and commercial anti-virus protection. Security tools often lag significantly in their availability on early high-performance network and computational systems. Sites generally have to use open source tools, such as Bro, to be able to perform security functions at top performance speeds. The network and host monitoring mechanisms, such as IDSs, IPSs, and firewalls, do not adequately scale to support the dynamic and high-speed networking environments that are delivered to early adopters of high-performance computing sites.

4. Accelerating Development

Few resources are currently directed to or planned for cyberinfrastructure for exascale computers. Clearly, we need to establish a plan designed to accelerate cyberinfrastructure R&D if we are to meet the needs of exascale science.

Workflow management. To maximize the productivity of scientists and extract the maximum amount of knowledge and discovery from exascale science, a focused effort on workflow management and the underlying tools and libraries is required; such an effort should include the following:

- Development of transparent and highly optimized data transport between exascale workflow stages
- Automation of the complex policy-driven and congestion-sensitive scheduling decisions that need to be made to keep an exascale complex operating at an acceptable level of utilization
- Extensions to workflow models that can take advantage of dynamic exploratory models enabled by vast computing resources (e.g., exploring parallel paths and branching a new, large-scale workflow from each viable result)
- Development of common workflow languages and pluggable implementations that allow scientists to specify their processes in a manner independent of the diverse architectures that may evolve, even within a single exascale complex

Collaboration frameworks and techniques. Highly productive exascale science will demand seamless access to scientists, data, and computational resources. The goal is to enable scientists to ascertain at a glance the status of the elements of collaborations that are relevant to their work at any moment and to share their work products with their collaborators synchronously and asynchronously. Achieving this goal will involve the development of a library of collaborative software components, including the following:

- Components to enable remote participants to hear and see each other comfortably and interact naturally. These components should be as easy and reliable to use as the telephone and, as much as possible,

Current cyberinfrastructure tools can handle systems of thousands of nodes, but new solutions are needed for machines with 100,000 nodes.

re-create the sensation of being in the same room with the remote participants. This will require advances in the human-computer interfaces to prevent these interactions from disrupting the real work of the scientists.

- Event distribution services that enable applications to share state interactively, so that all collaborators see the same representation of the application as one participant interacts with it. For example, in discussing and reviewing the DNA of an organism, the participants' view of the DNA on their screen would follow the interactions of the lead speaker with the application.

- Development of network storage for application states to support snapshots of collaborative applications in time, allowing collaborators to return to selected checkpoints in their interactions with applications and data.

- Application sharing by leveraging native platform event models. This will allow scientists to utilize software specific to their domain (e.g., molecule viewers, materials databases) without modification to the software.

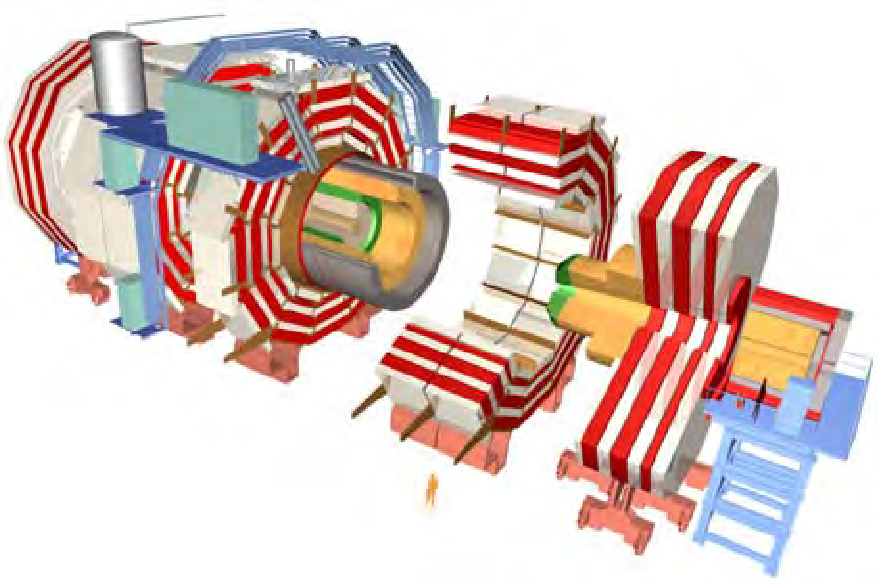


Figure 9.3 The Compact Muon Solenoid (CMS) experiment is a particle physics detector being built on the Large Hadron Collider (LHC) at CERN in Switzerland. Significant cyberinfrastructure will be required to support the thousands of physicists from around the world who will be analyzing the massive amounts of data produced by the CMS.

- Integration of exascale cyberinfrastructure developments with collaborative environments. This will allow remote collaborators to jointly access compute states, from scientists iterating on submission of science compute jobs to administrators assessing the health of the compute infrastructure.

- Integration of exascale data management and movement tools with collaborative environments. Such integration will allow remote collaborators to easily share, store, and access references to their datasets independent of physical location. Support for the relevant data transfer clients will enable users to interact with their datasets direct from their desktops with single-click and drag-and-drop access.

- Interfaces to the most prominent programming languages in use in exascale science applications. Such interfaces will facilitate language adoption by collaborations.

The availability of open tools for use as underlying infrastructure is key to the success of efforts to innovate at the application level, allowing the infrastructure to be modified by the domain scientists themselves according to their own needs and without regard for budgetary or legal constraints. Currently, the components described are largely not available as open source software. Cross-platform support is required to accommodate variation in the personal working environment of individual scientists.

New strategies are needed for cooperative work in the security-constrained exascale science network environment. Firewalls have long posed a difficult problem for science applications, and yet no suitable solution has been devised without significantly compromising security. Exascale science must employ strategies to securely enable application and data sharing and point-to-point interactions in constrained networks, while maintaining security standards compatible with the network policies of typical participating institutions.

Data management and movement. Deployment of exascale applications needs to be accelerated by support for research into data management and movement requirements at the exascale and for development of the tools or features that address these requirements, both in leading Data Grid management systems and as standalone tools.

Cyberinfrastructure management tools. In order to ensure that computational resources reach their full potential for exascale science, a substantial investment must be made in R&D on cyberinfrastructure management technology, including the following:

- *Tools and methods for scalable configuration management.* Development of a tiered configuration management strategy is essential.

- *Methodologies for monitoring tens of thousands of nodes.* Tools are needed to effectively monitor tens of thousands of nodes to ensure machine health and to indicate when a large portion of the machine fails. Monitoring should include aspects of the machine that are of interest to users, such as node availability and historical performance; aspects of interest to administrators, such as node failures; and aspects of interest to cyber security analysts, such as possible machine misuse or attack. Effectively managing data from hundreds of thousands of machine probes and terabyte-sized log files is critical.

- *Investigation and determination of proper methods for dealing with machine and component failures.* With tens of thousands of nodes, dealing with hardware failures will be a daily struggle. Strong relationships with hardware vendors and a clearly documented procedure will be essential.

Cyber security. Continued support is essential for the development of open source network intrusion detection tools, such as Bro. Equally important, the development of new approaches to the expected complex and distributed exascale cyberinfrastructure, could accelerate the capabilities to secure the exascale cyberinfrastructure soon after it becomes available. Tools to help detect unauthorized alteration of data (malicious or accidental) in data archival systems, Data Grid management systems, and file systems will be increasingly necessary as exascale datasets start to stress component (network, CPU, storage) data error rates. Underlying all of the cyber security capabilities is the need for more flexible, scalable, and verifiable authorization and authentication frameworks and tools.

Exascale science applications will need new strategies that enable data sharing while ensuring that individual institutional security policies are maintained.

5. Expected Outcomes

Sufficient and timely investment to support multidisciplinary teams in a focused program tailored to encourage and support the development of cyberinfrastructure in the five areas outlined here will enable increased scientific output, productive collaborations between team members, simplified management of data, increased reliability of resources, and overall security of cyberinfrastructure for all exascale science areas.

6. Required Investment

The effort will require multiple teams to advance the state of the art in each of the key areas identified here. Sizable investment also will be needed for designing, developing, testing, and deploying the cyberinfrastructure. In addition, some effort will have to be dedicated to maintaining the collaborations that will support crosscutting aspects of the development with the application areas and technology development teams.

7. Major Risks

Several risks and potential downside consequences are linked to exascale cyberinfrastructure R&D.

Workflow Management

- If we fail to invent and develop robust user-friendly workflow management environments, exascale scientists will not be able to process the vast amounts of data to get the most out their efforts and the investment in these large-scale systems.

- If we fail to sufficiently automate the resource management layers of workflow systems, vast hardware resources of exascale complexes will be underutilized and result in disappointing levels of speedup and productivity.

Collaboration Frameworks and Techniques

- If we do not invest in R&D tools and infrastructure to support the remote operation of global instruments, we will lose the active involvement of a large number of our scientists in the science done on these resources.

- If collaborative infrastructure does not support teams in a natural manner, productivity of the community will suffer, and efforts will be fragmented and duplicative because of inability to share results and collaborate effectively in large teams.

Data Management and Movement

- If we fail to address issues related to both short- and long-term management of data, we risk the loss of productivity of scientists as they deal with management and organization of data. Further, utilization of expensive resources will be reduced because of data movement into and out of these machines without high-performance capabilities and remote access to storage devices.

- If we do not develop strategies and infrastructure to manage the vast amounts of data that exascale science will produce, we risk losing intellectual property embedded in the data.

Cyberinfrastructure Management Tools

- If we do not invest in R&D on cyberinfrastructure management tools and techniques, we risk the investment in resources due to downtime and system errors as these complex resources are debugged.

- If we fail to address the usability of required security mechanisms, users will continue to be hindered, at a cost of productivity or a reduction in security as users find workarounds.

- If we inadequately connect the efforts in cyberinfrastructure to development in technology areas, we will see a duplication of effort to build the underlying infrastructure for each area.

Without the automation of the resource management layers of workflow systems, exascale resources will be underused, and both performance and productivity levels will be disappointing.

Cyber Security

- If mechanisms to validate the integrity of datasets and data collections are not developed, the accuracy and integrity of the research may be at risk.

- If significant log analysis tools to detect and investigate anomalies are not available, the security of the exascale-class and support systems will be at risk.

- If significant intrusion detection capabilities at line speed are not available, the security of the exascale class and support systems will be at risk.

- The current DOE policy is to certify systems to the highest level of risk for the system and implement the corresponding security control level, forcing all who desire to use the resource to conform regardless of the risk level of their individual work. If the proposed sandbox technology is not investigated and developed, open research may be at risk of moving to less capable systems because of the perception that security policies get in the way.

R&D in cyberinfrastructure is essential to the success of the exascale effort. It is a critical component, providing the foundation on which tools and applications for doing science will be built, the infrastructure by which scientists will collaborate, and the tools that ensure applications have resources on which to run, all in an environment that is secure.

References

Bill Allcock, Joe Bester, John Bresnahan, Ann L. Chervenak, Carl Kesselman, Sam Meder, Veronika Nefedova, Darcy Quesnel, Steven Tuecke and Ian Foster (2001), Secure, efficient data transport and replica management for high-performance data-intensive computing, p. 13 in *Proc. 18th IEEE Symposium on Mass Storage Systems*.

Charlie Catlett et al. (2007), TeraGrid: Analysis of organization, system architecture, and middleware enabling new types of applications, in *High Performance Computing and Grids in Action*, IOS Press, Advances in Parallel Computing series, Amsterdam.

Narayan Desai, Andrew Lusk, Rick Bradshaw and Remy Evard (2003), BCFG: A configuration management tool for heterogeneous environments, in *Proc. IEEE International Conference on Cluster Computing*, pp. 500, 2003.

Ian Foster, Jens Voeckler, Michael Wilde and Yong Zhao (2003), The virtual Data Grid: A new model and architecture for data-intensive collaboration, in *Proc. Conference on Innovative Data Systems Research*.

Veronika Nefedova, Robert Jacob, Ian Foster, Zhengyu Liu, Yun Liu, Ewa Deelman, Gaurang Mehta, Mei-Hui Su and Karan Vahi (2006), Automating climate science: Large ensemble simulations on the TeraGrid with the GriPhyN virtual data system, p. 32 in *Proc. 2nd IEEE International Conference on e-Science and Grid Computing*.

Rick Stevens (2003), Access Grid: Enabling group-oriented collaboration on the Grid, in The Grid: Blueprint for a New Computing Infrastructure, Ian Foster and Carl Kesselman, eds., Morgan Kaufmann.

Rick Stevens, Michael E. Papka and Terry Disz (2003), Prototyping the workspace of the future, *IEEE Internet Computing* 7(4): 51–58.

Gregor von Laszewski, Michael Hategan and Deepti Kodeboyina (2007), Java CoG Kit workflow, pp. 340-356 in *Workflows for Science*, I. Taylor, E. Deelman, D. B. Gannon, and M. Shields, eds., Springer.

Von Welch, Jim Barlow, James Basney, Doru Marcusiu and Nancy Wilkins-Diehr (2006), A AAAA model to support science gateways with community accounts, *Concurrency and Computation: Practice and Experience* 19(6):893-904.

Yong Zhao, Michael Wilde and Ian Foster (2007), Virtual Data Language: A Typed Workflow Notation for Diversely Structured Scientific Data, pp. 258-278 in *Workflows for eScience,* Springer.

Security policies on exascale systems must not get in the way of science.

Appendix A

Initiative Summary

U.S. Department of Energy
Office of Science

Simulation and Modeling at the Exascale for Energy, Ecological Sustainability and Global Security An Initiative

The objective of this ten-year vision, which is in line with the Department of Energy's Strategic Goals for Scientific Discovery and Innovation, is to focus the computational science experiences gained over the past ten years on the opportunities introduced with exascale computing to revolutionize our approaches to energy, environmental sustainability and security global challenges.

Executive Summary

The past two decades of national investments in computer science and high-performance computing have placed the DOE at the forefront of many areas of science and engineering. This initiative capitalizes on the significant gains in computational science and boldly positions the DOE to attack global challenges through modeling and simulation. The planned petascale computer systems and the potential for exascale systems shortly provide an unprecedented opportunity for science; one that will make it possible to use computation not only as an critical tool along with theory and experiment in understanding the behavior of the fundamental components of nature but also for fundamental discovery and exploration of the behavior of complex systems with billions of components including those involving humans.

Through modeling and simulation, the DOE is well-positioned to build on its demonstrated and widely-recognized leadership in understanding the fundamental components of nature to be a world-leader in understanding how to assemble these components to address the scientific, technical and societal issues associated with energy, ecology and security on a global scale.

In order to realize this leadership the DOE recognizes that the time-honored, or subsystems, approach in which the forces and the physical environments of a phenomenon are analyzed, is approaching a state of diminishing returns. The approach for the future must be systems based and simulation programs are developed in the context of encoding all known relevant physical laws with engineering practices, production, utilization, distribution and environmental factors.

This new approach will

- Integrate, not reduce. The full suite of physical, chemical, biological, chemical and engineering processes in the context of existing infrastructures and human behavior will be dynamically and realistically linked, rather than focusing on more detailed understanding of smaller and smaller components.

- Leverage the interdisciplinary approach to computational sciences. Current algorithms, approaches and levels of understanding may not be adequate. A key challenge in development of these models will be the creation of a framework and semantics for model interaction that allow the interconnection of discipline models with observational data. At the outset, specialized scientific groups will team with engineers, business experts, ecologists and human behavior specialists comprehensive models, that incorporate all known phenomena and have the capability to simulate systems characteristics under the full range of uncertainties.

- Capitalize on developments in data management and validation of ultra-large datasets. It will develop new approaches to data management, visualization and analysis that can treat the scale and complexity of the data and provide the insight needed for validation of the computations.

This new approach will enable DOE to exploit recent developments in commercially available computer architectures, driven by the implementation of first generation multi-core processors and the introduction of petascale computers within 18 months, and prepare it to take advantage of exascale computers in the next decade. This approach will also guarantee DOE's leadership in applying these computers to critical problems confronting the nation.

The initiative has four programmatic themes:

1. Engage top scientists and engineers, computer scientists and applied mathematicians in the country to develop the science of complexity as well as new science driven computer architectures and algorithms that will be energy efficient, extremely scalable, and tied to the needs of scientific computing at all scales. Correspondingly, recruit and develop the next generation of computational and mathematical scientists.

2. Invest in pioneering large-scale science, modeling and simulation that contribute to advancing energy, ecology and global security.

3. Develop scalable analysis algorithms, data systems and storage architectures needed to accelerate discovery from large-scale experiments and enable verification and validation of the results of the pioneering applications. Additionally, develop visualization and data management systems to manage the output of large-scale computational science runs and in new ways to integrate data analysis with modeling and simulation.

4. Accelerate the build-out and future development of the DOE open computing facilities to realize the large-scale systems-level science required to advance the energy, ecology and global security program. Develop an integrated network computing environment that couples these facilities to each other, to other large-scale national user facilities and to the emerging international network of high-performance computing systems and facilities.

The success of this fourth effort is built on the first three themes because exascale systems are, by themselves, among the most complex systems ever engineered.

U.S. Department of Energy
Office of Science

This initiative will enable DOE to address critical challenges in areas such as:

Energy- Ensuring global sustainability requires reliable and affordable pathways to low-carbon energy production, e.g. bio-fuels, fusion and fission, and distribution on a massive scale. The existing mix of energy supplies places global security at great risk. Acceptable solutions require rapid and unprecedented scientific and technologic advances. Unfortunately, existing analytical, predictive, control, and design capabilities will not scale. An objective of this initiative is to provide new models and computational tools with the functionality needed to discover and develop complex processes inherent in a new energy economy.

Ecological Sustainability- The effort toward sustainability involves characterizing the conditions for balance in the climate system. In particular, sustainable futures involve understanding and managing the balance of chemicals in the atmosphere and ocean. The ability to fit energy production and industrial emissions within balanced global climate and chemical cycles is the major scientific and technical challenge for this century.

Security- The internet, as well as the instrumentation and control systems for the energy infrastructure, is central to the well-being of our society. There are several potential opportunities relating to accurately modeling these complex systems: understand operational data, identify anomalous behavior to isolate the disturbance and automatically repair any damage.

For further information on this subject contact:

Dr. Michael Strayer, Associate Director
Office of Advance Scientific Computing Research
Michael.Strayer@science.doe.gov

Appendix B

Agendas

Simulation and Modeling at the Exascale for
Energy, Ecological Sustainability and Global Security (E3SGS)
Lawrence Berkeley National Laboratory
Town Hall Meeting

April 17 – 18, 2007

Tuesday, April 17, 2007

8:00 – 8:15 am	Welcome and Introduction...*Horst Simon, Associate Laboratory Director, Computing Sciences, LBNL*
8:15 – 8:45 am	Opening Remarks – Kick-off announcement on town hall meeting series.....*Michael Strayer, Associate Director, ASCR*
8:45 – 9:45 am	“The Computational Frontiers of Earth System Modeling”..................................*Bill Collins, LBNL and NCAR*
9:45 – 10:15 am	Morning Break
10:15 – 10:45 am	Panel Session..................................Panelists: *Horst Simon, Rick Stevens, Thomas Zacharia*
11:00 – 12 noon	Start breakout group discussions *B1 in 54-130 (Perseverance Hall)* *B2 in 66-Auditorium* *B3 in 66-316* *B9 in 62-203*
12:00 – 1:00 pm	Working lunch – pick up lunch and continue discussion
1:00 – 3:30 pm	Continue breakout group discussions
3:30 – 5:30 pm	Report back from breakout groups
5:30 pm	Adjourn

Wednesday, April 18, 2007

8:00 – 8:15 am	Welcome Back and Agenda for Day 2..*Horst Simon*
8:15 – 12:00 pm	Breakout group discussions *B4 in 62-255* *B5 in 66-Auditorium* *B7 in 62-203* *B8 in 66-316*
12:00 – 1:00 pm	“Solutions to the Energy Crisis” *(working lunch scheduled during this time).......Steve Chu, Laboratory Director, LBNL*
1:00 – 2:30 pm	Report back from breakout groups
2:30 pm	Adjourn

Simulation and Modeling at the Exascale for Energy, Ecological Sustainability and Global Security (E3SGS) Oak Ridge National Laboratory Town Hall Meeting

May 17-18, 2007

Wednesday, May 16, 2007

Arrival and dinner on your own

Thursday, May 17, 2007

Town Hall parking is in top two tiers of lot across Bethel Valley Road from ORNL Visitor Center

6:45 a.m. Buses pick up at hotels (Homewood and Springhill Suites in Turkey Creek area; Comfort Suites at Campbell Station Road; Comfort Inn, Double Tree, and Jameson Inn in Oak Ridge) One stop at each hotel

7:00-
7:15 a.m. Arrive Bldg. 8600 ... Check in/visitor badging
7:25 a.m. Local participants to meet buses at ORNL Visitor Center for transportation to Bldg. 8600
7:30 a.m. Morning networking session and refreshments ... Bldg. 8600, 1st floor atrium

8:00 a.m. Welcome and Introductions (Iran Thomas Auditorium) ... Thomas Zacharia
8:15 a.m. Opening remarks ... Michael Strayer, Assoc. Dir. ASCR

Climate Change Session... David Erickson, Session Chair
8:45 a.m. Climate change: policy perspectives ... Lincoln Pratson, Duke
9:30 a.m. Climate change: computational science perspectives ... David C. Bader, LLNL
10:15 a.m. Break

10:30 a.m. Laboratory Director's Welcome ... Jeff Wadsworth

Energy Session ... Doug Kothe, Session Chair
10:40 a.m. Energy: industry perspectives ... Jack Bailey, TVA
11:10 a.m. Energy: computational science perspectives ... Tom Downar, UC-Berkeley

12:00 p.m. Working lunch ... 2nd Floor Lobby, CNMS

Lunch will be picked up in the 2nd floor lobby of the CNMS –Center for Nanophase Materials Sciences-- just around the corner and a short walk away from the Iran Thomas Auditorium.

Sitting areas will be available on 1st, 2nd, 3rd floor lobbies, in adjacent conference rooms, on the patio, in room 156, the CNMS executive conference room, in small common areas, and in the SNS atrium – however, no food and beverages are allowed in the auditorium.

1:00 p.m. Applications breakouts ... Board buses at CNMS
B1. Climate ... SNS Room C156 (8600)
B6. Numerical Climate-Economic-Energy... Research Office Building L204 (5700)
B3. Biology ... SNS 354 (8600)
B2. Energy ... Iran Thomas Auditorium (8600)
B9. Astrophysics ... CNMS L183 (8610)
B10. Industrial Processes and Manufacturing ... CNMS L382 (8610)

3:00 p.m. Break
3:30 p.m. Continue applications breakout sessions

5:30 p.m. Adjourn

Board buses to return to main campus .... Flagpole lobby entrance

6:00 p.m. Dinner (Bldg. 5200, 2nd floor) .... Session Summaries

9:00 p.m. Board buses for return to hotel .... Visitor Center

* * * * * * * * * * * * * * *

Friday, May 18, 2007, sessions will be held on main campus in Bldg. 5200 (Conference Center)

7:30 a.m. Networking session and refreshments .... Bldg. 5200, 2nd floor lobby

8:00 a.m. Welcome back and agenda for day 2 (Rooms: Tennessee A, B, C) .... Thomas Zacharia
8:15 a.m. Begin infrastructure breakouts
B5. Hardware .... Room tbd
B7. Math .... Room tbd
B8. Software .... Room tbd
B4. Cyber security .... Room tbd

9:45 a.m. Break

10:15 a.m. Continue breakout sessions...

12:15p.m. Working lunch (pick up lunch and continue breakouts) .... 2nd floor lobby

1:45 p.m. Closing remarks

2:00 p.m. Adjourn

Board buses for return to hotels. .... Visitor Center

Simulation and Modeling at the Exascale for
Energy, Ecological Sustainability and Global Security (E3SGS)
Argonne National Laboratory
Town Hall Meeting

May 31 – June 1, 2007

Thursday, May 31, 2007

8:00 Welcome and Introduction..Rick Stevens,
Associate Laboratory Director, Computing and Life Sciences, ANL

8:15 Opening Remarks..Michael Strayer,
Associate Director, ASCR

8:45 "Energy, Environmental Sustainability and the Role of High-End Computing"............Robert Rosner,
Laboratory Director, ANL

10:00 Morning Break – Auditorium/Rotunda

10:30 "Potential Applications of Exascale Computing in Economics"..........Kenneth L. Judd, Paul H. Bauer
Senior Fellow, Hoover Institution

11:15 Agenda and Instructions for Town Hall ....Rick Stevens

11:30 – 3:30 Breakout group discussions

Noon – 1p.m. Working lunch scheduled noon-1p.m. – pick up lunches and continue discussions)
B1 Climate – Hosts: Kemner/Kotamarthi
B2 Energy – Hosts: Frank/Nowak
B3 Biology – Hosts: Edwards/Meyer
B6 Socioeconomic Modeling – Hosts: Foster/Jacob
B9 Astrophysics – Hosts: Fischer/Lamb
B10 Industrial Processes and Manufacturing – Hosts: Hassanein/Moré

2:00 - 2:30 Afternoon Break – Auditorium/Rotunda

3:30 Report back from breakout groups (each group 20min)

5:30 Adjourn

Friday, June 1, 2007

8:00 - 8:15 Welcome Back and Agenda for Day 2..Rick Stevens

8:15 – 12:00 Breakout group discussions
(working lunch scheduled noon-1p.m. – pick up lunches at Auditorium/Rotunda and continue discussions/report writing)
B4 Cyberinfrastructure – Hosts: Catlett/Papka
B5 Hardware – Hosts: Beckman/Gropp
B7 Math – Hosts: Hereld/Norris
B8 Software – Hosts: Lusk/Ross

10:00 - 10:30 Morning Break – Auditorium/Rotunda

12:00 Report back from breakout groups (each group 15min)

1:00 Next steps/Comments...Rick Stevens, Bill Kramer, Jeff Nichols

2:00 Adjourn

Appendix C

Attendees

Combined Attendees List

Abel, Tom	Kavli Institute for Particle Astrophysics and Cosmology
Agarwal, Deb	Lawrence Berkeley National Laboratory
Agarwal, Pratul K.	Oak Ridge National Laboratory
Ahern, Sean	UT-Battelle / Oak Ridge National Laboratory
Ahrens, James	Los Alamos National Laboratory
Alam, Sadaf R.	Oak Ridge National Laboratory/UT-Battelle
Alexiades, Vasilios	University of Tennessee Knoxville
Allan, Benjamin	Sandia National Laboratories
Altaweel, Mark	Argonne National Laboratory
Andrade, Jose	Northwestern University
Anitescu, Mihai	Argonne National Laboratory
Antypas, Katerina	Lawrence Berkeley National Laboratory / National Energy Research Scientific Computing Center
Apra, Edoardo	UT-Battelle
Aragon, Cecilia	Lawrence Berkeley National Laboratory
Archibald, Rick	Oak Ridge National Laboratory
Asanovic, Krste	Massachusetts Institute of Technology / University of California Berkeley
Ashby, Steven F.	Lawrence Livermore National Laboratory
Babnigg, Gyorgy	Argonne National Laboratory
Bader, David A.	Georgia Institute of Technology
Bader, David C.	Lawrence Livermore National Laboratory
Baertsch, Robert	University of California Santa Cruz
Bai, Zhaojun	University of California Davis
Bailey, David H.	Lawrence Berkeley National Laboratory
Bailey, Jack A.	Tennessee Valley Authority
Bair, Raymond A.	Argonne National Laboratory
Baker, Allen	University of Tennessee
Baldocchi, Dennis	University of California Berkeley
Banda, Michael	Lawrence Berkeley National Laboratory
Banks, David C.	University of Tennessee
Barker, Kevin	Los Alamos National Laboratory
Barrett, Richard	Oak Ridge National Laboratory
Barron, Tom O.	Oak Ridge National Laboratory
Bartels, Daniela	University of Chicago
Bashor, Jon	Lawrence Berkeley National Laboratory
Beavers, James	University of Tennessee
Beck, Micah	University of Tennessee
Beckman, Pete H.	Argonne National Laboratory
Bell, John	Lawrence Berkeley National Laboratory
Berger-Wolf, Tanya	University of Illinois-Chicago
Bernholc, Jerry	North Carolina State University
Bernholdt, David E.	Oak Ridge National Laboratory
Berry, Michael W.	University of Tennessee
Bethel, E. Wes	Lawrence Berkeley National Laboratory
Bland, Arthur S.	Oak Ridge National Laboratory
Blondin, John M.	North Carolina State University
Bode, Brett	Ames Laboratory
Bohn, Robert	NCO/NITRD
Borrill, Julian	Lawrence Berkeley National Laboratory / University of California Berkeley
Boudwin, Kathlyn	UT-Battelle
Bownas, Jennifer	Oak Ridge National Laboratory
Braam, Peter	Cluster File Systems, Inc.

Bramley, Randall	Indiana University
Branstetter, Marcia	UT-Battelle
Britt, Phillip F.	Oak Ridge National Laboratory
Bronevetsky, Greg	Lawrence Livermore National Laboratory
Brown, Nancy	Lawrence Berkeley National Laboratory
Brown , David L.	Lawrence Livermore National Laboratory
Brown , Maxine	University of Illinois-Chicago
Budiardja, Reuben	University of Tennessee
Buja, Lawrence E.	National Center for Atmospheric Research
Calafiura, Paolo	Lawrence Berkeley National Laboratory
Calder, Alan	State University of New York at Stony Brook
Canning, Andrew	Lawrence Berkeley National Laboratory
Cannon, William	Battelle Memorial Foundation
Canon, Richard	Oak Ridge National Laboratory
Cardall, Christian Y.	Oak Ridge National Laboratory
Carter, Jonathan	Lawrence Berkeley National Laboratory
Catlett, Charles E.	Argonne National Laboratory/University of Chicago
Chavarria, Daniel	Pacific Northwest National Laboratory
Chen, Jacqueline	Sandia National Laboratories
Cheng, Robert	Lawrence Berkeley National Laboratory
Chiang, John	University of California Berkeley
Childs, Henry	Lawrence Livermore National Laboratory
Chiu, George L.	IBM
Choudhary, Alok	Northwestern University
Christiansen, John H.	Argonne National Laboratory
Chu, Steve	Lawrence Berkeley National Laboratory
Cirillo, Richard	Argonne National Laboratory
Clarke, Leon E.	Pacific Northwest National Laboratory
Clarno, Kevin T.	Oak Ridge National Laboratory
Cobb, John W.	Oak Ridge National Laboratory
Coghlan, Susan	Argonne National Laboratory
Cohoon, Matthew	The University of Chicago
Colella, Phil	Lawrence Berkeley National Laboratory
Coleman-Smith, Christopher	Lawrence Berkeley National Laboratory
Collins, William	Lawrence Berkeley National Laboratory / University of California Berkeley
Collis, S. Scott	Sandia National Laboratories
Conway, Claudine	Dell
Cortis, Andrea	Lawrence Berkeley National Laboratory
Counce, Deborah	Oak Ridge National Laboratory
Crawford, Matt	Fermi National Laboratory
Culhane, Candace	Oak Ridge National Laboratory
Curfman McInnes, Lois	Argonne National Laboratory
Dale, Virginia H.	Oak Ridge National Laboratory
Davis, Kei	Los Alamos National Laboratory
Davis, Wayne	University of Tennessee
Davis, William	Lawrence Berkeley National Laboratory
D’Azevedo, Eduardo	Oak Ridge National Laboratory
de Almeida, Valmor	UT-Battelle
DeBenedictis, Erik	Sandia National Laboratories
Deiterding, Ralf	Oak Ridge National Laboratory
del-Castillo-Negrete, Diego	Oak Ridge National Laboratory
DePaoli, David	Oak Ridge National Laboratory
D’haeseleer, Patrik	Lawrence Livermore National Laboratory

Diachin, Lori A.	Lawrence Livermore National Laboratory
DiBennardi, Kathy	Argonne National Laboratory
Disz, Terrence	Argonne National Laboratory
Dobson, Patrick	U.S. Department of Energy Office of Basic Energy Sciences
Doctor, Richard	Argonne National Laboratory
Dongarra, Jack J.	University of Tennessee / Oak Ridge National Laboratory
Dosanjh, Sudip	Sandia National Laboratories
Downar, Thomas	Purdue University
Drake, John B.	Oak Ridge National Laboratory
Drugan, Cheryl	Argonne National Laboratory
Dubchak, Inna	Lawrence Berkeley National Laboratory
Duffy, Philip	Lawrence Livermore National Laboratory
Dupont, Todd	University of Chicago
Edwards, Robert	Argonne National Laboratory
Eigenmann, Rudolph	Purdue University
Ellis, David	LSI Corporation
Elnozahy, Elmootazbellah	IBM
Elwasif, Wael	Oak Ridge National Laboratory
Elwood, Jerry W.	U.S. Department of Energy
Endeve, Eirik	University of Tennessee / Oak Ridge National Laboratory
Erickson, David	Oak Ridge National Laboratory
Evans, Thomas	Oak Ridge National Laboratory
Fahey, Mark	Oak Ridge National Laboratory
Fann, George	Oak Ridge National Laboratory
Feng, Wu-chun	Virginia Tech
Feng, Zhili	Oak Ridge National Laboratory
Fernandez, Steven	Oak Ridge National Laboratory
Finsterle, Stefan	Lawrence Berkeley National Laboratory
Fischer, Paul	Argonne National Laboratory
Foley, Samantha	Oak Ridge National Laboratory
Folta, Peggy	Lawrence Livermore National Laboratory
Fossum, Barbara	Purdue University
Foster, Ian T.	Argonne National Laboratory / University of Chicago
Fowler, Robert	RENCI / University of North Carolina
Frank, Ed	Argonne National Laboratory
Fraser, Dan	Argonne National Laboratory
Friedman, Alex	Lawrence Livermore National Laboratory / Lawrence Berkeley National Laboratory
Fu, Joshua S.	University of Tennessee
Fu, Joshua Sing-Yih	University of Tennessee
Fujimoto, Richard M.	Georgia Tech
Fung, Inez	Lawrence Berkeley National Laboratory / University of California Berkeley
Galli, Giulia	University of California Davis
Ganguly, Auroop R.	Oak Ridge National Laboratory
Geffen, Charlette A.	Pacific Northwest National Laboratory
Gehin, Jess	UT-Battelle / Oak Ridge National Laboratory
Geist, George A.	Oak Ridge National Laboratory
Gerber, Richard	Lawrence Berkeley National Laboratory
Germain, Robert	IBM
Ghan, Steven	Pacific Northwest National Laboratory
Glass, Elizabeth	Argonne National Laboratory
Glendenin, Chad	University of Chicago
Goldberg, Stephen	Argonne National Laboratory
Golliver, Roger	Intel

Gorin, Audrey A.	Oak Ridge National Laboratory
Gracio, Debbie	Pacific Northwest National Laboratory
Graf, Peter	National Renewable Energy Laboratory
Graham, Richard	Oak Ridge National Laboratory
Graham, Robin	UT-Battelle
Gray, William	UT-Battelle
Grcar, Joseph	Lawrence Berkeley National Laboratory
Greene, Sherrell R.	Oak Ridge National Laboratory
Gropp, William	Argonne National Laboratory
Grossan, Bruce	Eureka Scientific / INPA
Grossman, Robert	University of Illinois-Chicago
Guidry, Michael	University of Tennessee / Oak Ridge National Laboratory
Gustafson, William	Pacific Northwest National Laboratory
Hanson, Donald	Argonne National Laboratory
Hanson, Paul J.	Oak Ridge National Laboratory
Hartman-Baker, Rebecca	Oak Ridge National Laboratory
Hauser, Loren	UT-Battelle
Hazlewood, Victor G.	UT-Battelle / Oak Ridge National Laboratory
Head-Gordon, Martin	Lawrence Berkeley National Laboratory
Head-Gordon, Teresa	Lawrence Berkeley National Laboratory
Helland, Barbara	U.S. Department of Energy Office of Science
Henry, Christopher	Northwestern University
Hereld, Mark	Argonne National Laboratory
Heroux, Michael	Sandia National Laboratories
Hetrick, David M.	Oak Ridge National Laboratory / UT-Battelle
Hix, William	Oak Ridge National Laboratory
Hoffman, Forrest M.	Oak Ridge National Laboratory
Hoisie, Adolfy	Los Alamos National Laboratory
Homen-de-Mello, Tito	Northwestern University
Hovland, Paul	Argonne National Laboratory
Howell, Louis	Lawrence Livermore National Laboratory
Huang, Jian	University of Tennessee
Iancu, Costin	Lawrence Berkeley National Laboratory
Iordache, Maria	IBM
Jackson, Keith	Lawrence Berkeley National Laboratory
Jacob, Robert	Argonne National Laboratory
Jacobs, Gary	Oak Ridge National Laboratory
Jain, Prashant	University of Illinois Urbana-Champaign
Janssen, Curtis	Sandia National Laboratories
Jemian, Pete	Argonne National Laboratory
Johnson, David	UT-Battelle
Johnson, Fred	U.S. Department of Energy Office of Science
Jones, Phillip	Los Alamos National Laboratory
Jones, Wesley	National Renewable Energy Laboratory
Joo, Kyungseon	University of Connecticut
Joseph, Anthony	University of California Berkeley
Jouline, Igor B.	Oak Ridge National Laboratory
Joy, Ken	University of California Davis
Jubin, Robert	Oak Ridge National Laboratory
Judd, Kenneth	Hoover Institution
Kalyanaraman, Anantharaman	Washington State University
Kamath, Chandrika	Lawrence Livermore National Laboratory
Karonis, Nick	Northern Illinois University

Karpov, Eduard	University of Tennessee
Kasdorf, James	Pittsburgh Supercomputing Center
Kaushik, Dinesh	Argonne National Laboratory
Keffer, David	University of Tennessee
Kendall, Ricky	Oak Ridge National Laboratory
Kerstein, Alan	Sandia National Laboratories
Kettimuthu, Raj	Argonne National Laboratory
Khaleel, Moe A.	Pacific Northwest National Laboratory
Khamayseh, Ahmed K.	Oak Ridge National Laboratory
Khokhlov, Alexei M.	University of Chicago
Khomami, Bamin	University of Tennessee
Kibblewhite, Ed	University of Chicago
Kiefer, David	Cray, Inc.
Klasky, Scott A.	Oak Ridge National Laboratory
Knotek, Michael L.	Knotek Scientific Consulting
Koester, David P.	MITRE Corporation
Kohl, James	Oak Ridge National Laboratory
Konerding, David	Lawrence Berkeley National Laboratory
Kotamarthi, Veerabhadra	Argonne National Laboratory
Kothe, Douglas B.	Oak Ridge National Laboratory
Kramer, William	Lawrence Berkeley National Laboratory / National Energy Research Scientific Computing Center
Krueger, Paul	Cray, Inc.
Kuehn, Jeffrey	Oak Ridge National Laboratory
Kumaran, Kalyan	Argonne National Laboratory
Kumfert, Gary	Lawrence Livermore National Laboratory
Kyrpides, Nikos	Lawrence Berkeley National Laboratory
Laguna, Pablo	Penn State University
Lamb, Donald	University of Chicago
Land, Miriam	Oak Ridge National Laboratory
Langston, Michael A.	University of Tennessee
Larzelere, Alexander	Los Alamos National Laboratory
Lattimer, James	State University of New York at Stony Brook
Lazaro, Michael	Argonne National Laboratory
Lee, Jason	Lawrence Berkeley National Laboratory
Lee, Lie-Quan	Stanford Linear Accelerator Center
Leininger, Matt	Lawrence Livermore National Laboratory
Leszczynski, Jerzy R.	Jackson State University
Levesque, John	Cray, Inc.
Levine, Michael	Carnegie Mellon University
Li, Fangxing	University of Tennessee
Li, Sherry	Lawrence Berkeley National Laboratory
Lightstone, Felice	Lawrence Livermore National Laboratory
Lindsay, Robert	U.S. Department of Energy Office of Science
Liu, Fang	Indiana University at Bloomington
Liu, Xiaohong	Pacific Northwest National Laboratory
Livny, Miron	University of Wisconsin-Madison
Locascio, Phil	Oak Ridge National Laboratory
Loebl, Andrew	UT-Battelle
Loft, Rich	National Center for Atmospheric Research
Lu, Guoping	Lawrence Berkeley National Laboratory
Lucas, Robert	University of Southern California
Ludaescher, Bertram	University of California Davis
Lusk, Ewing L.	Argonne National Laboratory

Lynch, Vicki E.	Oak Ridge National Laboratory
Ma, Xiaosong	North Carolina State University / Oak Ridge National Laboratory
Macal, Charles	Argonne National Laboratory
Mansfield, Betty K.	Oak Ridge National Laboratory
Marchesini, Stefano	Lawrence Berkeley National Laboratory
Markidis, Stefano	University of Illinois Urbana-Champaign
Markowitz, Victor	Lawrence Berkeley National Laboratory
Marquez, Andres	Pacific Northwest National Laboratory
Marronetti, Pedro	Florida Atlantic University
Martin, Dan	Lawrence Berkeley National Laboratory
Martinez, Manuel	University of Louisville
Matarazzo, Celeste	Lawrence Livermore National Laboratory
May, Elebeoba	Sandia National Laboratories
McCoy, Debbie D.	Oak Ridge National Laboratory
McFarlane, Joanna	UT-Battelle
McHardy, Alice	IBM Research
McMahon, James	Lawrence Berkeley National Laboratory
McParland, Charles	Lawrence Berkeley National Laboratory
Menon, Surabi	Lawrence Berkeley National Laboratory
Messer, Bronson	Oak Ridge National Laboratory
Meyer, Folker	Argonne National Laboratory
Meza, Juan	Lawrence Berkeley National Laboratory
Mezzacappa, Anthony	Oak Ridge National Laboratory
Michaels, George	Pacific Northwest National Laboratory
Mickelson, Sheri	University of Chicago / Computation Institute
Middleton, Don E.	National Center for Atmospheric Research
Miller, Norman	Lawrence Berkeley National Laboratory
Mills, Richard	Oak Ridge National Laboratory
Monroe, Laura	Los Alamos National Laboratory
Moore, Shirley	University of Tennessee
Moore, Terry	University of Tennessee
Moré, Jorge J.	Argonne National Laboratory
Mundy, Chris	Pacific Northwest National Laboratory / Battelle
Munson, Todd	Argonne National Laboratory
Murphy, Richard C.	Sandia National Laboratory
Myra, Eric	State University of New York at Stony Brook
Myrick, Virginia L.	Oak Ridge National Laboratory
Najm, Habib	Sandia National Laboratories
Neaton, Jeffrey	Lawrence Berkeley National Laboratory
Neitzel, Bryon	Cluster File Systems
Ng, Cho	Stanford Linear Accelerator Center
Ng, Esmond	Lawrence Berkeley National Laboratory
Nichols, Jeffrey A.	Oak Ridge National Laboratory
Nikravesh, Masoud	University of California Berkeley / Lawrence Berkeley National Laboratory
Norman, Michael	University of California San Diego
Norris, Boyana	Argonne National Laboratory
North, Michael	Argonne National Laboratory
Nowak, David	Argonne National Laboratory
Nugent, Peter	Lawrence Berkeley National Laboratory
Nukala, Phani K.	Oak Ridge National Laboratory
Nutaro, James	UT-Battelle
Oefelein, Joseph	Sandia National Laboratories
Oehmen, Christopher	Pacific Northwest National Laboratory

Okojie, Felix A.	Jackson State University
Oliker, Leonid	Lawrence Berkeley National Laboratory
Olson, Douglas	Lawrence Berkeley National Laboratory
Olson, Robert	Argonne National Laboratory
O’Maonaigh, Heather C.	Oak Ridge Associated Universities
Ostrouchov, George	Oak Ridge National Laboratory
Otoo, Ekow	Lawrence Berkeley National Laboratory
Overbeek, Ross	Fellowship for Interpretation of Genomes
Pagerit, Sylvain	Argonne National Laboratory
Pan, Jerry	Oak Ridge National Laboratory
Pannala, Sreekanth	UT-Battelle / Oak Ridge National Laboratory
Papka, Michael E.	Argonne National Laboratory / University of Chicago
Pascucci, Valerio	Lawrence Livermore National Laboratory
Peery, James	Sandia National Laboratories
Pennington, Rob	National Center for Supercomputing Applications / University of Illinois
Penumadu, Dayakar	University of Tennessee
Perumalla, Kalyan	Oak Ridge National Laboratory
Peters, Mark	Argonne National Laboratory
Peterson, Cynthia B.	University of Tennessee
Peterson, Gregory D.	University of Tennessee
Petravick, Don	Fermi National Laboratory
Pieper, Gail	Argonne National Laboratory
Pinar, Ali	Lawrence Berkeley National Laboratory
Pindzola, Michael	Auburn University
Platt, Darren M.	Joint Genome Institute
Plechac, Petr	University of Tennessee
Polys, Nicholas	Virginia Tech
Poole, Stephen W.	Oak Ridge National Laboratory
Pordes, Ruth	Fermi National Laboratory
Post, Wilfred	Oak Ridge National Laboratory
Potok, Thomas	Oak Ridge National Laboratory
Pratson, Lincoln F.	Duke University
Price, Nathan	Institute for Systems Biology
Pugmire, David	Los Alamos National Laboratory
Quinn, Thomas	University of Washington
Racunas, Stephen	Stanford University
Radhakrishnan, Bala	UT-Battelle
Raghavan, Padmasani	Pennsylvania State University
Ray, Jaideep	Sandia National Laboratory
Ricker, Paul	University of Illinois at Urbana-Champaign
Rintoul, Mark	Sandia National Laboratories
Robert, Sahoff	Oak Ridge National Laboratory
Robinson, Sharon M.	Oak Ridge National Laboratory
Roche, Kenneth	Oak Ridge National Laboratory
Rogers, Jim	Oak Ridge National Laboratory
Rosen, Benjamin	Dell, Inc.
Roskies, Ralph Z.	Pittsburgh Supercomputing Center
Rosner, Robert	Argonne National Laboratory
Ross, Robert B.	Argonne National Laboratory
Rotem, Doron	Lawrence Berkeley National Laboratory
Runolfsson, Thordur	University of Oklahoma
Rushton, James E.	Oak Ridge National Laboratory
Ryne, Robert	Lawrence Berkeley National Laboratory

Sabau, Adrian	Oak Ridge National Laboratory
Saied, Faisal	Purdue University
Sale, Michael	Oak Ridge National Laboratory
Salve, Rohit	Lawrence Berkeley National Laboratory
Samatova, Nagiza	Oak Ridge National Laboratory
Sankaran, Ramanan	Oak Ridge National Laboratory
Sarma, Gorti	Oak Ridge National Laboratory
Scheraga, Harold	Cornell University
Schissel , David	General Atomics
Schopf, Jennifer	Argonne National Laboratory
Schryver, Jack	Oak Ridge National Laboratory
Schulthess, Thomas	Oak Ridge National Laboratory
Schulz, Karl W.	Texas Advanced Computing Center
Schwartz, Peter	Lawrence Berkeley National Laboratory
Scott, Stephen	Oak Ridge National Laboratory
Scott, Steve	Cray
Sears, Mark	U.S. Department of Energy
Sekine, Yukiko	U.S. Department of Energy Office of Science
Sethian, James	Lawrence Berkeley National Laboratory
Shalf, John	Lawrence Berkeley National Laboratory
Shaw, Henry	Lawrence Livermore National Laboratory
Sheetz, R. Michael	University of Kentucky
Sheldon, Frederick	UT-Battelle
Shelton, Robert B.	University of Tennessee
Shelton, William A.	Oak Ridge National Laboratory
Shen, Han-Wei	Ohio State University
Shoshani, Arie	Lawrence Berkeley National Laboratory
Siegel, Andrew	Argonne National Laboratory
Simon, Horst	Lawrence Berkeley National Laboratory
Simunovic, Srdjan	Oak Ridge National Laboratory
Sjolander, Kimmen	University of California Berkeley
Skinner, David	Lawrence Berkeley National Laboratory
Skow , Dane	Argonne National Laboratory
Smith, Barry	Argonne National Laboratory
Smith, Jeff W.	UT-Battelle / Oak Ridge National Laboratory
Smith, Jeremy	University of Tennessee
Smith, Melissa C.	Clemson University
Snavely, Allan	San Diego Supercomputing Center
Snyder, Seth	Argonne National Laboratory
Spada, Mary E.	Argonne National Laboratory
Spentzouris, Panagiotis	Fermi National Laboratory
Sterling, Thomas	Louisiana State University
Stevens, Rick	Argonne National Laboratory / University of Chicago
Storaasli, Olaf	Oak Ridge National Laboratory
Straatsma, Tjerk P.	Pacific Northwest National Laboratory
Strayer, Michael	U.S. Department of Energy Office of Science
Strohmaier, Erich	Lawrence Berkeley National Laboratory
Studham, R. Scott	Oak Ridge National Laboratory
Swain, William	University of Tennessee
Swesty, Douglas	State University of New York at Stony Brook
Szeto, Ernest	Lawrence Berkeley National Laboratory / BDMTC
Tang, William M.	Princeton University
Tautges, Timothy	Argonne National Laboratory

Taylor, Mark A.	Sandia National Laboratories
Tentner, Adrian	Argonne National Laboratory
Thakur, Rajeev	Argonne National Laboratory
Tieman, Brian	Argonne National Laboratory
Tierney, Brian	Lawrence Berkeley National Laboratory
Tobis, Michael	University of Texas
Towns, John	National Center for Supercomputing Applications / University of Illinois Urbana-Champaign
Trebotich, David	Lawrence Livermore National Laboratory
Uddin, Rizwan	University of Illinois Urbana-Champaign
Uhrig, Robert	University of Tennessee
Vaiana, Andrea C.	Los Alamos National Laboratory
Vaidya, Sheila	Lawrence Livermore National Laboratory
Vay, Jean-Luc	Lawrence Berkeley National Laboratory
Vazhkudai, Sudharshan	Oak Ridge National Laboratory
Verastegui, Becky	Oak Ridge National Laboratory
Vetter, Jeffrey S.	Oak Ridge National Laboratory
Vogt, David Paul	Oak Ridge National Laboratory
von Laszewski, Gregor	Argonne National Laboratory
Wadsworth, Jeffrey	Oak Ridge National Laboratory
Wang, Lin-Wang	Lawrence Berkeley National Laboratory
Wang, Paul T.	Mississippi State University
Wang, Yuson	Argonne National Laboratory
Ward, Robert C.	University of Tennessee
Warren, Michael	Los Alamos National Laboratory
Weaver, Nicholas	International Computer Science Institute
Wehner, Michael	Lawrence Berkeley National Laboratory
Wei, Yajun	Northwestern University
Weigand, Gilbert G.	Oak Ridge National Laboratory
Werner, Janet	Argonne National Laboratory
Wheat, Stephen	Intel
White III, James	Oak Ridge National Laboratory
Whitfield, David	University of Tennessee
Whitney, Paul	Pacific Northwest National Laboratory
Wilbanks, Thomas J.	Oak Ridge National Laboratory
Wilde, Michael	University of Chicago / Argonne National Laboratory
Wilke, Andreas	University of Chicago
Williams, Dean	Lawrence Livermore National Laboratory
Wing, William	Oak Ridge National Laboratory
Witek, Rich	AMD
Wolf, Matthew	Georgia Tech
Woodward, Paul	University of Minnesota
Worley, Brian	UT-Battelle
Wu, Kesheng	Lawrence Berkeley National Laboratory
Wullschleger, Stan D.	Oak Ridge National Laboratory
Xia, Fangfang	University of Chicago
Yang, Yunfeng	Oak Ridge National Laboratory
Yelick, Kathy	University of California Berkeley
Young, Glenn	Oak Ridge National Laboratory
Yu, Weikuan	Oak Ridge National Laboratory
Zacharia, Thomas	Oak Ridge National Laboratory
Zhang, Yingqi	Lawrence Berkeley National Laboratory
Zinner, Jenifer	University of Chicago

Appendix D

Abbreviations and Terminology

ABBREVIATIONS AND TERMINOLOGY

3D	three-dimensional
aa	amino acid
AGN	active galactic nucleus
AMIGA	All Modular Industry Growth Assessment
AMR	adaptive mesh refinement
ANL	Argonne National Laboratory
AOGCM	Atmosphere-ocean general circulation model
ASCR	Advanced Scientific Computing Research
BBH	binary black hole
BES	Basic Energy Sciences
BG	Blue Gene
BHNS	black hole and neutron star
BNS	binary neutron star
CAF	Co-Array Fortran
CGE	computable general equilibrium
CGRO	Compton Gamma-Ray Observatory
CMS	Compact Muon Solenoid
DAE	differential algebraic equation
DBA	design basis accident
DETF	Dark Energy Task Force
DFT	density functional theory
DVM	dynamic vegetation model
E3	Simulation and Modeling at the Exascale for Energy and the Environment
ELM	edge-localized mode
EMF	Energy Modeling Forum
EOS	equation of state
ESM	earth system model
EVLA	Enhanced Very Large Array
EXIST	Energetic X-ray Imaging Survey Telescope
FFT	fast Fourier transform
flops	floating point operations per second
FPGA	field programmable gate array
FUSE	Far Ultraviolet Spectroscopic Explorer
Gbps	gigabits per second
GIS	geographic information system
GK	gyrokinetic
GLAST	Gamma-ray Large Area Space Telescope
GMT	Giant Magellan Telescope
GNEP	Global Nuclear Energy Partnership
GRB	gamma-ray burst
GTC	Gyrokinetic Toroidal Code
HCCI	homogeneous charge compression ignition
HPC	high-performance computing

HPCS	High-Productivity Computer Systems
HST	Hubble Space Telescope
I/O	input/output
IDS	intrusion detection system
IEA	International Energy Agency
INTEGRAL	International Gamma-ray Astrophysics Laboratory
IPCC	Intergovernmental Panel on Climate Change
IPS	intrusion prevention system
ISO	Infrared Space Observatory
IUE	International Ultraviolet Explorer
JDEM	Joint Dark Energy Mission
JFNK	Jacobian-free Newton-Krylov
JWST	James Webb Space Telescope
LBNL	Lawrence Berkeley National Laboratory
LES	large eddy simulation
LHC	Large Hadron Collider
LSST	Large Synoptic Survey Telescope
LTC	low-temperature compression
LWR	light water reactor
M&S	modeling and simulation
MHD	magnetohydrodynamic
MIT	Massachusetts Institute of Technology
MPI	message-passing interface
MPP	massively parallel processing (or processors)
MW	megawatts
NASA	National Aeronautics and Space Administration
NCSU	North Carolina State University
NEMS	National Energy Modeling System
NK	Newton-Krylov
NPP	nuclear power plant
NRC	Nuclear Regulatory Commission
NSF	National Science Foundation
OASCR	Office of Advanced Scientific Computing Research
ODE	ordinary differential equation
OLG	overlapping generations
ORNL	Oak Ridge National Laboratory
PB	petabytes
PDE	partial differential equation
PF	petaflops
PIC	particle in cell
R&D	research and development
RF	radio frequency
s-process	slow neutron capture process
SciDAC	Scientific Discovery through Advanced Computing
SitAware	situational awareness

SKA	Square Kilometer Array
SN	supernova
SNF	spent nuclear fuel
SNP	single-nucleotide polymorphism
SOA	secondary organic aerosol
SOC	system on chip
SOS	system of systems
SVD	singular value decomposition
T_c	critical temperature
TDDFT	time-dependent density functional theory
TRU	transuranic
UCSD	University of California, San Diego
UPC	Unified Parallel C
V&V	verification and validation
VIRGO	Variability of Solar Irradiance and Gravity Oscillations
VO	virtual organization
WEO	World Energy Outlook

Appendix E

References

Section 1: Climate

R. Alley, T. Berntsen, N. L. Bindoff, Z. Chen, A. Chidthaisong, P. Friedlingstein, J. Gregory, G. Hegerl, M. Heimann, B. Hewitson et al. (2007), Climate change 2007: The physical science basis, IPCC, Working Group 1 for the Fourth Assessment, WMO.

CNA Corporation (2007), National security and the threat of climate change. National Research Council (2001), Improving the effectiveness of U.S. climate modeling, National Academies Press.

S. Doney, ed. (2004), Ocean carbon and climate change: An implementation strategy for U.S. ocean carbon research, U.S. Carbon Cycle Science Scientific Steering Group.

Ian Foster, ed. (2007), Exascale global socioeconomic modeling enabling comprehensive climate change impact and response analysis, DOE.

L. Ma, M. J. Shaffer and L. R. Ahuja (2001), Application of RZWQM for soil nitrogen management, pp. 265–301 in M. J. Shaffer, L. Ma and S. Hansen, eds., Modeling Carbon and Nitrogen Dynamics for Soil Management, Lewis Publ., Boca Raton, Florida.

S. Stich, B. Smith, C. I. Prentice, A. Arneth, A. Bondeau, W. Cramer, J. Kaplan, S. Levis, W. Lucht, M. Sykes, K. Thonicke and S. Venevsky (2003), Evaluation of ecosystem dynamics, plant geography, and terrestrial carbon cycling in the LPJ dynamic global vegetation model, *Glob. Change Biol.* 9, 161–185.

Section 2A: Energy—Combustion

DOE Office of Basic Energy Sciences 2006. Basic Research Needs for Clean and Efficient Combustion of 21st Century Transportation Fuels (http://www.sc.doe.gov/bes/reports/files/CTF_rpt.pdf).

DOE Office of Fossil Energy 2004. FutureGen Integrated Hydrogen, Electric Power Production and Carbon Sequestration Research Initiative (http://www.fossil.energy.gov/programs/powersystems-/futuregen/futuregen_report_march_04.pdf).

Section 2B: Energy—Nuclear Fusion

D. A. Batchelor, M. Beck, A. Becoulet, R. V. Budny, C. S. Chang, P. H. Diamond, J. Q. Dong, G. Y. Fu, A. Fukuyama, T. S. Hahm, D. E. Keyes, Y. Kishimoto, S. Klasky, L. L. Lao, K. Li, Z. Lin, B. Ludaescher, J. Manickam, N. Nakajima, T. Ozeki, N. Podhorszki, W. M. Tang, M. A. Vouk, R. E. Waltz, S. J. Wang, H. R. Wilson, X. Q. Xu, M. Yagi and F. Zonca (2007), Simulation of fusion plasmas: Current status and future direction. *Plasma Sci. Technol.* 9:312–387.

S. Ethier, W. M. Tang and Z. Lin (2005), Gyrokinetic particle-in-cell simulations of plasma microturbulence on advanced computing platforms. *J. Phys. Conf. Series* 16, 1–15.

S. Ethier, W. M. Tang, R. Walkup and L. Oliker (2007), Large scale gyrokinetic particle simulation of microturbulence in magnetically confined fusion plasmas, *IBM J. Res. Dev.*, accepted for publication.

W. W. Lee (1983), Gyrokinetic approach in particle simulation. *Phys. Fluids* 26(2) 556–562.

W. W. Lee (1987), Gyrokinetic particle simulation model. *J. Comput. Phys.* 72, 243–269.

Z. Lin, T. S. Hahm, W. W. Lee, W. M. Tang and R. B. White (1998), Turbulent transport reduction by zonal flows: Massively parallel simulations. *Science* 281, 1835–1837.

Z. Lin, T. S. Hahm, W. W. Lee, W. M. Tang and R. B. White (2000), Gyrokinetic simulations in general geometry and applications to collisional damping of zonal flows. *Phys. Plasmas* 7(5) 1857–1862.

Leonid Oliker, Andrew Canning, Jonathan Carter, John Shalf and Stephane Ethier (2004), Scientific computations on modern parallel vector systems. In *Proc. SC2004*, Pittsburgh, PA..

L. Oliker, A. Canning, J. Carter, C. Iancu, M. Likewski, S. Kamil, J. Shalf, H. Shan, E. Strohmaier, S. Ethier, and T. Goodale (2007), Scientific application performance on candidate petaScale platforms. In *Proc. IPDPS'07*, Long Beach, CA.

L. Oliker, J. Carter, M. Wehner, A. Canning, S. Ethier, B. Govindasamy, A. Mirin, D. Parks, P. Worley, S. Kitawaki, Y. Tsuda (2005), Leading computational methods on scalar and vector HEC platforms. In *Proc. SC2005*, Seattle, WA.

W.M. Tang and V. S. Chan (2005), Advances and challenges in computational plasma science. *Plasma Phys. Control. Fusion* 47(2) R1–R34.

Section 2C: Energy—Solar

DOE office of Basic Energy Sciences 2005. Basic Research Needs for Solar Energy uti¬lization: Report of the Basic Energy Sciences Workshop on Solar Energy Utilization, April 18-21, 2005. (http:www.sc.doe.gov/bes/re¬ports/files/SEU_rpt.pdf).

Section 2D: Energy—Nuclear Fission

DOE Office of Basic Energy Sciences (2006), *Basic Research Needs for Advanced Nuclear Energy Systems: Report of the Basic Energy Sciences Workshop on Basic Research Needs for Advanced Nuclear Energy Systems,* July 31–August 3, 2006 (http://www.sc.doe.gov/bes/reports/files/ANES_rpt.pdf).

DOE Office of Nuclear Energy, Science, and Technology (2006), *Global Nuclear Energy Partnership Technology Development Plan,* Global Nuclear Energy Partnership Technology Development Program Integration Office, January 9, 2007.

Section 3: Biology

S. S. Goens, S. Botero, A. Zemla, C. E. Zhou and M. L. Perdue (2004), *J. Gen. Virol.* 85, 3195.

E. Klipp, W. Liebermeister, A. Helbig, A. Kowald and J. Schaber (2007), *Nature Biotechnol.* 25, 390.

Section 4: Socioeconomic Modeling

H. M. Amman, D. A. Kendrick and J. Rust, eds. (1996), Handbook of Computational Economics, vol. 1. Elsevier.

A. Dixit and R. Pindyck (1992), Investment Under Uncertainty. MIT Press.

J. Edmonds, M. Wise, H. Pitcher, R. Richels, T. Wigley and C. MacCracken (1997), An integrated assessment of climate change and the accelerated introduction of advanced energy technologies: An application of MiniCAM 1.0, *Mitigation and Adaptation Strategies for Global Change* 1(4) 311–339.

EMF Study-21 (2006), Special issue on multi-greenhouse gas mitigation and climate policy, EMF Study 21. *The Energy Journal*.

B. C. English, D. G. De La Toore Ugarte, K. Jenson, C. Hellwinckel, J. Menard, B. Wilson, R. Roberts and M. Welsh (2006), 25% renewable energy for the United States by 2025: Agricultural and economic impacts. Department of Agricultural Economics, University of Tennessee.

S. W. Hadley, D. J. Erickson III, J. L. Hernandez, C. T. Broniak and T. J. Blasing (2006), Responses of energy use to climate change: A climate modeling study. *Geophys. Res. Lett.*, 33 (L17703).

D. A. Hanson and J. A. Laitner (2006), Technology policy and world greenhouse gas emissions in the AMIGA modeling system, *The Energy Journal* (Special Issue on Multi-Greenhouse Gas Mitigation and Climate Policy).

I. Held and B. Soden (2006), Robust responses of the hydrological cycle to global warming. *Journal of Climate* 19:5686–5699.

IEA (2007), International Energy Agency. World Energy Outlook, www.worldenergyoutlook.org.

IMAGE 2.2 Model Flow Diagram (2007), www.mnp.nl/image/model_details.

IMAGE team (2001), The IMAGE 2.2 implementation of the SRES scenarios: A comprehensive analysis of emissions, climate change and impacts in the 21st century, RIVM CD-ROM publication 481508018, National Institute for Public Health and the Environment, Bilthoven, the Netherlands.

IPCC (2007), Intergovernmental Panel on Climate Change (IPCC) 4th assessment report, www.ipcc.ch, 2007.

K. Judd (1998), Numerical Methods in Economics, MIT Press.

J. A. Laitner and D. A. Hanson (2006), Modeling detailed energy-efficiency technologies and technology policies within a CGE framework, *The Energy Journal* (Special Issue on Hybrid Modeling of Energy-Environmental Policies: Reconciling Bottom-up and Top-down).

A. J. McMichael, D. H. Campbell-Lendrum, C. F. Corvalán, K. L. Ebi, A. K. Githeko, J. D., Scheraga and A. Woodward, eds. (2003), Climate change and human health: Risks and Responses. World Health Organization, Geneva.

B. C. Murray, A. J. Sommer, B. Depro, B. L. Sohngen, B. A. McCarl, D. Gillig, B. D. Angelo and K. Andrasko (2005), Greenhouse gas mitigation potential in US forestry and agriculture. EPA Report 430-R-05-006.

NAPAP Integrated Assessment Report. National Acid Precipitation Assessment Program (NAPAP), Office of the Director, 1990.

The National Energy Modeling System: An overview (2003). Energy Information Administration, U.S. Department of Energy.

A. P. Sokolov, C. A. Schlosser, S. Dutkiewicz, S. Paltsev, D. W. Kicklighter, H. D. Jacoby, R. G. Prinn, C. E. Forest, J. Reilly, C. Wang, B. Felzer, M. C. Sarofim, J. Scott, P. H. Stone, J. M. Melillo and J. Cohen (2005), The MIT Integrated Global System Model (IGSM) version 2: Model description and baseline evaluation. MIT.

M. L. Stein (1999), Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Sustainable Bioenergy: A Framework for Decision Makers (2006), UN Energy.

M. Q. Wang (2001), Development and use of GREET 1.6 fuel-cycle model for transportation fuels and vehicle technologies (2001), Center for Transportation Research, Argonne National Laboratory.

Section 5: Astrophysics

Andreas Albrecht, Gary Bernstein, Robert Cahn, Wendy L. Freedman, Jacqueline Hewitt, Wayne Hu, John Huth, Marc Kamionkowski, Edward W. Kolb, Lloyd Knox, John C. Mather, Suzanne Staggs and Nicholas B. Suntzeff (2006), Report of the Dark Energy Task Force, arXiv.org:astro-ph/0609591.

John G. Baker, Joan Centrella, Dae-Il Choi, Michael Koppitz and James van Meter (2006), Gravitational wave extraction from an inspiraling configuration of merging black holes. *Phys. Rev. Letters* 96:111102.

J. M. Blondin and A. Mezzacappa (2007), Pulsar spins from an instability in the accretion shock of supernovae, *Nature* 445:58–60.

A. C. Calder, D. M. Townsley, I. R. Seitenzahl, F. Peng, O. E. B. Messer, N. Vladimirova, E. F. Brown, J. W. Truran and D. Q. Lamb (2007), Capturing the fire: flame energetics and neutronization for Type Ia supernova simulations, *ApJ* 656:313–332.

M. Campanelli, C. O. Lousto, P. Marronetti and Y. Zlochower (2006), Accurate evolutions of orbiting black-hole binaries without excision, *Phys. Rev. Letters* 96: 111101.

V. N. Gamezo, A. M. Khokhlov and E. S. Oran (2005), Three-dimensional delayed-detonation model of Type Ia supernovae, *ApJ* 623:337–346.

E. J. Hallman, B. W. O'Shea, J. O. Burns, M. L. Norman, R. Harkness and R. Wagner (2007), The Santa Fe Light Cone Simulation Project, I: Confusion and the WHIM in upcoming Sunyaev-Zel'dovich effect surveys, ArXiv e-prints 704:0704.2607.

Wayne Hu and Scott Dodelson (2002), Cosmic microwave background anisotropies, *Ann. Rev. Astronomy and Astrophysics* 40:171.

Hans-Thomas Janka, K. Langanke, A. Marek, G. Martinez-Pinedo and B. Mueller (2007), Theory of core-collapse supernovae, *Phys. Rept.* 442:38–74.

D. Lynden-Bell (1969), Galactic nuclei as collapsed old quasars, *Nature* 223:690.

Anthony Mezzacappa (2005), Ascertaining the core-collapse supernova mechanism: The state of the art and the road ahead, *Ann. Rev. Nucl. Part. Sci.* 55, 467–515 (2005).

Carlos F. Sopuerta, Ulrich Sperhake and Pablo Laguna (2006), Hydro-without-hydro framework for simulations of black hole-neutron star binaries, *Classical and Quantum Gravity* 23:S579.

S. E. Woosley and J. S. Bloom (2006), The supernova – gamma-ray burst connection, *Ann. Rev. Astronomy and Astrophysics* 44:507.

Section 6: Math & Algorithms

A. T. Adai, S. V. Date, S. Wieland and E. M. Marcotte (2004), LGL: Creating a map of protein function with an algorithm for visualizing very large biological networks, *J. Mol. Biol.* 340(1):179-190.

D. A. Bader (2004), Computational biology and high-performance computing, special issue on bioinformatics, *Communications of the ACM* 47(11):34–41.

D. A. Bader, A. Snavely and G. Jacobs (2006), NSF workshop report on petascale computing in the biological sciences, August 29–30, Arlington, VA.

D. H. Bailey (2005), High-precision arithmetic in scientific computation, *Computing in Science and Engrg.*, May–June, 54–61.

M. Barad and P. Colella (2005), A fourth-order accurate local refinement method for Poisson's equation, *J. Comp. Physics* 209(1) 1–18.

M. J. Berger and S. H. Bokhari (1987), A partitioning strategy for nonuniform problems on multiprocessors, *IEEE Trans. Computers* C-36(5): 570–580.

J. Birge and F. Louveaux (1997), Introduction to Stochastic Programming, Springer.

U. Catalyurek and C. Aykanat (1996), Decomposing irregularly sparse matrices for parallel matrix-vector multiplications, *Lecture Notes in Computer Science* 1117: 75–86.

U. Catalyurek and C. Aykanat (1999), Hypergraph-partitioning-based decomposition for parallel sparse-matrix vector multiplication, *IEEE Trans. Parallel Dist. Systems* 10(7): 673–693.

S. Chandra, M. Parashar and J. Ray (2007), Analyzing the impact of computational heterogeneity on runtime performance of parallel scientific components, in Proc. 15th High Performance Computing Symposium (HPC-07), SCS Spring Simulation Multiconference, Norfolk, VA.

B. Christianson and M. Cox (2005), Automatic propagation of uncertainties, pp. 47–58 in M. Buecker, G. Corliss, P. Hovland, U. Naumann, and B. Norris, eds., Automatic Differentiation: Applications, Theory, and Implementations. *Lecture Notes in Computational Science and Engineering* vol. 50, Springer.

K. Devine, E. Boman, R. Heaphy, R. Bisseling and U. Catalyurek (2006), Parallel hypergraph partitioning for scientific computing, in *Proc. IPDPS 2006*.

V. Eijkhout (1998), Overview of iterative linear system solver packages, *NHSE Review* 3(1).

J. Gondzio and R. Sarkissian (2003), "Parallel Interior-Point Solver for Structured Linear Programs." *Mathematical Programming* 96: 561-584.

B. Hayes (2003), A lucid interval, *American Scientist* 91(6) 484–488.

B. Hendrickson and R. Leland (1995), A multilevel algorithm for partitioning graphs. in *Proc. Supercomputing ‘95*.

H. Johansson and J. Steensland (2006), A performance characterization of load balancing alogrithms for parallel SAMR applications, Technical Report 2006-047, Uppsala University, Dept. of Information Technology.

G. Karypis and V. Kumar (1998), A fast and high quality multilevel scheme for partitioning irregular graphs, *SIAM J. Sci. Computing* 20(1) 359–392.

G. Karypis and V. Kumar (1997), A coarse-grain parallel multilevel *k*-way partitioning algorithm, in *Proc. 8th SIAM Conf. Parallel Processing for Scientific Computing*.

G. J. Klir (1994), The many faces of uncertainty, pp. 3–19 in B. M Ayyub and M. M. Gupta, eds., Uncertainty Modeling and Analysis: Theory and Applications, *Elsevier Science*.

X. Li (2006), Direct solvers for sparse matrices, http://crd.lbl.gov/~xiaoye/SuperLU/SparseDirectSurvey.pdf, September.

G. Nemhauser and L. Wolsey (1988), Integer and Combinatorial Optimization, John Wiley & Sons.

J. Nocedal and S. Wright (2006), Numerical Optimization, 2^{n} ed., Springer.

M. J. North and C. M. Macal (2007), Managing Business Complexity: Discovering Strategic Solutions with Agent-Based Modeling and Simulation, Oxford.

A. Patra and J. T. Oden (1995), Problem decomposition strategies for adaptive hp finite element methods, Computing Systems in Eng. 6(2) 97–109.

J. R. Pilkington and S. B. Baden (1994), Partitioning with spacefilling curves, CSE Technical Report CS94-349 Dept. Computer Science and Engineering, University of California – San Diego.

J. Ray, C. A. Kennedy, S. Lefantzi and H. N. Najm (2007), Using high-order methods on adaptively refined block-structured meshes derivatives, interpolations, and filters, *SIAM J. Sci. Computing* 29(1):139–181.

P. Schwartz, M. Barad, P. Colella and T. Ligocki (2006), A Cartesian grid embedded boundary method for the heat equation and Poisson's equation in three dimensions, *J. Comput. Phys.* 211(2) 531–550.

H. D. Simon (1991), Partitioning of unstructured problems for parallel processing, *Computing Systems in Engrg.* 2: 135–148.

J. Steensland (2002), Efficient partitioning of dynamic structured grid hierarchies, Uppsala University Library, Uppsala, Sweden.

J. Steensland and J. Ray (2003), A heuristic remapping algorithm reducing inter-level communication in SAMR applications, in *Proc. 15th IASTED International Conference on Parallel and Distributed Computing and Systems 2003* (PDCS03).

R. W. Walters and L. Huyse (2002), Uncertainty analysis for fluid mechanics with applications. Technical report, NASA, Feb.

M. Woolridge (2002), An Introduction to Multi-Agent Systems. Wiley & Sons.

Section 7: Software

Ray Bair, Lori Diachin, Stephen Kent, George Michaels, Anthony Mezzacappa, Richard Mount, Ruth Pordes, Larry Rahn, Arie Shoshani, Rick Stevens, and Dean Williams (2003), Planning ASCR/Office of Science data-management strategy. http://www-conf.slac.stanford.edu/dmw2004/docs/DM-strategy-final.doc.

J. Duell, P. Hargrove and E. Roman (2002), The Design and Implementation of Berkeley Lab's Linux Checkpoint/Restart, Berkeley Lab Technical Report LBNL-54941.

R. L. Graham, S.-E. Choi, D. J. Daniel, N. N. Desai, R. G. Minnich, C. E. Rasmussen, L. D. Risinger and M. W. Sukalksi (2003), A network-failure-tolerant message-passing system for terascale clusters, *Int. J. Parallel Programming* 31(4) 285-303.

W. Gropp and Ewing Lusk (2004), Fault tolerance in Message Passing Interface programs, *Int. J. High Perform. Comput. Appl.* 18(3):363-372.

J. Li, W. Liao, A. Choudhary, R. Ross, R. Thakur, W. Gropp, R. Latham, A. Siegel, B. Gallagher, and M. Zingale (2003), Parallel netCDF: A high-performance scientific I/O interface, in *Proc. SC2003*, Phoenix, AZ.

E. Lusk and Kathy Yelick (2007), Languages for high-productivity computing: The DARPA HPCS language project, *Parallel Processing Letters* 17(1), March.

J. Parker, T. Engelsiepen, R. Ross, R. Thakur, R. Latham, and W. Gropp (2006), High-performance file I/O for the BlueGene/L supercomputer, in *Proc. 12th International Symposium on High-Performance Computer Architecture* (HPCA-12).

R. Ross, Jose Moreira, Kim Cupps, and Wayne Pfeiffer (2006), Parallel I/O on the IBM Blue Gene /L system, BlueGene Consortium Quarterly Newsletter, February.

Rob Ross, Evan Felix, Bill Loewe, Lee Ward, Gary Grider and Rob Hill (2005), HPC file systems and scalable I/O: Suggested research and development topics for the fiscal 2005–2009 time frame, ftp://ftp.lanl.gov/public/ggrider/HEC-IWG-FS-IO-Workshop-08-15-2005/FileSystems-DTS-SIO-FY05-FY09-R&D-topics-final.pdf.

Douglas Thain, Todd Tannenbaum and Miron Livny (2005), Distributed computing in practice: The Condor experience, *Concurrency – Practice and Experience* 17(2–4) 323-356.

Section 8: Hardware

J. Blyler (2005), Navigating the Silicon Jungle: FPGA or ASIS?, Chip Design, June/July, http://www.chipdesignmag.com/display.php?articleId=115&issueId=11

C. Coarfa, Y. Dotsenko, J. Mellor-Crummey, F. Cantonnet, T. El-Ghazawi, A. Mohanty and Y. Yao (2006), An evaluation of global address space languages: Co-array Fortran and Unified Parallel C, in *Proc. Principles and Practice of Parallel Programming (PPoPP),* New York.

M. Elnozahy (2006), IBM has its PERCS, HPCWire 15(14), April. http://www.hpcwire.com/hpc/614724.html.

F. Johnson (2006), DARPA HPCS Program, *SciDAC Review 2*, Fall, http://www.scidacreview.org/0602/html/news2.html

A. Krasnov, Andrew Schultz, John Wawrzynek, Greg Gibeling and Pierre-Yves Droz *(2007),* RAMP Blue: A message-passing manycore system in FPGAs, in *Proc. FPL 2007 - International Conference on Field Programmable Logic and Applications*, Amsterdam.

C. Maxcer (2007), D-Wave claims quantum computing breakthrough, *TechNewsWorld*, www.technewsworld.com/story/55801.html.

M. Reilly, L. C. Stewart, J. Leonard and D. Gingold (2006), SiCortex technical summary, www.sicortex.com/whitepapers/sicortex-tech_summary.pdf.

D. E. Shaw et al. (2007), Anton, a special-purpose machine for molecular dynamics simulations, in *Proc. 34th Annual International Symposium on Computer Architecture,* San Diego, pp. 1-12.

M. Wehner, L. Oklier, and J. Shalf (2007), Towards ultra-high performance resolution models of Climate and Weather, *Intl. J. High-Performance Comp. Apps.*, to appear.

Section 9: Cyberinfrastructure

Bill Allcock, Joe Bester, John Bresnahan, Ann L. Chervenak, Carl Kesselman, Sam Meder, Veronika Nefedova, Darcy Quesnel, Steven Tuecke and Ian Foster (2001), Secure, efficient data transport and replica management for high-performance data-intensive computing, p. 13 in *Proc. 18th IEEE Symposium on Mass Storage Systems.*

Charlie Catlett et al. (2007), TeraGrid: Analysis of organization, system architecture, and middleware enabling new types of applications, in *High Performance Computing and Grids in Action,* IOS Press, Advances in Parallel Computing series, Amsterdam.

Narayan Desai, Andrew Lusk, Rick Bradshaw and Remy Evard (2003), BCFG: A configuration management tool for heterogeneous environments, in *Proc. IEEE International Conference on Cluster Computing*, pp. 500, 2003.

Ian Foster, Jens Voeckler, Michael Wilde and Yong Zhao (2003), The virtual Data Grid: A new model and architecture for data-intensive collaboration, in *Proc. Conference on Innovative Data Systems Research.*

Veronika Nefedova, Robert Jacob, Ian Foster, Zhengyu Liu, Yun Liu, Ewa Deelman, Gaurang Mehta, Mei-Hui Su and Karan Vahi (2006), Automating climate science: Large ensemble simulations on the TeraGrid with the GriPhyN virtual data system, p. 32 in *Proc. 2nd IEEE International Conference on e-Science and Grid Computing.*

Rick Stevens (2003), Access Grid: Enabling group-oriented collaboration on the Grid, in The Grid: Blueprint for a New Computing Infrastructure, Ian Foster and Carl Kesselman, eds., Morgan Kaufmann.

Rick Stevens, Michael E. Papka and Terry Disz (2003), Prototyping the workspace of the future, *IEEE Internet Computing* 7(4): 51–58.

Gregor von Laszewski, Michael Hategan and Deepti Kodeboyina (2007), Java CoG Kit workflow, pp. 340-356 in *Workflows for Science*, I. Taylor, E. Deelman, D. B. Gannon, and M. Shields, eds., Springer.

Von Welch, Jim Barlow, James Basney, Doru Marcusiu and Nancy Wilkins-Diehr (2006), A AAAA model to support science gateways with community accounts, *Concurrency and Computation: Practice and Experience* 19(6):893-904.

Yong Zhao, Michael Wilde and Ian Foster (2007), Virtual Data Language: A Typed Workflow Notation for Diversely Structured Scientific Data, pp. 258-278 in *Workflows for eScience,* Springer.

DISCLAIMER

This report was prepared as an account of a workshop sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees or officers, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of document authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

E^3

Office of Science
U.S. DEPARTMENT OF ENERGY

ASCR