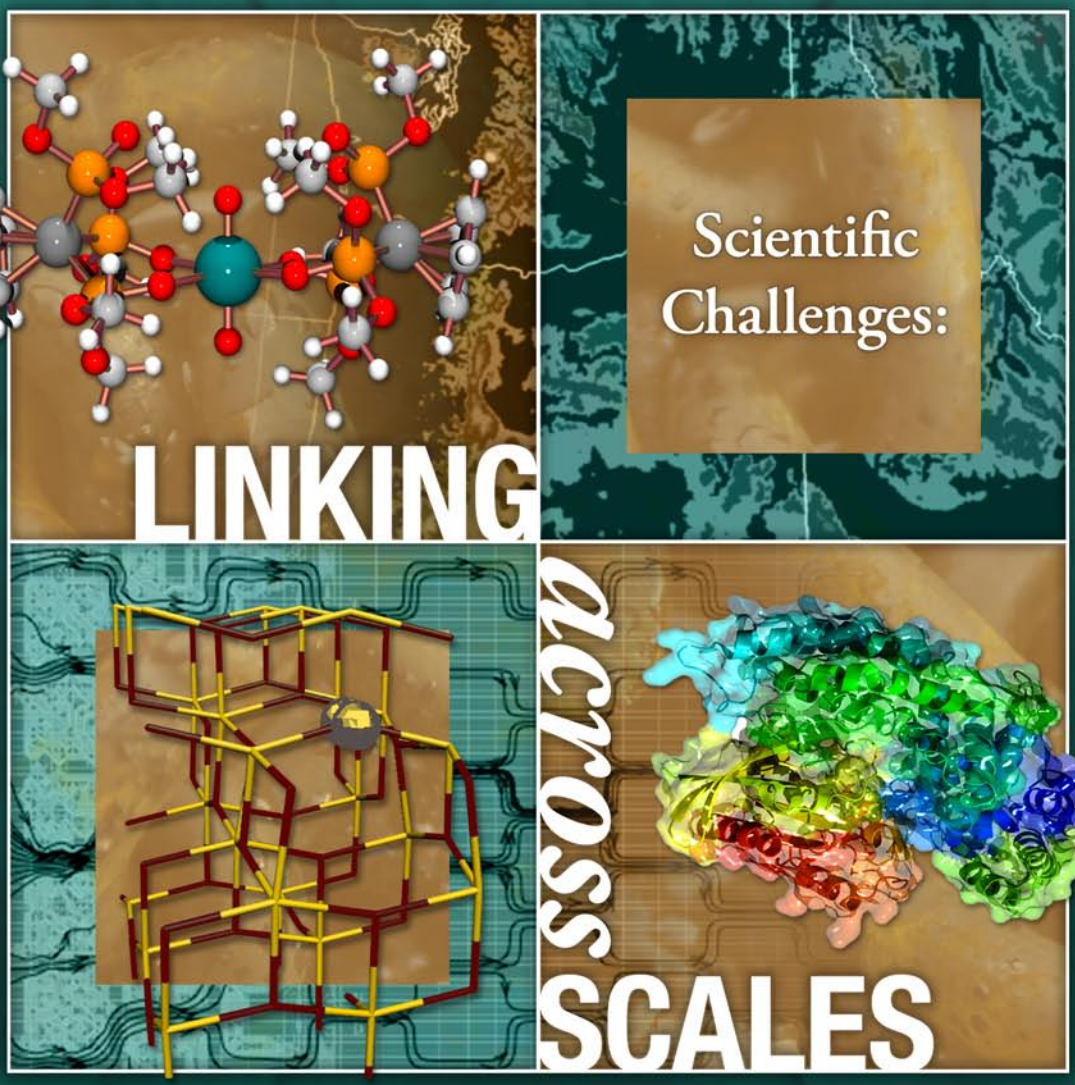


William R. Wiley Environmental Molecular Sciences Laboratory

MOLECULAR SCIENCE COMPUTING FACILITY

Pacific Northwest National Laboratory

Prepared for the U.S. Department of Energy under Contract DE-AC05-76RL01830

July 2005

William R. Wiley Environmental Molecular Sciences Laboratory

Molecular Science Computing Facility

Scientific Challenges:
LINKING *across* SCALES

Edited by
Wibe A. de Jong
Theresa L. Windus

July 2005

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830
Pacific Northwest National Laboratory
Richland, WA 99352

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

PNNL-15144

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service,
U.S. Department of Commerce, 5285 Port Royal Rd., Springfield, VA 22161
ph: (800) 553-6847
fax: (703) 605-6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/ordering.htm

This document was printed on recycled paper.
(05/2005)

Acknowledgments

The Molecular Sciences Computing Facility document "Scientific Challenges: Linking Across Scales" reflects contributions by white paper authors and workshop participants from universities and national laboratories listed in Appendices A and B. The editors wish to thank all who provided invaluable input to this document. Special thanks goes to Dave Dixon, University of Alabama; Bruce Garrett, Bruce Palmer, Steve Yabusaki, and Steve Ghan, PNNL; Chris Oehmen and Erich Vorpagel, EMSL, for their help in the compilation and writing of the scientific drivers chapter of this document; Stephan Elbert, PNNL, for his review and comments to the document; and to Julia White, EMSL, for her contributions to the document introduction.

The editors also wish to thank the PNNL staff who provided support in the production of the document: Joanne Stover and Mary Ann Showalter for technical editing and Chris DeGraaf for graphic design and layout. The following EMSL staff were instrumental in the organization and success of the workshop: Lisa Hobson, Becky Wattenburger, Tina Foley, and Jessica Foreman.

Wibe A. de Jong

Theresa L. Windus

On the Cover

Top left image: Uranyl-Klaui complex (courtesy of Wibe de Jong and Jun Li, EMSL).

Bottom left image: Orbital density of H^+ adsorped to TiO_2 (courtesy of Eric Bylaska, PNNL).

Bottom right image: Ras-RasGAP protein complex (courtesy of Yuri Alexeev and Marat Valiev, EMSL).

Background of top two boxes: Simulated snowpack in Washington state (courtesy of Steve Ghan and Ruby Leung, PNNL).

Background of bottom two boxes: Simulation of reactive fluid flow through minerals (courtesy of Peter Lichtner, LANL).

Table of Contents

Executive Summary

The previous 50 years have seen tremendous leaps in the development of advanced computing capabilities. Where once scientists were limited to simulating molecular systems on the order of tens of atoms, researchers now model microscopic systems on the order of thousands of atoms, such as cell membranes. This has led to new insight and an increasing ability to address complex national issues related to our environment. National leaders recognize the environmental impact and need for remediation of legacy waste from weapons production activities of previous generations as well as the critical need to understand and mitigate the effects of current energy production on the environment.

"Computational science—the use of advanced computing capabilities to understand and solve complex problems—is now critical to scientific leadership, economic competitiveness, and national security."

John H. Marburger III
Presidential Science Advisor and Director of the Office of Science and Technology Policy

Significant progress has been made in our ability to model systems of limited temporal and spatial extent, but now these models must be linked across scales in both time and space in order to fully address the biological, chemical, and environmental behavior inherent to such complex processes. Simulation of scientific challenges across scales will require leading-edge computing facilities and capabilities uniquely tailored toward solving large multiscale problems.

One national facility—the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL)—is poised to implement these advanced computational capabilities and combine them with leading-edge experimental research to provide a cross-disciplinary environment for world-class scientists to obtain solutions to current and future environmental research challenges.

The purpose of this document is to define the evolving science drivers for performing computational environmental molecular research at EMSL and to provide guidance associated with the next-generation high-performance computing center that must be developed at EMSL's Molecular Science Computing Facility (MSCF) in order to address this critical research. The MSCF is the pre-eminent computing facility—supported by the U.S. Department of Energy's (DOE's) Office of Biological and Environmental Research—tailored to provide the fastest time-to-solution for current computational challenges in chemistry and biology, as well as providing the means for broad research in the molecular and environmental sciences. The MSCF provides resources and expertise to emerging EMSL Scientific Grand Challenges and Collaborative Access Teams that are designed to leverage the multiple integrated research capabilities of EMSL, thereby creating a synergy between computation and experiment to address environmental molecular science challenges critical to DOE and the nation.

The MSCF must maintain leading-edge capabilities so the scientific community can continue to address new and more complex environmental scientific challenges.

This document describes such future challenges in the areas of biological and chemical sciences and environmental systems, the role that MSCF computing and expert staff resources will play, and how the MSCF and its recommended upgraded computational resources will positively impact the environmental missions of DOE. This document is comprised of input received from white papers submitted by prestigious biology, chemistry, and environmental systems researchers and results from a two-day workshop where top scientific community representatives provided insight into the major scientific drivers of the next three to five years. This document is not just an extrapolation of ongoing science, but a visionary analysis of new scientific possibilities that have yet to take full advantage of high-performance computing. It also provides a roadmap for the next stage of development of the MSCF, uniquely positioning this facility to provide solutions for environmental challenges via a holistic approach that couples phenomena across temporal and spatial scales.

To enable the science described in this document, a *balanced* architecture is recommended with respect to processor, memory hierarchy, interprocessor communication, and disk access and storage. A single architecture could satisfy the needs of all science areas described, although it is recognized that some science areas can take greater advantage of certain aspects of the architecture. In addition to hardware, it is recommended that the MSCF continues to provide a complete collaborative production environment with dedicated expert staff and sophisticated software that enables researchers to solve the large-scale scientific challenges described in this document. The proposed next-generation computing resources will increase EMSL's scientific ability to enable innovative research with cross-cutting contributions to the nation's most challenging environmental and molecular problems. Maintaining state-of-the-art tools for simulation is as critical as tools for experiment, with the resonance between theory and experiment amplified by the quality of simulation.

Acronyms

BER	DOE Office of Biological and Environmental Research
BLAS	Basic Linear Algebra Subprograms
CAM	National Center for Atmospheric Research's Community Atmosphere Model
CAT	Collaborative Access Team
CCl_4	carbon tetrachloride
CCSP	U.S. Climate Change Science Plan
CO_2	carbon dioxide
CPU	central processing unit
CRM	cloud resolving model
CCSD(T)	coupled cluster with singles, doubles, and approximate triples
DFT	density functional theory
DOE	U.S. Department of Energy
Ecce	Extensible Computational Chemistry Environment
EMSL	William R. Wiley Environmental Molecular Sciences Laboratory
ERSD	Environmental Remediation Sciences Division
FPGA	field-programmable gate arrays
GA Tools	Global Array Tools
GVL	Graphics and Visualization Laboratory
I/O	input/output
IR	infrared
MD GRAPE	Molecular Dynamics GRAvity PipE (special purpose hardware)
MMF	Multiscale Modeling Framework
MP2	Møller-Plesset perturbation theory
MPP2	EMSL's Massively Parallel Processing System-2

MS³	Molecular Science Software Suite
MSCF	Molecular Science Computing Facility
NWChem	Northwest Computational Chemistry Software
NWfs	Northwest file system
PES	potential energy surface
PETSc	Portable Extensible Toolkit for Scientific computation
PNNL	Pacific Northwest National Laboratory
RRKM	Rice-Ramsperger-Kassel-Marcus theory
SMnase	*Serratia marcescens Endonuclease*
TCE	tensor contraction engine
VisUS	Visualization and User Services Group

1. Introduction

The Molecular Science Computing Facility (MSCF) is a key element of the William R. Wiley Environmental Molecular Science Laboratory (EMSL), a national user facility located at the Pacific Northwest National Laboratory (PNNL). EMSL and MSCF operations are funded by the U.S. Department of Energy (DOE) Office of Biological and Environmental Research (BER). The MSCF is a unique facility equipped with a high-performance supercomputer, computational resources, and expert staff tailored to address computational grand challenges in the environmental molecular sciences. The MSCF is the pre-eminent DOE computing facility for environmental chemistry and biology research.

The BER program supports world-class fundamental research in the biological, chemical, and environmental sciences to provide innovative solutions to the nation's challenges related to environment and energy production. Environmental research funded by BER focuses on improving the understanding and reliable prediction of climate change and providing science-based solutions for environmental remediation. Programs in the biological sciences support research in genomics and systems biology, where the goal is to understand how living organisms work and interact with and react to their environment. This research will enable the development of biological solutions to produce clean energy, clean up metals and radionuclides in the environment, and reduce carbon dioxide in the atmosphere. BER's Environmental Remediation Sciences Division (ERSD) is the sponsor of EMSL and has a core mission to advance the fundamental science that will lead to solutions for complex environmental problems such as remediation of DOE tank waste sites.

Research at the EMSL is aligned with DOE missions and is focused on gaining a more thorough understanding of the physical, chemical, and biological processes that govern environmental processes starting at the molecular scale and propagating into the larger scales. Specifically, EMSL's vision is to continue to establish distinctive science signatures in the areas of biogeochemistry and subsurface science, interfacial chemistry and catalysis, structure/dynamics of biomolecules and biomolecular complexes, biochemical pathways, aerosol chemistry, and spectra signatures and trace detection.

The William R. Wiley Environmental Molecular Sciences Laboratory (EMSL) at Pacific Northwest National Laboratory is a U.S. Department of Energy Biological and Environmental Research national scientific user facility.

The MSCF is well positioned to contribute extensively to these distinctive science signatures. The facility supports a wide range of computational modeling activities, from benchmark calculations on small molecules to reliable calculations on large molecules, from solids to simulations of large biomolecules, and from reactive chemical transport modeling to multiscale climate modeling. Results of this research serve as a foundation for new science-based solutions to environmental challenges critical to DOE and the nation.

The MSCF is tightly integrated with the other five experimental facilities in EMSL. This integration enables fundamental research on the physical, chemical, and biological processes that underpin scientific issues of interest to DOE and the nation.

The MSCF is tightly integrated with the other five experimental facilities in EMSL. Integration of theory, modeling, and simulation with experiment provides multidisciplinary teams of scientists the advanced experimental and computational resources for fundamental research on the physical, chemical, and biological processes that underpin scientific environmental issues of interest to DOE and the nation.

The purpose of this document is to define the evolving science drivers for performing computational environmental molecular research at EMSL and to provide guidance associated with the next-generation high-performance computing center that must be developed at the MSCF in order to address this critical research. This document describes future environmental challenges in the areas of biological and chemical sciences and environmental systems, the role that MSCF computing and expert staff resources will play, and how the MSCF and its recommended upgraded computational resources will positively impact the environmental missions of DOE.

MSCF

The Molecular Science Computing Facility provides computational resources for projects in basic and applied environmental molecular science that address the environmental problems and research needs facing the U.S. Department of Energy and the nation.

Overview

The MSCF is an integrated production computing environment with specifically designed hardware architecture, software resources, visualization tools, and a dedicated team of hardware operations, scientific consultants, and software developers to accomplish DOE environmental mission goals, with links to external facilities and laboratories within the DOE system, collaborating universities, and industry. Unlike typical computing facilities, strong collaborations exist within the MSCF among experimental researchers, fostered in no small part by the coexistence of the facility with five experimental research facilities at EMSL. This cross-cutting atmosphere positions the MSCF to provide the fundamental discoveries that advance macroscopic theories.

Since research at the MSCF is focused on gaining a more thorough understanding of the physical, chemical, and biological processes that govern environmental processes, the user facility supports a wide range of computational activities in biology, chemistry, and environmental systems, including:

- ***Accurate thermodynamic and kinetic simulations***
- ***Simulations of molecular interactions with environmental materials involving large molecular systems and different scales***
- ***Simulations of biomolecular systems that include protein interactions and functionality, multiprotein molecular machines, signaling networks, and interactions of cells with both their environment and other cells***
- ***Simulation of long-term natural and human impacts on the environment, ranging from the subsurface to the atmosphere, to provide a scientific foundation for policy decisions.***

To meet the demands of current environmental science drivers in the areas of biology, chemistry, and environmental systems, MSCF hardware resources are comprised of a high-performance computing system—referred to in this document as MPP2, the EMSL data storage file system (NWfs), and the Graphics and Visualization Resource Laboratory (GVL). MPP2 is a balanced Hewlett-Packard supercomputer composed of 1,960 1.5-GHz Intel Itanium-2 (Madison) processors with a theoretical peak performance of 11.8 teraflops, 6.8 terabytes of random access memory, 450 terabytes of disk, a Quadrics QsNetII interconnect, and a Linux operating system. Application software resources include the internally developed Molecular Science Software Suite (MS^3)—which consists of Northwest Computational Chemistry Software (NWChem), Extensible

The 11.8-teraflop supercomputer housed in the MSCF is one of the fastest open systems in the United States.

Computational Chemistry Environment (Ecce), Global Array Tools (GA Tools)—and various external software packages.

While the computing architecture is tailored toward providing the fastest time-to-solution on large-scale chemistry problems, the balanced machine has shown good performance for biology, subsurface, and atmospheric cloud modeling as well. These scientific fields greatly benefit from data stored on a machine with a large memory capability, a high-bandwidth interconnect to scale the modeling calculations, and a 53-terabyte global file system to store simulation results.

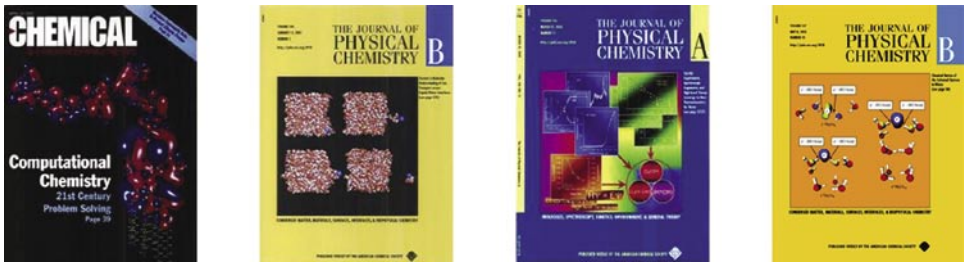
In addition, PNNL has established a high-speed, high-bandwidth optical fiber link from its campus in Richland, Washington, to Seattle, Washington, to connect to DOE's new UltraScience Network and other networks across the nation. This new network connection will increase by as much as 300 times the amount of data that PNNL (and MSCF users) can transmit to and from national and international science communities.

A more comprehensive overview of MSCF capabilities and research thrusts is available on the MSCF website (*http://mscf.emsl.pnl.gov*) and in Appendix C.

Scientific Impact

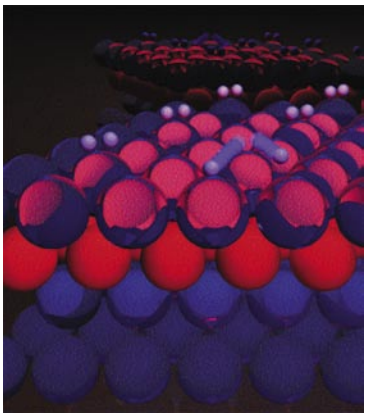
The MSCF originated the concept of Computational Grand Challenges—projects that address complex, large-scale scientific and engineering problems with broad scientific and environmental or economic impacts whose solution can only be advanced by applying high-performance scientific techniques and that use large computational resources. A three-year, externally peer-reviewed Computational Grand Challenge project involves researchers from universities, national laboratories, and industry working together as teams.

With more than 340 publications in peer-reviewed scientific journals in the past three years resulting from Computational Grand Challenges and smaller research projects, the MSCF is clearly making a significant scientific impact in the nation's research community. For example, in Fiscal Year 2004, MSCF users published in more than 100 prestigious scientific journals, such as *Science*, *Nature Materials*,



Research performed at the MSCF has been highlighted on various covers of scientific journals.

Proceedings of the National Academy of Sciences, *Journal of the American Chemical Society*, and *Journal of Physical Chemistry*, and research using MSCF resources was featured in *Chemical and Engineering News* and on the covers of scientific journals. To exemplify the scientific impact generated by the research using MSCF high-performance computing resources, various scientific accomplishments from EMSL Computational Grand Challenge teams are highlighted below.

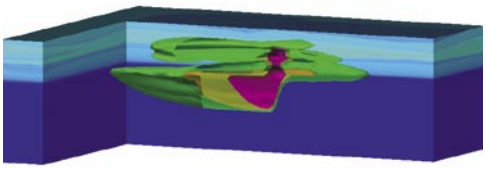


H_2 dissociation on near-surface alloys (image courtesy of Greeley and Mavrikakis, University of Wisconsin-Madison).

Catalysis on Metal Alloys

Jeff Greeley and Manos Mavrikakis of the University of Wisconsin-Madison used MSCF resources to study a new class of bimetallic alloys that have greatly improved catalytic activity and selectivity. They show that some near-surface alloys bind atomic hydrogen as weakly as noble metals, but activate the dissociation of molecular hydrogen much more easily. Near-surface alloys can be prepared using advanced nanomaterial synthesis protocols, and their catalytic properties promise to be useful for a variety of applications, including pharmaceuticals production, hydrogen storage, and anodes for low-temperature fuel cells. This work was published in *Nature Materials* (Volume 3, page 810) and featured in the November 29, 2004, edition of *Chemical and Engineering News*.

Modeling Carbon Tetrachloride Migration in the Subsurface



Simulation of carbon tetrachloride migration and remediation beneath the 216-Z-9 Trench on the Hanford Site using soil vapor extraction (image courtesy of White, PNNL).

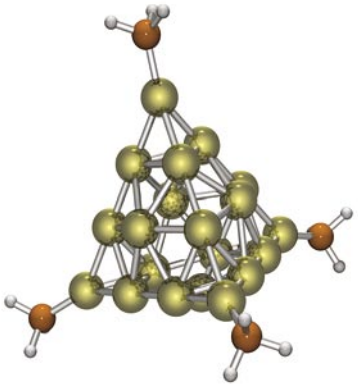
Long-term, scientifically defensible predictions of subsurface contaminant fate are critical to the development of remediation alternatives that can accelerate the cleanup of DOE waste sites and reduce costs and risks to human health and the environment. For example, efficient and cost-effective cleanup of the toxic carbon tetrachloride (CCl_4) plume in the subsurface sediments of the 200 West Area at the Hanford Site largely depends on understanding the distribution and state of this organic contaminant. Mark White and coworkers from PNNL and the Idaho National Laboratory used the MSCF supercomputer and their state-of-the-art, high-resolution, multiple-phase subsurface flow and transport model to simulate the historical migration of CCl_4 at the Hanford Site's 200 West Area from the disposal period (beginning in 1954) to the present, including the soil vapor extraction activities, to predict the current distribution and fate, and to make credible predictions of the distribution and state of the subsurface CCl_4 at the Hanford Site. The new knowledge and capabilities developed by this research have enhanced DOE's ability to understand, assess, and manage risks from contaminated soil and groundwater.

Structure of a subunit of SMnase protein (image courtesy of Pettitt, University of Houston).

Protein Binding and Solvation

The research group lead by Monty Pettitt at the University of Houston is using the MSCF to obtain a fundamental understanding of the physics and chemistry of biochip systems. Biochip technologies use molecular probes tethered on surfaces for identification of gene mutation, DNA-protein binding, pollutant effects on gene and protein expression, contaminated water analysis, and various other applications. One of the group's current studies is that of the role of solvents, such as water, on

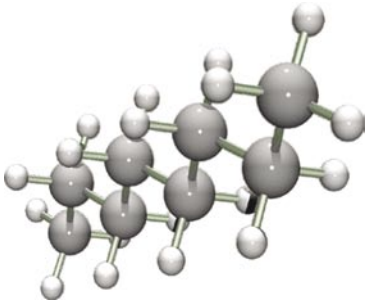
nucleotide/protein interactions. The scientists use molecular dynamics and multiscale computational modeling to study *Serratia marcescens Endonuclease* (SMnase) complexes with DNA. SMnase exists in both monomeric and dimeric forms, both being functional but with the dimer being the conformation of choice in nature. These long-time simulations, impossible without the high-performance computing resources of the MSCF, revealed that the solvent water molecules play an important role in this biomolecular system. Evaluation of water clusters and pathways in the protein led to the rationalization of a possible mechanism for cleavage of the enzyme.



Geometry of Au_{20} cluster (image courtesy of de Jong and Li, EMSL).

Au_{20}—Combining Theory and Experiment

Modeling and experimental research capabilities in EMSL were uniquely combined to synthesize, identify, and characterize 20-atom gold clusters. Experimentalists Xi Li, Hua-Jin Zhai, and Lai-Sheng Wang of Washington State University detailed the synthesis and identification of the Au_{20} clusters, while EMSL researcher Jun Li provided extensive theoretical insight into the electronic structure of the clusters. These researchers found that the cluster could have high structural stability and may have novel optical and catalytic properties. Subsequent computational modeling guided the choice of stabilizing ligands that maintained the structural and electronic properties of Au_{20} but allowed for bulk quantity production of these clusters. Published in *Science* (Volume 299, page 864), their work was also featured on the cover of the *Journal of Physical Chemistry B* (Volume 108, page 12259).



Octane hydrocarbon molecule.

Hydrocarbon Chemistry at Chemical Accuracy

Highly accurate determination of the heats of formation for key alkane components used in models of hydrocarbon fuel combustion demonstrates optimal performance benefits of a balanced computing architecture with a tuned computation code. Led by David Dixon of the University of Alabama and supported by EMSL researchers Bert de Jong and Theresa Windus, researcher Lisa Pollack performed highly accurate *ab initio* correlated coupled cluster with singles, doubles, and approximate triples [CCSD(T)] quantum mechanics calculations on the MSCF supercomputer using an optimally tuned version of NWChem software developed by EMSL researchers. To make maximum use of the computational resources, the (T) triples correction to the CCSD was run on 1,400 processors, where it required 23 hours of elapsed time. Since the majority of this calculation consisted of matrix multiplications, an impressive sustained performance of 75 percent of peak performance, or 6.3 teraflops, was achieved.

Current Science Projects and Use

In the current fiscal year (2005), 15 million central processing unit (CPU) hours have been allocated to 15 Computational Grand Challenges and approximately 30 small one-year MSCF pilot projects that will potentially evolve into a Computational Grand Challenge proposal. In addition, five percent of the resources were allocated for use by projects assigned by the DOE Office of Science. Approximately 13 million CPU

hours, or 87 percent of the total allocation, were assigned to the 15 Computational Grand Challenge projects, each averaging about 865,000 hours.

The 15 Computational Grand Challenges currently running on MSCF high-performance computing resources cover a wide range of environmental scientific directions important to DOE and the nation. These projects, loosely grouped by science area, show the breadth of science currently being performed at the MSCF.

Biology:

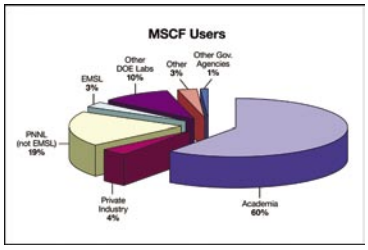
- ***Multiscale Modeling of Biochip Systems***
- ***Complex Enzymatic Reactions Studied by Molecular Modeling and Electronic Structure Calculations***
- ***Image Processing, Modeling, and Simulation of Complex Biological Systems Using Volume Filling and Boundary Fitted Mesh Based Methods***
- ***Structure and Recognition in Microbial Membranes, Proteins, and DNA***
- ***Bioinformatics Tools to Define the Proteomic State of the Cell***

Chemistry:

- ***Reliable Electronic Structure Prediction of Molecular Properties***
- ***Nanostructure Formation, Aggregation, and Reactivity***
- ***Direct Dynamics Simulations: From Molecules to Macromolecules and Condensed Phases***
- ***Computational Design of Catalysts: The Control of Chemical Transformation***
- ***Molecular Computational Studies in Environmental Chemistry, Geochemistry, and Biogeochemistry***
- ***Computational Design of Materials for Hydrogen Storage***
- ***Reliable Relativistic Quantum Chemistry Calculations for Molecules with Heavy Elements***
- ***New Theoretical Developments and Computational Studies of Complex Processes in Environmental Chemistry, Waste Containment, and Biochemistry***

Environmental Systems:

- ***Superparameterization: A New Paradigm for Climate Modeling***
- ***Multifluid Flow and Multicomponent Reactive Transport in Heterogeneous Subsurface Systems***



2004 MSCF user affiliation.

The MSCF high-performance computing resources have supported the research of approximately 420 users on average for the last 2 years, or about 30 percent of the entire EMSL user community. Fifty-nine of the one hundred distinguished EMSL users, as determined by the essential science indicators and endowed chairs, are associated with the MSCF. Researchers external to PNNL make up 78 percent of the user community, while the remaining 22 percent is comprised of PNNL and EMSL staff, postdoctoral fellows, and students. A distribution of users by affiliation shows that 60 percent of the scientists come from universities, whereas 32 percent of the users are from DOE-sponsored laboratories. In addition, EMSL software user agreements have been granted to almost 1,300 sites for use of NWChem and Ecce.

The MPP2 computing system has maintained high availability and utilization, as exemplified by 93 percent availability and 87 percent utilization in the first quarter of fiscal year 2005. In addition, there has been high usage of the storage capability, with approximately 75 terabytes of the more than 85 terabytes of disk space being used to store scientific data.

EMSL Scientific Grand Challenges: Combining Modeling and Experiment

EMSL is recognized within the scientific community for its outstanding and one-of-a-kind instrumentation used for fundamental experimental, theoretical, and computational research in the environmental molecular sciences. Two EMSL Scientific Grand Challenges being implemented are designed to leverage integrated sets of EMSL research capabilities, creating a synergy between computation and experiment, in the study of highly significant scientific problems. These EMSL Scientific Grand Challenges, which will last from three to five years, will create new benchmarks for integrated research. These challenges are aligned with DOE mission areas, are driven by users, and take full advantage of EMSL's unique capabilities, resources, and technical expertise. They are also designed to attract and involve users who are among the best scientists in the world in the research area of the challenge. Each EMSL Scientific Grand Challenge contains significant experimental and computational components that are necessary to accomplish its goals, and the MSCF plays a major role in the two projects already underway. Each of these projects contains science that will require significant interplay between different length and time scales and, therefore, will require new algorithms and computational methods to succeed.

In 2004, EMSL launched Scientific Grand Challenges in biogeochemistry and biology. The EMSL Biogeochemistry Scientific Grand Challenge, "Understanding the Molecular Basis for Electron Transfer at the Microbe-Mineral Interface," was the first to commence and is focused on answering the questions:

- ***How do microorganisms accomplish electron transfer with mineral phases containing polyvalent metal ion centers such as oxides and clays?***

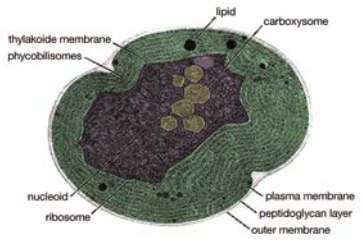
- ***What molecular interactions occur at the interface of the bacterial cell envelope and hydrated mineral surfaces to regulate electron flux to and away from microorganisms?***

Understanding the mechanism by which electrons are transferred from the microbial outer membrane directly to terminal electron acceptors outside the microbial cell is key to realizing the bioremediation potential of the model organism in this study, *Shewanella oneidensis*. Experimental characterization of the proteins and minerals must be combined with theoretical work to provide an essential fundamental understanding of the biochemical and biophysical aspects of metal reduction at microbe-mineral interfaces. Molecular modeling and simulation studies in combination with electron transfer theory calculations will be carried out to characterize the cytochrome-mediated electron transfer process across the outer membrane of *Shewanella oneidensis* and the resulting charge transfer and migration in mineral substrates.

The EMSL Membrane Biology Scientific Grand Challenge, "Elucidating the Fundamental Changes in Key Biological Membrane Processes in Cyanobacteria under Environmental Perturbations," is focused on answering the questions:

- ***What are the consequences of environmental changes on growth and physiology?***
- ***How does the transcriptome respond to environmental changes?***
- ***How does the membrane proteome change and how does such a dynamic pattern relate to that of the transcriptome?***
- ***How does the cellular ultrastructure respond to environmental changes?***
- ***What is the relationship between the plasma and thyllakoid membranes to intracellular membrane systems?***
- ***What are the structures and dynamics of key molecular complexes?***
- ***What are the underlying networks that predict forms and functions of membranes and their components?***

Currently, 500,000 CPU hours have been specifically allocated to the EMSL Grand Challenges with additional time associated with two of the Computational Grand Challenges that are related to these areas of research. It is anticipated that the allocation will grow to one million hours in fiscal year 2006 and to approximately one million hours per project in the following years.



A transmission electron micrograph (false colored) of a dividing cell of Synechocystis 6803 (image from the CYANOBASE website).

In addition to the Scientific Grand Challenges, EMSL has formed Collaborative Access Teams (CATs) that conduct high-impact science to demonstrate the capabilities and expertise of EMSL and maintain EMSL at the forefront of science. CATs provide a mechanism to attract high-impact users in a focused research

environment and build new capabilities for use by the CAT and general users. MSCF interactions with the CATs are described as appropriate in Section 2 of this document. Currently, all CAT usage of the MSCF is captured in one of the existing Computational Grand Challenges and, therefore, requires no extra resources other than those described earlier.

2. MSCF Science Drivers and Their Impact to DOE

Research at EMSL is focused on finding new solutions to environmental problems critical to DOE through a more fundamental understanding of underlying physical, chemical, and biological processes. Specifically, EMSL is committed to significant scientific discovery and progress in the distinctive science signature areas of biogeochemistry and subsurface science, interfacial chemistry and catalysis, structure/dynamics of biomolecules and biomolecular complexes, biochemical pathways, and spectra signatures and trace detection. To address this research, biological and chemical interactions with physical processes are investigated from the subsurface to the atmosphere using the computational resources available in the MSCF. The scientific drivers and direction for this research have been established through BER guidance and broad input and support from the scientific community, and represent the science directions envisioned by the current and future EMSL user community. The scientific direction and resulting general computational resource requirements presented in this document were derived from the EMSL vision and distinctive science signatures, solicited white papers, and a technical workshop on scientific computing held at EMSL in December 2004. In addition, scientific leaders associated with the upcoming EMSL Scientific Grand Challenges were interviewed to assess computational needs for their research.

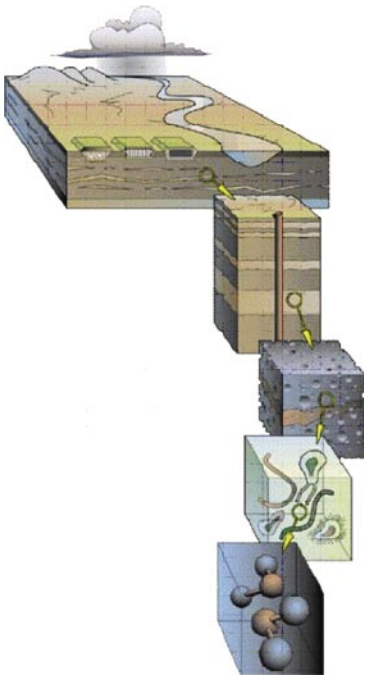
The scientific vision for the coming 3 to 5 years —and the computational resources to achieve that vision—is tantamount to DOE's environmental objective of understanding ***physical, chemical,*** *and* ***biological*** *processes.*

In August 2004, a call for white papers was issued by the MSCF to the current EMSL user community and a broad scientific audience representing potential users and leaders in areas of key interest to BER and DOE. Scientists were asked to share their scientific visions for the next three to five years and to discuss the computational resources that would be required to accomplish that vision, as well as the role MSCF high-end computing resources will play in achieving their future scientific goals. Twenty-five white papers were received (available as part of the Supporting Documents accompanying this document)—84 percent authored by non-EMSL/PNNL staff. After an internal review, authors of papers that showed the clearest scientific vision in appropriate scientific areas were invited to present their paper at an MSCF technical workshop, held December 10 and 11, 2004, at EMSL. Fifty-five scientists and stakeholders from the United States and abroad attended the workshop. The morning of the first day consisted of information and motivating talks (also available as part of the Supporting Documents) to set the tone of the workshop. The remaining portion of the day and the next morning were devoted to discussions of science drivers and computing needs in three parallel breakout sessions: biological sciences, chemical sciences, and environmental systems.

The summaries from the three breakout sessions were then combined with information in the white papers and input from stakeholders, such as BER, to form the basis of this document, which describes the science drivers, the associated

requirements for enhancing high-performance computing resources at the MSCF, and the role that a next-generation MSCF high-performance computing facility will play within three to five years across the scientific community and for DOE.

The following subsections address in more detail the drivers and challenges encountered in the biological, chemical, and environmental systems research areas, their impact on DOE environmental missions, their ties to the EMSL distinctive science signature areas, and the computational needs that will be required to greatly advance the research. A common theme among the research areas is the need to develop models that describe a wide range of length and time scales. Environmental science is directly impacted by all three science areas and exemplifies the need to cross scales.



Environmental science, including linking scales at the field, meso, pore, micro, and molecular scales. Solving problems in this field will involve linking chemical and biological interactions from the subsurface to the atmosphere.

Environmental Science Across Scales

The goal of environmental science is to understand the current state of the environment based on knowledge of the past and then use this information to predict the future state. For example, given a current production process, what is its long-term environmental impact? Given a potential environmental remediation strategy, what will it lead to? Environmental chemical science addresses delicate and heterogeneous systems maintained by a myriad of nonlinearly coupled processes in a self-organized state frequently very far from statistical equilibrium. Moreover, those processes occur over a vast range of time and space scales. The issue of scale dominates environmental science because the science of the whole of what takes place is far richer than the sciences of the pieces. In other words, an understanding of the chemical and biological interactions from the subsurface to the atmosphere requires an understanding of the couplings among the various contributing phenomena. Although scientists are interested in the results at large spatial and temporal scales, detailed insight into behavior at the molecular scale and how it is coupled to the macro scale is key to understanding 1) how humans have impacted the environment, 2) how to remediate anthropogenic impacts on the environment, and 3) how to minimize future anthropogenic impacts on the environment. Computational environmental science, in combination with experimental investigations, provides the fundamental integrated understanding of the complex interactions of materials from anthropogenic sources with the environment. The issues of complexity in terms of the systems to be studied and their interactions in a poorly characterized environment have made it difficult for this area to advance as rapidly as needed. For example, understanding the function of biomolecules in the living organisms envisioned for bioremediation of contaminated DOE nuclear production sites requires knowledge of its basic chemical processes and the biological processes and networks.

Models of contaminant fate and transport in the subsurface are built on detailed knowledge of the binding and chemical reaction of contaminants on soil particles as well as transport and reaction in groundwater. Reliable models of the direct impact of mobile contaminants on humans and the risk of proposed remediation technologies will be critical for developing the safest and most cost-effective approaches to site cleanup and for public acceptance. Such a research effort needs to span the interfacial regime from bare solid surfaces to complex, solution-phase

surface chemistry, covering a range of time and length scales. An understanding of the linkages between different temporal and spatial scales is an important focus of the overall computational science effort. No matter how detailed the physics in a reactive-transport model, if the critical underlying physical, chemical, and biological data are missing or unreliable, the accurate predictive capability of a model is decreased. High-quality data are needed, and great care must be taken to minimize the errors in the calculated data used in a sophisticated environmental or chemical process model so that errors do not accumulate, propagate, and ultimately invalidate the macroscopic-scale model.

Likewise, understanding the radiation processes in the atmosphere is very dependent on understanding the formation of aerosols and clouds. Chemical and physical processes of varying size and time scales are crucial to developing atmospheric models that reduce the error between the observed and computed temperatures, as well as other observable conditions, to ensure accurate, as well as trusted, future predictions. These in turn will enable policy makers and scientists to determine safe levels of greenhouse gases and ensure a viable living environment for future generations.

"We cannot imagine the kinds of problems that the supercomputers of tomorrow will be able to solve, but we can imagine the kind of problems we will have if we fail to provide researchers in the U.S. with the best computing resources."

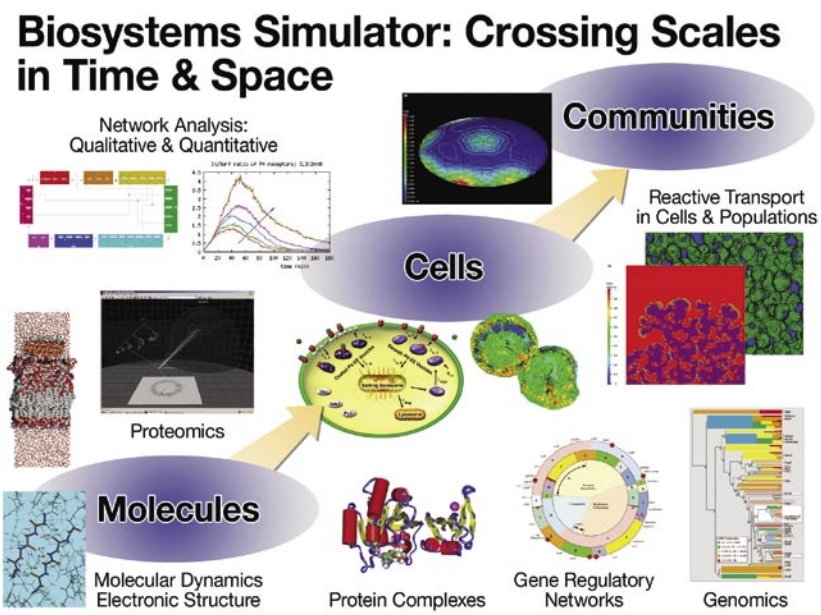
Judy Biggert (R-IL)
Chairman of the Science Subcommittee on Energy

Simulation: Linking Scales

As the preceding section suggests, a holistic approach that can couple phenomena across scales from the very small (molecular) to the large (ecosystems) is critical to progress in environmental molecular science. Along with theory and experiment, simulation is a critical element of a successful approach to systems science. Simulation can be used where a real experiment is too costly or dangerous to perform. It can act as a time machine to slow down processes that may be too fast to study in sufficient detail (chemical reactions) or speed up processes that are too slow to produce results in our lifetimes (climate change). Simulation is frequently the only method available to interpret complex theories in a manner that can be compared to experiment. The understanding gained by such comparisons drives the development of both theory and experiment.

Biological Sciences

Advances in biology, from a detailed understanding of molecular machines built of proteins to the workings of complex biological systems, are playing an important role in the discovery of groundbreaking solutions to DOE's challenges in energy, security, and the environment. The combination of the many facets of multiscale computational biology will lead to an intricate understanding of biological systems and the ability to control and modify the machinery of nature. Understanding how organisms respond to and interact with their environments could lead to the development of modified organisms that would assist DOE with its mission areas



The many facets of multiscale systems biology, which, combined, will provide the critical scientific knowledge base needed to make significant advances toward DOE missions in energy, security, and the environment.

of bioremediation and environmental cleanup, carbon sequestration and understanding the natural carbon cycle, and biological alternatives for fuel and energy production.

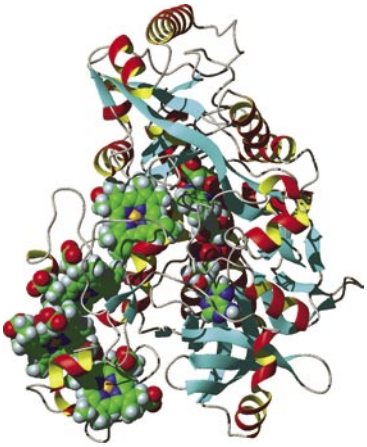
This section discusses scientific areas within multiscale computational biology that are important to DOE's environmental mission and are aligned with EMSL's distinctive science signatures of biogeochemistry and subsurface science, structure/dynamics of biomolecules and biomolecular complexes, and biochemical pathways. These are the areas where the computational capabilities of the MSCF will significantly contribute to increasing the understanding of the complex world of biology. The main focus is on the molecular modeling and simulation of biological and biochemical processes, although other scientific areas in systems biology are discussed where high-performance computational resources could make a significant impact.

Molecular Modeling of Basic Biochemical Processes

Biochemical molecular modeling simulations help researchers predict and understand the mechanics of proteins and membranes, the behavior of complexes of proteins, and the interaction of membrane proteins with intracellular proteins and external minerals or molecules in order to provide the fundamental knowledge of the basic biological processes at the smallest length scale of computational biology. Identifying, characterizing, and understanding these molecular machines of life comprise one of the goals outlined in DOE's Genomics:Genomes to Life Program (http://doegenomestolife.org). This area of computational biology has strong ties to the area of computational chemical sciences, and some of the approximate methodologies developed and used in the chemical sciences are used to study biological systems. The use of fully quantum mechanical dynamics methods on large biomolecular systems consisting of thousands of atoms—a brute force attempt to achieve accurate results—would require computational resources far beyond any currently conceivable computing environment, with the possible exception of a quantum computer. Instead, biochemical molecular modeling simulations are relying on classical molecular dynamics and Monte Carlo methods, which are based on an atomistic representation of the biomolecular system with a force field defining and driving the atomic motions over time.

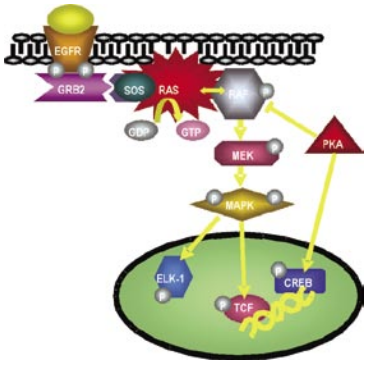
In biological systems, host-guest interactions are of primary importance to control the tertiary structure of biomolecules, determine the rugged energy landscapes involved in protein folding, influence the function of proteins by regulating access to active sites, as well as playing a critical role in controlling the reactions in the active site, often by properly positioning the reacting species to effectively control both the entropy and enthalpy of the reaction process. Host-guest interactions also play an important role in the chemical sciences, such as the uptake of metal ions by host ligands (this is described in more detail in the Chemical Sciences section of this document). Knowledge of these host-guest interactions is crucial to understanding the function of biomolecules in living organisms envisioned for bioremediation of contaminated DOE nuclear production sites. These often weak interactions are critical to the interactions that occur at the membrane surface of a cell in terms of interactions with a mineral surface or at a receptor site where a ligand binds to initiate cell signaling processes. These types of interactions are central to separations systems, in which the competition between ion-solvent, ion-ligand, and ligand-solvent interactions controls the selectivity and efficiency of separations systems used to extract specific species from mixed wastes. Finally, these types of interactions are important in sensors being developed to detect contaminants and toxic materials.

Cells harvest energy through mechanisms such as photosynthesis, respiration, and the fermentative degradation of complex metabolites to simpler products. In the respiration process, energy is extracted from the transfer of electrons through a series of protein complexes, the electron transport chain, to a final electron acceptor. In aerobic respiration, the final electron acceptor of the electron transport chain is oxygen, whereas the final electron acceptors of anaerobic respiration are usually nitrate- or sulfur-containing compounds. However, as an example, *Shewanella* species have been shown to use a much broader range of electron acceptors, including fumarate, nitrate, nitrite, trimethylamine n-oxide, dimethyl sulfoxide, sulfite, thiosulfate, elemental sulfur, Fe(III), Mn(IV), and uranium, thereby affecting the solubility of these substrates in the environment (this has direct impact on the EMSL Biogeochemistry Grand Challenge and the EMSL biogeochemistry and subsurface science distinctive science signature). Due to the classical nature of the molecular dynamics and Monte Carlo methods, processes such as chemical reactions or electron transfer that require the explicit presence of electrons cannot be studied, nor can bonds be easily made or broken during a molecular dynamics simulation. Drawing on developments in the last decade in the chemical sciences, computational methods that combine quantum mechanical techniques with molecular mechanics/dynamics have been developed and used with some success. Each implementation of such quantum mechanics/molecular mechanics approaches has its own set of limitations. In general, the larger the quantum mechanical region, the more accurate the simulation. The simulation of an enzymatic catalytic reaction in a biological system that contains proteins, membranes, and water requires that a very large section of the biomolecular system be included in the quantum mechanics region. For example, modeling the ion channel of a so-called G7 protein in a membrane would require the whole channel, adjacent amino acid residues, and contents to be included in the quantum mechanics region. This could easily add



Modeling the cytochrome electron transfer in the flavocytochrome c_3 fumarate reductase of Shewanella frigidimarina requires quantum mechanical methods. The ferric and ferrous hemes in the protein can be modeled quantum mechanically, whereas the majority of the protein is accounted for with molecular mechanics (image courtesy of Straatsma, PNNL).

up to a thousand atoms or more, which is beyond the limit of currently available teraflop-scale computing resources. New algorithms that describe the interaction among proteins and metal ions and protons need to be expanded before simulating such systems.



Modeling signaling pathways, Ras acts as the GTP-activated molecular switch in the cell signaling pathway for gene expression, controlling cell proliferation and differentiation (image courtesy of Straatsma, Resat, Miller, Soares, and Dixon, PNNL).

At both the spatial and temporal scale, the accuracy of biochemical simulations is limited by available computing power. For example, current teraflop-scale computing resources are able to simulate basic processes in proteins and membranes for periods of tens of nanoseconds. Even rapid biological processes can take place at time scales that are three to five orders of magnitude longer. Simulations at the microsecond time scale are needed to model the interaction between proteins, which is crucial to obtaining a fundamental understanding of cell signaling processes. To get an idea of the length of these simulations, consider that a molecular dynamics simulation of ten nanoseconds currently requires five days. Looking forward, this means it would require approximately 28 years to complete a ten microsecond simulation on current computers. Obviously, not only will additional compute resources be required, but improvements in long time scale theory and algorithms will be required for microsecond simulations. This microsecond time scale is also needed to simulate the interaction of membrane proteins with cell proteins and with external influences such as mineral surfaces and molecules. Understanding how membrane proteins respond and interact with minerals will lead to insights and developments that can contribute to DOE's mission areas of bioremediation and environmental cleanup and EMSL's signature area of biogeochemistry and subsurface science. Fundamental research into interfaces between hard (inorganic) and soft (organic) matter also have implications in the areas of biosensors (another EMSL distinctive science signature) and even "wet" computing device design. Currently, the complexity required for an understanding in such areas means that computing power access must grow faster than Moore's law for this community.

Simulations at the millisecond time scale are needed to model slow conformational changes that determine the biological function in large biomolecular systems. Biological processes in membranes, such as the transport of simple ions and small molecules through the membrane, and the formation of vesicles and membrane fusion also occur at this time scale. The modeling of these processes provides insight in the way microbes acquire materials (nutrients) from their environment. Millisecond molecular modeling simulations also are used to predict and understand protein folding and unfolding, as discussed below.

The computational requirements for molecular dynamics simulations are well defined, and many of these simulation methods already make use of teraflop-scale computers. In massively parallel architectures, simulations depend on the speed of the processor, but their scaling characteristics are strongly affected by interprocessor communication latency and speed (interconnect) relative to the processor speed. The general approach is to distribute the atoms in the biomolecular system of interest over the processors. However, at each of the large number of time steps, information about (possibly all) other atoms in the system is needed, and this information can only be obtained via the interconnect. With the great

advancement of processor speed relative to the interconnect, communication, especially latency, is becoming a bottleneck in large molecular dynamics simulations. For example, even the low 3- to 5-microsecond latency of the current Elan4 interconnect on the MSCF machine is a bottleneck for large molecular dynamics simulations on large processor counts, requiring further improvements to be made in algorithms for prefetching and overlap of communication and computation to minimize time to solution.

An equally important computational requirement is the ability to efficiently store and handle multiterabytes of data. Even today, long molecular dynamics or Monte Carlo simulations that are run on massively parallel computers create as-yet unresolved data management issues from the large amounts of data generated. For example, a single time step in the simulation of a protein consisting of 50,000 residues or, with each residue containing about 20 atoms, a million atoms easily manipulates tens of megabytes of data. The current strategy for a nanosecond simulation of small biomolecular systems of about 50,000 atoms is to store trajectory data for every 10th or 50th time step for post-processing analysis. One such simulation can easily generate single data files of 200 gigabytes. Larger 100-nanosecond simulations of a biomolecular system of ten million atoms could potentially generate 4,000 terabytes of data based on current state-of-the-art methods. From this example, it is clear that tools and methodologies that enable data analysis "on the fly" must be developed and refined. The question of reaching macroscopic time scales from molecular dynamics simulations cannot be solved solely by increases in hardware capacity, since there are fundamental limitations on how many time steps can be executed per second on a computer, whether parallel or serial. To address this severe challenge, new, theoretically sound, time-coarsening, and multiscale methodologies must be developed that will permit dynamics-based methods to traverse longer time scales.

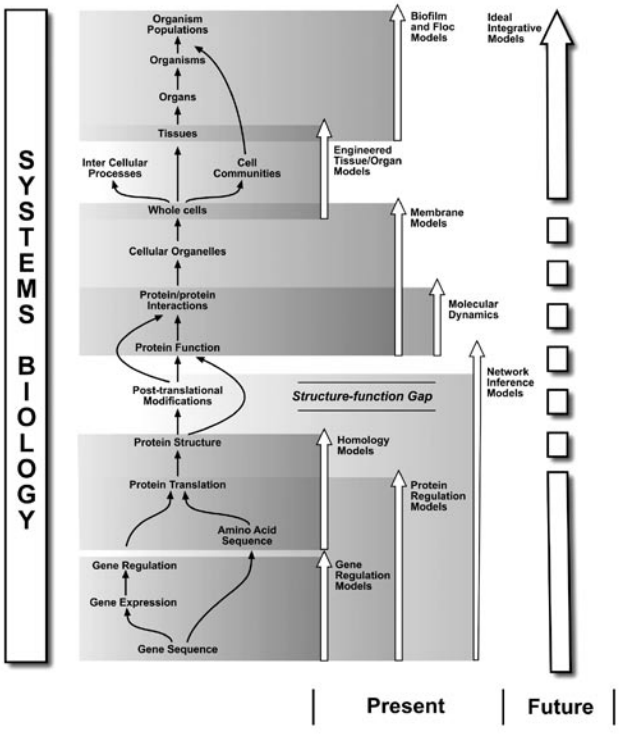
The force fields that lie at the heart of molecular dynamics and Monte Carlo methods generally are empirical in nature and range from single-component bonded and non-bonded atomic interactions to multicomponent-bonded and higher-order, non-bonded interactions (e.g., polarizable atomic charges). These methods have improved steadily over the past 20 years, but it is generally acknowledged by developers and users of these force fields that much improvement is still needed. Especially for biological systems, the difficulty of finding minima on complex conformational or reaction surfaces is not a solved problem and has direct impact on the force field development. A need exists to develop an improved set of tools that are far more efficient for dealing with atomic motion within the energy and time domains.

Modeling of Protein Folding for Structure Prediction

Large arrays of proteins and multiprotein molecular machines are the workhorses of all living cells. The function of each protein is defined in part by its three-dimensional folded structure in the cell, which is often unknown. It is crucial to obtain a detailed knowledge of the protein structure before one can start to simulate

the biochemical processes described above and be able to determine and understand the function of a protein, its role in cells, regulatory complexes, and interaction with the environment. Hence, determining the structures of proteins is a high-priority task for all biological sciences in the near future.

Proteins are macromolecules with a backbone structure and a unique side chain sequence which result from the translation of gene products. They can vary in size from a few of the 20 standard amino acids to tens of thousands or more. The chemistry and electrostatic behavior of these side chains defines the functionality of the protein and enable proteins to perform almost all the specialized tasks required by cells, from forming the cytoskeletal support to providing transfer pathways across the membrane and transferring energy, signals, and materials. Assuming one can accurately define the open reading frames from gene sequence data, the primary (linear) structure or amino acid sequence of an unfolded protein can be determined. The next step, predicting the fully folded three-dimensional structure of a protein from the sequence of amino acids, is essential to obtaining a fundamental understanding of its function. Classically, sequence homology has been used to assign proteins of unknown function to a class or family of proteins with known function. Sequence homology calculations rely almost exclusively on primary structure information to make these assignments. However, some proteins with poor sequence homology have been shown to have nearly identical functions. It is here where molecular modeling of the processes involved in three-dimensional folding will play an important role in the EMSL's structure/dynamics of biomolecules and biomolecular complexes distinctive science signature. Because of the sheer size of these proteins, which contain hundreds of thousands or even millions of atoms, and the long time scales (milliseconds), next-generation computing resources reaching towards petascales and novel algorithms will be required.

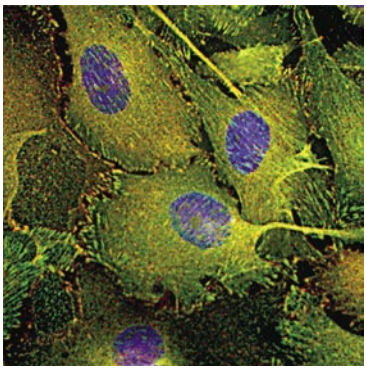


Integrative multiscale modeling combines the many facets of computational biology, systems biology, and bioinformatics.

Multiscale Modeling of Biological Processes

Twenty-first century biology will be dominated by a convergence of research in the biological, information, and physical sciences. Currently, biological science is undergoing a major revolution—driven by technological breakthroughs such as microarrays and other high-throughput data production capabilities that enable global gene expression experiments which provide snapshots of gene activities in living cells. Thus, modern biology is moving from a reductionist approach, where the behavior of the individual pieces, such as genes or proteins, is understood, to one where understanding of the whole system behavior is paramount. Biology is being transformed into a systems science analogous to

the study of systems in physical sciences, where "… biology will surely be presented to future generations of students as a set of basic systems that have been duplicated and adapted to a very wide range of cellular and organismic functions following basic evolutionary principles constrained by Earth's geological history"[1]. Currently, the biological science field is far from fully representing information that can describe and predict cellular behavior in general. However, the completion of multiple genome sequences sets the stage for a complete understanding of the biomolecular machinery of life. Integrative multiscale biological modeling attempts to understand structures and functions at a systems level based on knowledge acquired from genomics, proteomics, metabolomics, and other high-throughput technologies. This clearly represents a true scientific grand challenge as it requires fundamental changes in the traditional reductionist approach to biological studies.



Understanding complex cell systems and their functions is based on measuring proteins within an organism. Specialized instruments identify and quantify proteins expressed under specific environmental conditions.

The future of biology will be driven by the fundamental paradigm shift from today's hypothesis-driven research to data-driven discovery research, employing the massive amounts of available biological data. The enabling technology in this paradigm shift is the emerging ability to integrate through many levels of conventional knowledge by formalizing the *interface* of processes from the molecular level to the organism to a population. This will require substantial computing power, a large storage infrastructure, the development of new data reduction and analysis algorithms, implementation of these algorithms in efficient and easy-to-use software, visualization tools, and development of standards for data representation and quality control. The end result will be the ability of a researcher to use all of the available information to analyze a given system at a range of scales. This includes information obtained from computational chemistry simulations, systems biology, regulatory network modeling, cells and cell network modeling, tissue and organ models for more complex organisms, cell signaling models, and population interaction models. These different levels of information must be linked so that changes at one level can propagate both upward and downward in the model, from protein interactions in complexes to transformations in the ecosphere.

An ultimate goal of multiscale biology is to establish a virtual cell computational environment that will enable modeling and simulation of the complex phenomena occurring in cells. These models would be linked to key biological experiments, used to analyze the data and, ultimately, drive the design of new experiments.

Proteomics and Genomics: Building Blocks for Multiscale Models

Detailed analysis of genome and proteome sequence data and characterization of the expressed proteins are the foundation on which much of biological modeling, simulation, and multiscale systems biology research are built and play a significant role in the EMSL Membrane Biology Grand Challenge.

1 Smith TF. 2002. "The Challenges Facing Genomic Informatics," in *Current Topics in Computational Molecular Biology*, Chapter 1. Eds. T Jiang, Y Xu and Michael Zhang, MIT Press. Cambridge, MA.

The amount of genome and proteome data is growing at an exponential rate, providing researchers with a rich set of resources for analyzing protein expression. In tandem with the increased size of the information base, scientists are beginning to ask new questions that cannot be answered without the use of efficient, massively parallel, sequence alignment tools on high-performance computing resources. Of particular interest is the ability to classify protein families or superfamilies for annotation. Creating a protein family designation requires the comparison of whole proteomes with each other. Genome-to-genome or proteome-to-proteome comparisons will require high-performance computing due to the sheer size of the data set. Because of the importance of these analyses to systems biology and biomolecular modeling, high-performance memory and data storage-driven genome and proteome sequence analysis capacity is a critical enabling technology for the future of biology, by reducing time-to-solution from weeks or months to minutes. This application of high-performance computing to sequence data is only the first step. To turn this wealth of knowledge into information of practical value, it must be integrated into larger modeling frameworks. Genome and proteome data are thus one of the fundamental building blocks on which integrated multiscale modeling in biology will be built. This low-level data forms the basis for further exploration into molecular, cellular, and higher-level interactions among biosystems.

A second tier of knowledge, important to EMSL's biological and environmental biomarkers signatures, on which multiscale biological methodologies will be built results from high-throughput and other peptide identification technology. Because of its invaluable place in bridging the gap between gene expression data and protein translation data, high-throughput proteomics, such as that performed at EMSL's signature proteomics pipeline user facility, will be an indispensable tool for biologists. This facility generates massive amounts of data daily that should be analyzed quickly but is currently stored and processed offline. Efforts are underway to improve the peptide identification process using mass spectrometry data by fusing the peptide identification process to the spectral acquisition. Real-time analysis, using data and memory-intensive methods such as hypothetical "fingerprints" to assign peptides to spectra or statistics, would improve the experimental throughput of useful data to researchers. The development of these analysis methods has benefited significantly—and will continue to benefit—from the MSCF computational resources. This application of high-performance computing to peptide identification is not in itself an end. Rather, it should be viewed as a second layer of data from which the identity and possibly the relationship between proteins in biosystems under varying conditions can be surmised. These data motivate detailed exploration into the integration of protein identity, function, protein complex interactions, and gene regulation with higher-level cell and cell community processes.

Putting Gene and Protein Data to Work: Multiscale Integrative Models

Given the emerging wealth of gene and protein data, biologists must now focus on harnessing this information for practical use. Many science areas in biology of relevance to DOE are emerging with questions requiring integrative multiscale

modeling. These include: How do we detect agents of bioterrorism and prevent their use against a population? What is the relationship between the genetic sequence of a given organism and its ability to remediate heavy metals? What is the environmental and biological impact of emerging energy sources? These key scientific questions clearly align with many of DOE's missions in energy production, environmental remediation, and global climate change mitigation and with EMSL distinctive science signatures. Tackling these important scientific questions requires modeling very complex biological systems over a wide range of length and time scales. This range runs from modeling molecules, proteins, protein-protein interactions, and signaling pathways, to subsystems such as membranes, cells, and the relationship and interactions among different cells or organisms and their environments. For example, the cycle whereby the cell impacts its environment and the environment impacts the cell can be extremely complex, having many dynamic interactions that occur from the nanometer scale to the meter scale and beyond.

Biology has historically resorted to a reductionist approach, wherein each field obtains a specialized understanding and knowledge base for certain subpieces of a biological system of interest. As the knowledge bases of each of these facets in biology continue to focus on subpieces of subpieces of systems, it becomes more and more difficult to make generalizations about the interactions among the pieces of the larger biological systems. To fully understand the biological machinery of a cell and its interactions with other cells and its environment, a fundamental paradigm shift in biology is needed in which the knowledge base of the many facets of biology will be combined into a new holistic computational modeling approach. Understanding how various components of a cell interact with each other or how proteins interact with each other to perform a biological function, will require the development of very complex and heterogeneous simulations of complex biological systems. Moreover, this will require full integration of models at various length and time scales with dynamic and nonlinear feedback mechanisms, simultaneously coupling them with large experimental and simulation databases. These computational models are not yet available, but in many cases, the constituent subsystem models have been well studied and validated. Fully integrated multiscale models using those well-defined pieces are expected to be developed in the next few years as increased high-performance computing capabilities (possibly into the petascale) and mathematical formalisms for integrating the individual model pieces become available.

The absolute necessity for this paradigm shift is easy to imagine, even in the context of simple problems. For instance, if one desires to harness the power of *Shewanella* (a known heavy metal reducer) to assist in environmental cleanup of nuclear waste sites, it is critical to understand all the diverse mechanisms of this organism's ability to remediate the material. But if the technology is to be of practical use, great confidence is needed in the ability of this organism to function in a wide range of environmental conditions, and in the presence of many other organisms, any of which might interfere with the desired behavior of this organism. This type of knowledge requires a much broader understanding

Dissimilatory metal-reducing *Shewanella* on hematite (Fe_2O_3).

of *Shewanella*, including a holistic view of how its protein expression will change in, for example, varying climates and chemical environments. Since environment impacts the organism, and vice versa, there is a dynamic and nonlinear feedback system at work that must be understood to fully benefit from this organism. The MSCF has the opportunity, and even the responsibility, to play and maintain a leadership role in understanding and designing life forms tailored to bioremediation. To effectively maintain leadership and rise to such grand challenges, a tight coupling between algorithm development and computational/facility upgrades will be required.

High-Performance Computing Requirements

Of all the computational biology science areas discussed in this section, only the high-performance computing requirements for molecular dynamics simulations have been well characterized. Molecular dynamics simulation methods already make use of teraflop-scale computers and will strongly benefit from larger-scale high-performance computing resources as previously discussed. With the increase in processor flop count outpacing the speedup in processor networks, the interprocessor communication or interconnect has become the cause of the scaling bottleneck of these simulations on massively parallel architectures. Currently, the general approach for distributing the atoms of the molecular system over the available processors requires that information be interchanged between those processors at each of the large number of time steps of the simulation.

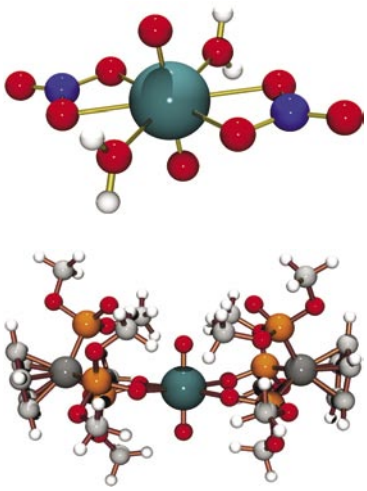
The high-performance computing and architectural needs of the other computational biology science areas are still being defined. Many of these areas have only begun to define and develop simulation tools to answer their scientific questions and have not reached the point where they can address high-performance computing resource and architectural needs. The consensus is that multiscale modeling will be operation-intensive (floating point and, in some cases, integer) as well as demanding of memory size and bandwidth, and disk storage capacity and input/output (I/O) bandwidth, but no quantification is possible now. However, the requirements for these research areas are not expected to exceed the extensive computational resource needs in molecular dynamics and the chemical and environmental systems sciences (such as processors with high flop counts and large memory and data-storage requirements).

Chemical Sciences

Chemistry Research Challenges

Chemistry, the science of molecules, is also the science of the everyday world. For the world at large, molecules are the fundamental units of matter; smaller than this, matter loses its real-world identity and, hence, its defining characteristics. An understanding of the structure, interactions, and reactions of molecules is of critical importance to a wide range of phenomena, from the fate of contaminants in the

environment, to the production of plastics from crude oil, to the occurrence and treatment of genetic diseases. Chemistry is a key science in a world attempting to balance energy use, environmental quality, human health, and economic prosperity. Advances in the chemical sciences—such as understanding chemical reactions in solution, assembling unique nanostructured materials that can be used to sequester contaminants, and predicting biological activity at the molecular level of complex biomolecules such as proteins, DNA, and membrane structures—are playing a central role in DOE's mission to advance the national, economic, energy, and environmental security of the United States as well as its role in cleaning up past production practices in the nuclear cycle.



Modeling structure and energetics of actinide complexes in the environment, such as uranyl-nitrate-water (above, top) and with complexing agents, such as uranyl-Klaui complex (above, bottom) contribute to DOE's environmental cleanup mission (images courtesy of de Jong, EMSL).

During the past two decades, computing has revolutionized the way in which science is practiced, facilitated by advances in computer hardware and software as well as new mathematical and theoretical approaches. As in all areas of science, the chemical sciences currently are undergoing a paradigm shift in which computation is viewed as the third branch of science, along with the well-established theory and experiment branches. Computation can be used to solve complex chemical problems for which current experimental technologies may prove too expensive or dangerous and allows scientists to explore temporal and/or spatial domains that are not accessible by present experimental methods. Some examples of chemistry research areas important to DOE's and EMSL's environmental mission are discussed below; computational chemistry is significantly contributing to these research areas with the promise—pending continued investments in computational capabilities—to continue contributing to important discoveries and scientific tools that transform the nation's understanding of energy and matter.

Computational Environmental Molecular Science

Computational environmental molecular science is key to addressing the complex environmental cleanup problems facing DOE's nuclear production sites, as well as other polluted sites in the nation, since it provides critical information about molecular properties. Four decades of nuclear weapons production at DOE facilities have resulted in the interim storage of millions of gallons of highly radioactive wastes in hundreds of underground tanks, extensive contamination of the soil and groundwater at thousands of locations, and hundreds of buildings that must be decontaminated and decommissioned. The single most challenging environmental issue confronting DOE, and perhaps the nation, is the safe and cost-effective management and remediation of these wastes. DOE invests approximately $6 billion a year in environmental cleanup activities. The sole use of conventional approaches to remediation and control (e.g., excavation, treatment, recovery, and disposal of residual waste for contaminated soils, or pumping and treating contaminated aquifers) is cost prohibitive. Remediation strategies for DOE sites require a combination of conventional remediation approaches with the increasing effectiveness and decreased cost of emerging technologies—such as *in-situ* techniques like natural bioremediation—to meet future remediation goals within budget constraints. Incorporating *in-situ* remediation technologies into

overall remediation strategies poses significant cost savings by leveraging physical, chemical, and biological subsurface processes to enhance the natural recovery of vadose zone and groundwater systems. Obtaining fundamental knowledge about the chemical properties and interactions of the wastes with the environment is a key and cost-effective ingredient to the success of advanced remediation.

Modeling contaminant fate and transport in the subsurface, a component of the EMSL biogeochemistry and subsurface science signature, requires detailed knowledge of the binding and reaction of contaminants on soil particles as well as transport and reaction in groundwater. Computational environmental molecular science can provide the underlying thermodynamic, kinetic, and structural properties data needed for such models, and can provide data that are difficult—or at times even impossible—to obtain in the laboratory or field due to the cost, time constraints, or danger of the experiment. The reliable calculation of molecular interactions of chemicals, including those containing radioactive elements, with environmental matrices such as soils is incredibly complex and will require tens to hundreds of sustained teraflops to perform. High-quality data are needed, and great care must be taken to minimize the errors in the calculated data used in a sophisticated environmental or chemical process model so that errors in the data do not accumulate, propagate, and ultimately invalidate the macroscopic-scale model. The requirement of accuracy means that we must be able to predict thermodynamic quantities, such as heats of formation (ΔH_f), to better than 1 kcal/mol and activation energies to within a few tenths of a kcal/mol—a daunting computational task for the size and complexity of the systems under consideration. Computational molecular science also can be used to aid in the design of new processes, such as new compounds for efficient separation of radionuclides (e.g., technetium) or the rational design of enzymes to enhance the biodegradation of organic wastes or the immobilization of radionuclides.

Over 30 million gallons of high-level radioactive waste is stored in tanks at the Hanford Site in the state of Washington.

Some of the most hazardous materials in the underground tanks, soils and groundwater, and affected buildings contain radioactive isotopes of the actinides and lanthanides, making fundamental knowledge of actinide and lanthanide chemistry a requirement. Because of the difficulty and expense involved in conducting experiments with radioactive materials, computational methodologies can be used to provide a safe and alternative research approach. Further, such methodologies must include relativistic effects for accurate calculations in order to guide the choice of experiments, reliably extend the available experimental data into all of the regimes of interest, and minimize the need for experimental work on radioactive materials. Obtaining accurate calculations of heavy elements is very difficult and requires the use of large basis sets, treatment of relativistic effects including spin-orbit, treatment of correlation effects for d and f orbitals, and the presence of many states. Truly reliable predictions require the combination of the theory of relativity with quantum mechanics and will require 100 to 1000 times the current performance.

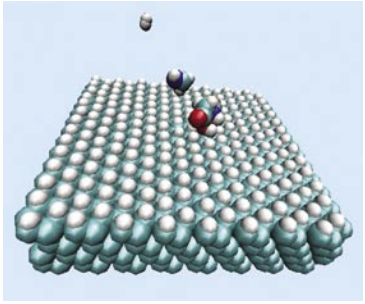
Addressing waste tank and nuclear waste disposal issues requires a wide range of predictive capabilities, including thermodynamics and kinetics for many elements

in the Periodic Table. Such tools will enable reliable predictions about processes occurring in the tanks to facilitate an understanding for optimizing separation processes and long-term storage of the wastes—especially for systems at high pH or high ionic strength and with different organic complexants or competing metal ions. Computation in these cases is critical to developing speciation models, interpreting experimental results, and developing thermodynamic models of complex mixtures.

To facilitate research and remediation efforts addressed in this section, improvements in the MSCF computational hardware and software are needed for 1) accurate thermodynamic and kinetic simulations for idealized models that can be used to benchmark and scale less accurate methods, 2) simulations of molecular interactions with environmental materials involving large molecular systems and different scales, and 3) simulations of actinides and lanthanide complexes and interactions, including proper relativistic treatments. These computationally intensive methods require access to large quantities of processors for significant lengths of time—but will largely impact remediation efforts at various DOE sites, such as the Hanford Site, in part by reducing design and development time for new remediation strategies. Ultimately, cross-disciplinary teams such as those in the EMSL Grand Challenges involving computation and experiment will be necessary to solve these difficult challenges.

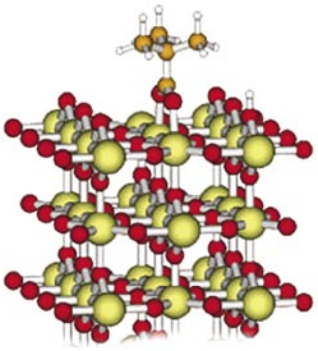
Chemical Transformations

A detailed understanding of the bond-breaking and/or -making processes that will allow for control of chemical reactions will impact a number of areas important to DOE's mission, ranging from catalysis (including processes for disposing of tank wastes) to acid-base chemistry and redox processes important in the tanks and subsurface as well as in biochemical systems to electron transfer processes in biochemical systems for radiation biology, bioremediation, and energy production systems.



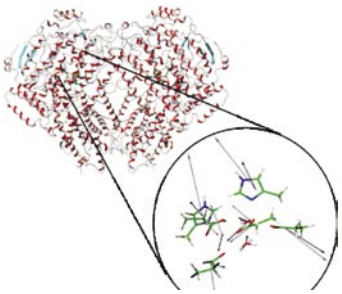
Understanding chemical transformation of molecules on surfaces provides insight into catalytic processes.

Electron transfer reactions and redox chemistry play important roles in addressing a number of challenges important to DOE. For example, photocatalysis on titanium dioxide surfaces can be used for the destruction of contaminants at DOE production sites, splitting of water, and mineralization of harmful organic compounds in polluted air and wastewaters. Understanding the fundamental chemistry of the electron transfer process is critical because the solubility of a given substrate in the environment affects the solubility of other substrates, which is the key issue associated with the EMSL Biogeochemistry Grand Challenge. In addition, redox chemistry is important in the containment and cleanup of nuclear wastes at DOE production sites; the ability to predict and control oxidation states of actinide elements, which control their solubility, is critical for disposal issues at DOE production sites—from performance assessment calculations of the safety of nuclear waste repositories to the remediation of subsurface contamination. Redox chemistry also is critical to understanding the flow and transport of contaminants in the subsurface, as well as in the production of fuel cells.



Researchers at PNNL use state-of-the-art computational methods in chemistry, solid-state physics, and materials science to interpret experimental results and design new materials with favorable catalytic properties.

Reactions of radical species are central in many processes important to DOE, such as atmospheric chemistry, photochemical degradation, plasma processes, radiation biology, reactions in tanks, and combustion of fossil fuels (currently the nation's primary energy source). For example, the major reactive species in the troposphere is OH radical. Its reactions with by-products of energy use, such as hydrocarbons and SO_2, result in atmospheric pollutants such as ozone, secondary organic aerosols, and acid rain. Understanding the impacts of other compounds generated by energy production and use requires a detailed knowledge of the reaction mechanisms in the atmosphere. These types of reactions are critical to understanding research in the Atmospheric Chemistry CAT and EMSL's aerosol chemistry distinctive science signature. Radical reactions also are important in condensed-phase systems, where the reactions typically are created by ionizing radiation. Understanding the factors controlling radical reactions in condensed phases is essential for mitigating corrosion of nuclear reactors, hydrogen gas generation in nuclear wastes, and biological effects of radiation. A detailed understanding of the complex chemical mechanisms enables the design of cleaner, more efficient combustion processes for energy use.



Researchers at Columbia University use the MSCF to study the efficient oxidation of alkanes by oxygen with methane monooxygenase as a catalyst. Design of catalysts that mimic this biomolecule would be of great interest in the fuel industry (image courtesy of Gherman and Friesner, Columbia University).

Catalysts modify the rates of elementary chemical reaction steps without being consumed, thereby making it possible to control chemical transformations, use energy efficiently, and maximize the yields of desirable products as well as prevent the formation of waste streams that pose environmental impacts. Catalytic processes, one of EMSL's science signatures, are essential to energy production and conservation, which are central to DOE's mission. According to the Basic Energy Sciences Advisory Committee report[2], "The Grand Challenge for catalysis science in the 21st century is to understand how to design catalyst structures to control catalytic activity and selectivity." Computational methods hold the key to providing the fundamental understanding of catalytic processes, thus enabling true first-principles catalyst design. True *ab initio* catalyst *design* will require quantitative information about *transition states* for critical reaction processes in catalysis—which are *only* accessible today using computational methods. The marriage of theory and experiment will lead to quantitative design principles and methodologies, enabling scientists to develop a common language for homo-, hetero-, and bio-catalysis.

For homogeneous catalysis, studies will provide detailed information about the effects of different transition metal atoms and ligands on both the geometric and electronic structure of catalysts, reactive intermediates, and transition states. For heterogeneous catalysis, of particular interest to the Catalysis CAT, computational studies can be used to identify the active sites on catalytic surfaces and characterize how structure and composition (e.g., oxides versus pure metals, dopants) affect reaction energetics and dynamics. For biocatalysts, computational studies can be used to identify the reaction pathway that can then be used in genetic engineering studies to modify enzymatic rates.

2 U.S. Department of Energy (DOE) Basic Energy Sciences Advisory Committee, 2005. *Opportunities for Discovery: Theory and Computation in Basic Energy Sciences*. DOE. Washington, D.C.

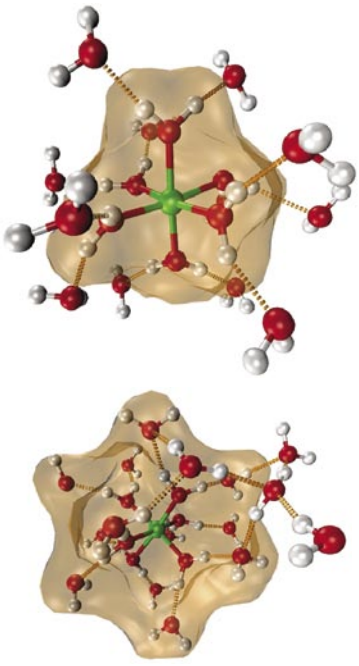
Chemical transformations also will play essential roles in all areas of developing an economically viable hydrogen economy, currently an important focus of the President and DOE. Employing hydrogen as an energy carrier requires the development of technologies to produce, store, and use it efficiently. Hydrogen production relies on chemical transformations, typically water or hydrocarbons, to form H_2. New catalysts will be needed to facilitate these transformations. Transportation of hydrogen will require high densities not afforded by gas-phase or pressurized H_2 and calls for the design of new materials, such as chemical or metal hydrides, that will reversibly evolve and take up H_2 through chemical reactions, most likely using catalysts. Fuel cells, in which O_2 and H_2 undergo electrochemical processes to convert chemical energy to electricity, rely on electrocatalytic processes as well as proton and electron transport. Understanding chemical reactions sufficiently to be able to control them is essential to the successful implementation of hydrogen as a fuel carrier.

Thus, computational chemistry is well positioned to provide fundamental understanding to realize control of chemical transformations. Predictions of reaction mechanisms and rate constants require:

- ***Accurate electronic structure theories to obtain information about the potential energy surfaces of large systems***
- ***Statistical mechanical approaches for sampling relevant configurations contributing to the reaction***
- ***Dynamical theories for calculating the overall rates of reaction***
- ***Multiscale theories, including master equations, for inferring overall rate coefficients and their pressure dependences from the above.***

Although challenges do still exist, significant progress has been made in all of these areas, and reliable predictions of gas-phase reaction rates are now possible. A major scientific challenge for the future is to develop computational tools that provide the understanding required to control chemical reactions in condensed-phase environments and at complex interfaces at the same level of detail that gas-phase reactions are understood and controlled today.

New theoretical capabilities must be exploited to develop a quantitatively predictive understanding of the chemical reactions and rearrangements in condensed phases and complex environments that underpin new technologies in the chemistry of environmental remediation (e.g., contaminant degradation in the environment by natural and remedial processes, degradation of materials used for nuclear waste entombment, destruction of hazardous waste by supercritical oxidation, environmental and downstream reactions that determine the fate of by-products of energy-producing and -using processes), photosynthesis, catalysis for efficient energy production, and combustion.



Modeling "weak" second (above, top) and third (above, bottom) hydrogen bonding interactions around highly charged ion in water at 320 Kelvin (images courtesy of Bylaska and Valiev, PNNL; Weare and Rustad, UCSD).

Host-Guest Interactions

Weak interactions, such as hydrogen bonds and van der Waals interactions, and stronger interactions, such as ion-to-ion interactions and bonds between ligands and metal atoms, are crucial to several areas important to DOE, including bioremediation, separations science, and sensor design, and to many of EMSL's distinctive science signatures. These types of host-guest interactions are responsible for forming molecular complexes without creating a chemical bond (i.e., the amount of electron sharing between the interacting species is small compared to covalent bonds). Knowledge of these host-guest interactions is crucial to understanding the function of biomolecules in the living organisms envisioned for bioremediation of contaminated DOE nuclear production sites. A strong overlap and interaction exists between chemistry and biological simulations, as indicated in the Biological Sciences section of this document.

Because these interactions are relatively weak, except in the case of cation/anion interactions, the processes of making and breaking these bonds often are more facile and reversible than are chemical bonds. Processes regulated by host-guest interactions generally involve multiple interactions between the guest (e.g., a metal ion) and the host (e.g., an interaction site on a macromolecule). Therefore, the dynamics of the process involves collective effects and more complicated reaction coordinates than simple bond breaking or formation. An important aspect of these types of interactions is high-level electronic structure calculations do not need to be used to address the bond-breaking aspects. However, a large number of small interactions often are not additive in their effects.

Computational chemistry already is significantly contributing to the study of host-guest interactions. Computational tools allow accurate characterization of these interactions, including providing an understanding of how the interaction energy changes with structure of the molecular complex and how the molecular environment affects this interaction. Computational approaches that account for dynamics and fluctuations of the molecular complexes are essential to understanding the thermodynamically stable conformations and to unravel the dynamical mechanism of complex formation. Computational approaches also are needed to design new separation systems through the use of high-throughput screening approaches and genetic algorithms.

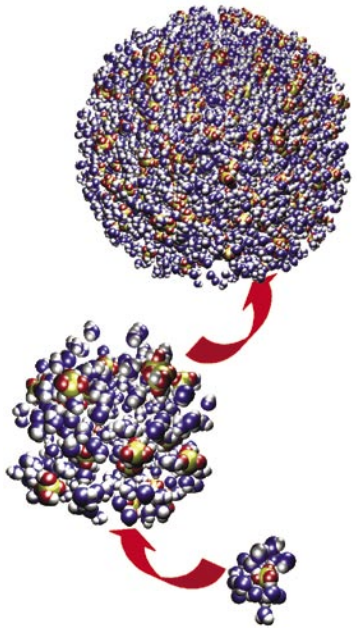
Aerosol Chemistry

Aerosol particles generated by energy use (e.g., combustion) significantly impact the atmosphere, affecting climate change, human health, and visibility (smog). Research leading to a comprehensive understanding of aerosol formation and transformation processes is of vital interest to DOE. Primary emissions of particulates from combustion include soot and organic aerosols created from condensation of lubrication oils in diesel exhaust. Secondary aerosols are created by emission of gas-phase precursors, such as hydrocarbons from incomplete combustion or SO_2 from coal-fired power plants, which are oxidized in the atmosphere to semivolatile

compounds that condense to form nanoscale particles. In addition, there are substantial natural contributions to organic-based aerosols, such as emissions from trees in the southeast United States. Aerosols affect the Earth's radiation balance by scattering and absorbing solar radiation and by acting as cloud condensation nuclei to influence the formation, lifetime, and radiative properties of clouds. Atmospheric aerosols also are important because they provide reaction pathways for oxidant production and destruction that are not present in the gas phase. Evidence also exists that aerosol particles less than 10 μm in size can cause diseases in humans. In addition, aerosols provide mechanisms for the removal of pollutants from the atmosphere that could lead to terrestrial and aquatic ecosystem damage, and are known to affect visibility in urban and regional areas. All of these research areas are directly related to science of interest in the Atmospheric Chemistry CAT and to EMSL's aerosol chemistry distinctive science signature.

A comprehensive understanding of aerosol formation and transformation processes is of vital interest to DOE.

The mechanisms of aerosol formation, such as the factors controlling soot production in combustion processes, are poorly understood. Similarly, the mechanisms for gas-to-particle nucleation in the atmosphere, where many gas-phase components such as water, sulfuric acid, nitric acid, ammonia, and organics can contribute, are not known well enough to enable reliable predictions of particle formation. The nucleation processes are coupled to chemical reaction mechanisms and other aerosol processes such as growth and coagulation. For example, OH in the troposphere oxidizes SO_2 emitted from coal burning to form sulfuric acid, which nucleates with water to form aqueous sulfuric acid droplets. The presence of other gas-phase species such as ammonia is thought to increase the particle production rate significantly, although the evidence to date is anecdotal. The mechanisms of aerosol aging are equally poorly understood. Over time periods of hours to days, the composition of atmospheric aerosols can be altered by the addition of new species, loss of components (such as water by evaporation as the relative humidity decreases), or chemical reactions at the aerosol surface or in the bulk of liquid particles, and these changes can significantly modify aerosol properties, such as optical properties and the ability to act as cloud condensation nuclei. The formation, growth, and aging of aerosol particles are truly multiscale, with molecular-scale processes intimately coupled to macroscopic phenomena on very large scales.



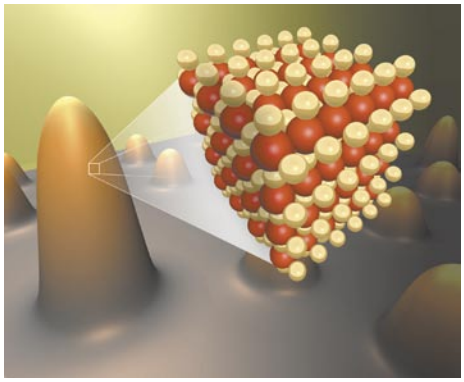
Modeling nucleation for sulfuric acid and water from a minimal critical size cluster for nucleation (above, bottom), to a nucleated particle with growth (above, middle) that can be studied on MPP2, where the future goal is to model a nanodroplet of sulfuric acid (above, top).

Computational approaches combined with experimental studies have led to unprecedented advances in the understanding of aerosol chemistry. As discussed earlier, computational tools can accurately predict chemical reactions that produce gas-phase precursors, and a molecular-level understanding may now be achieved for the complex process of soot formation. New molecular-scale computational approaches have been developed to describe nucleation processes and are being extended to treat nucleation in systems in which multiple gas-phase components participate. However, high accuracy is required for reliable predictions, and the need for accuracy always substantially increases computational cost.

Advances currently are being made in the capabilities to model collisions of radical species such as OH with soot particles and organic compounds at the gas/particle interface. Calculations will provide detailed information about the properties of aerosols, including their reactivity, as a function of aerosol composition. The distribution of different components in aerosol particles should also be understood, since it can affect properties such as reactivity with other atmospheric molecules. For example, enhancement of surface concentration can significantly increase reaction probabilities and open up new mechanisms that might not occur in the bulk phase. Molecular simulations are beginning to provide valuable insight into the distribution of ions in aqueous droplets and the impact of increased surface concentrations of heavy halide ions on reactions with gas-phase OH radicals. Moreover, computational approaches are essential for predicting how light will interact with aerosols, including scattering, absorption, and photochemistry.

Nanoscience

Nanoscience—research on systems functioning at nanometer-length scales—has attracted enormous interest since size constraints often produce qualitatively new behavior. For particles of 1-nm^3 volume, the size regime is that of molecular systems and properties for isolated particles can be calculated with reasonable reliability using current methodologies, albeit with significant computational cost. Nanoparticles have a larger surface-to-volume ratio and often different structural features compared to bulk materials. The properties of nanostructures strongly depend on size, shape, and composition and can differ significantly from either bulk materials or isolated molecules. This area of research has grown explosively in the past decade and holds the promise of revolutionary advances in areas of interest to DOE. For example, tailoring nanoscale catalysts could potentially lead to controlled chemical reactions important in contaminant destruction as well as with chemical and energy production. So, while nanoscience is not explicitly one of EMSL's distinctive science signatures, the role of chemical processes at the nanoscale is likely to be critical to, for example, understanding contaminant fate and transport in the subsurface; clear evidence exists that chemistry in microenvironments is unmistakably important in such processes. The deliberate design of nanoscale materials with novel and enhanced functionalities may also impact other areas important to DOE, such as solar energy conversion, energy-efficient lighting, chemical and biological sensors, and optical and magnetic properties of materials. In addition, nanoparticles created as primary and secondary by-products of energy use (e.g., soot and secondary organic aerosol particles) significantly affect human health and the environment. Finally, it should be recognized that biology predominantly occurs on the nanoscale in terms of chemical processes controlled by enzymes and the critical role of molecular complexes and machines.



Three-dimensional nanostructures, or nanodots, are so small that about 100,000 of them would fit on the head of a pin. Also known as quantum dots, nanodots are metal oxide crystals that are like artificial atoms with unique electronic properties. Unlike normal atoms, the properties of the nanodots can be changed by changing the material size, material composition, and how they interact with the substrate.

Realizing the potential of nanoscience will require knowledge of how properties of nanoparticles change with size, structure, and composition. For example, changes in electronic structure caused by confinement of nanoscale

semiconductor materials to sizes less than the de Bröglie wavelength can lead to large shifts of optical and photocatalytic properties. A systematic understanding of the factors that control these properties does not yet exist. Also not understood is why changes in the size of small gold nanoclusters lead to large variations in reactivity. Yet, such properties are critical for the design of biosensors for the environment and human health, as evidenced by the recent work to develop sensors for Alzheimer's Disease based on decorated gold particles[3]. Another major challenge of nanoscience is the ability to control the synthesis of desired nanoscale materials. Experimental techniques have been developed with limited control of fabrication that produce cubes, spheres, prisms, rods, polypods, polyhedra, and discs with reliable size and shape distributions. However, the underlying chemistry and physics controlling the specific structures formed remains unknown. Structured nanomaterials can be formed by self-assembly of nanoscale building blocks with properties tailored to meet a variety of technological needs, but little is known about the formation of nanoparticles under solution conditions. This has implications for both synthetic materials and the formation of natural particles. Understanding the mechanisms of creating macroscopic objects from nanoscale building blocks remains a significant scientific challenge.

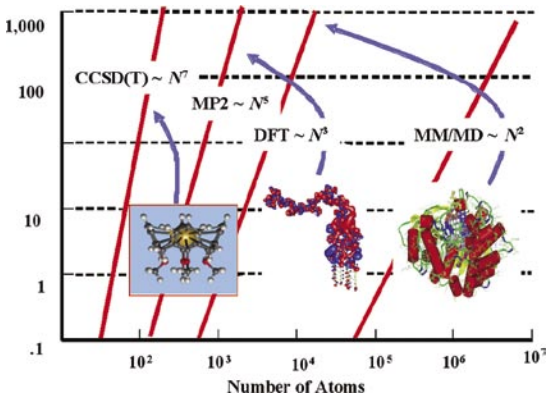
New theoretical and computational approaches must be implemented to understand the emergence of new phenomena at the nanoscale. Computational chemistry methods can be extended to nanoscale systems to provide information about preferred structures that is difficult to obtain experimentally—for example, knowledge of the electronic structure underlying optical and electronic properties and a detailed understanding of the correlation among structure, size, and those properties. However, this will require larger computational resources than currently are available—ten to one hundred times the current computational capability. Computational methods also will be important in understanding the impact of dynamics and fluctuations on nanoscale properties and phenomena. Understanding these phenomena will require a systems approach where many simulations are required. Finally, theory and computation are needed to unravel the complex mechanisms of nucleation, growth, and self-assembly of structured nanomaterials. New methods must be developed to make reliable predictions of this behavior, requiring substantial computational resources.

Common Needs and Requirements for High-Performance Computing in Chemistry Research

Several common needs arise for chemistry research. Chemistry, as the science of atoms and molecules, and computational chemistry at its most basic level, requires knowledge of *molecular and atomic interactions* or energies as a function of interatomic distances (i.e., the potential energy surface [PES]) for a collection of atoms and molecules. Knowledge of portions of the PES allows predictions of equilibrium

3 Georganopoulou DG, L Chang, JM Nam, CS Thaxton, EJ Mufson, WL Klein, and CA Mirkin. 2005. "Nanoparticle-Based Detection in Cerebral Spinal Fluid of a Soluble Pathogenic Biomarker for Alzheimer's Disease." *Proceedings of the National Academy of Sciences* 10.1073/pnas.0409336102.

structural, *spectroscopic*, and *thermodynamic properties*, and determination of *dynamic properties* (e.g., rate constants). In addition, coupling processes occurring from molecular scales to spatial and temporal scales of macroscopic processes will require *multiscale/multiresolution* methods.



To model larger, more realistic systems with higher fidelity, a combination of lower-scaling methods and increased computing power is needed.

The drive to perform simulations of more realistic models (i.e., those that represent the actual systems probed in experimental observations with higher fidelity) with fewer approximate methods requires increased computational resources. Depending on the method used, scaling of computational effort with system size can vary from nearly linear to exponential and increases in computational effort with system size ultimately limit the size of the system that can be studied. Today, it is now possible using quantum mechanical approaches to make predictions of molecular properties of *modest-sized* systems (i.e., tens of atoms) that are as reliable as the most detailed experiments, especially for thermodynamics, structure, and many spectroscopic properties. However, condensed-phase chemical problems, which essentially are models of the "real" physical system, and difficult chemical systems, such as those that include transition metals or heavy elements, have the potential to exhaust all computational and algorithmic resources in the quest for increased reliability of the simulations. Approximate methods often are employed to reduce the computational effort; therefore, there is a need for *benchmark calculations* to establish the validity of computational methods and models, as well as the uncertainty and limitations of the methods.

Each of these cross-cutting themes—molecular and atomic interactions; structural, spectroscopic, and thermodynamic properties; dynamic properties; multiscale/multiresolution methods; and benchmarks—is discussed below in the context of the science areas presented earlier. Also presented are the requirements for high-performance computing, again, relative to the cross-cutting theme areas.

Molecular and Atomic Interactions

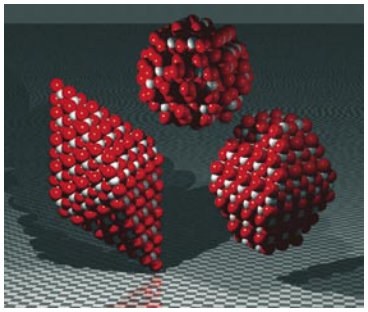
Electronic structure methods are used for characterizing molecular interactions. However, limitations exist in scaling of electronic structure methods required for the accurate calculation of portions of a PES. A specific example of this limitation is in the area of catalysis. Although intermediate-level computations using methods such as density functional theory (DFT) and second-order Møller-Plesset perturbation theory (MP2) can often provide insight into how a catalyst works, the true computational design of real catalysts requires the ability to predict accurate thermodynamic and kinetic results. At this time, only CCSD(T) or multireference configuration interaction methods, with basis sets that allow extrapolation to the complete basis set limit, provide the necessary accuracy. Such calculations are extremely difficult, as they scale as N^7 and use large basis sets, and N increases linearly with both the number of electrons in the system

and the size of the basis sets used. In addition, calculations of rate constants for kinetics require more points on the PES than just single minima for reactants and products, substantially raising computational costs. An example of the state of the art with the current MSCF computational hardware and software is the calculation of the heat of formation of octane, C_8H_{18}. The most computationally intensive part of this calculation required 1,400 processors for 23 hours, yielding 75 percent CPU efficiency for a sustained performance of 6.3 teraflops. With the scaling of current accurate methods ranging from N^5 (MP2) to N^7 [CCSD(T)], the need for access to large computational resources for accurate predictions is evident. Keep in mind, this example involved a single energy at a single nuclear configuration. A simple kinetics calculation at this level of accuracy for a similar sized system would require finding the stationary points (reactant, transition state, and product), which requires approximately thirty energy evaluations as a conservative estimate. Hence, one would need approximately 200 teraflops to obtain the kinetics of one reaction. Of course, more approximate methods that have been appropriately benchmarked could decrease this estimate.

Many systems, including heterogeneous and homogenous catalysts and biological systems (e.g., enzymes), involve transition metals, which require the accurate treatment of multivalent, open-shell atoms, thus presenting a challenge for applications of electronic structure methods. The high-accuracy possible for systems containing first- and second-row elements is currently not attainable for transition metals. To improve accuracy, *simultaneous* improvements in basis set, treatment of electron correlation, and the inclusion of relativistic effects are essential. Unfortunately, advanced treatment of all three factors *simultaneously* is far beyond current computational capabilities. For heavier atoms, including actinides, additional complexity arises from the necessity to simultaneously include scalar and spin-orbit relativistic effects.

Accurate calculations of weak interactions make different demands of the electronic structure calculations. Interactions between closed-shell species often are dominated by electrostatics and dispersion for which DFT methods with current approaches and functionals are not optimal. MP2 calculations with basis sets extrapolated to the complete basis set limit have been shown to provide accurate accounting of hydrogen bonding interactions. Although the size of systems for which accurate calculations can be performed is growing (e.g., accurate energetics can be obtained for clusters with more than 20 water molecules), extension of these types of calculations to more reliable models of condensed-phase systems will require significant increases in computational resources. For example, explicit calculations of up to 100 water molecules would require an approximate 600- to 3,000-fold increase in the computational resources depending on the method used. Of course, implicit methods decrease this significantly, but also tend to decrease the accuracy of the results.

Extending electronic structure calculations to treat large biogeochemical systems, nanoscale systems, and solid-state materials (e.g., those important in interfacial chemistry) place an even larger demand on computational resources. Calculations



Predicting titanium oxide anatase nanocrystal shapes using density functional computations. From left to right: bypyramid, sphere, bifrustum (image courtesy of Barnard, Zapol, and Curtiss; ANL).

of the electronic, optical, and structural properties of nanoparticles containing thousands to tens of thousands of atoms currently rely on approximate electronic structure methods such as those based upon tight-binding approaches at either the molecular orbital theory or DFT levels. Planewave methods, especially those involving periodicity, also are commonly used for these types of simulations. Calculations of accurate reaction energetics, kinetics, and excited states (band gaps for solid-state systems) require improved methods (e.g., DFT functionals that are more accurate than those currently available) as well as increased computational capabilities.

Electronic excitations in molecules and solid-state systems are important in photochemistry, photocatalysis, and radiation chemistry. The most accurate, non-empirical approaches to calculating excitation energies require large, diffuse basis sets and a balanced, multiconfigurational representation of the electronic wave functions. These approaches are currently applicable only to relatively small molecular systems. The treatment of photochemical and radiation processes in condensed phases requires new approaches to treating large molecular ensembles as well as improved computational resources. The long-range effects of condensed phases, particularly polarizable media, on the electronic structure of ground and excited states can be profound, yet these are not readily included in current approaches due to a lack of theoretical techniques and limited computational resources. Another complication in treating dynamical processes of excited states is the need to evaluate additional terms that couple the different electronic states. Accurate calculations of the couplings require highly accurate wave functions and their derivatives with respect to coordinates. In addition, nuclear motion can be closely coupled to the electronic states, and this adds additional complexity to the calculation, (i.e., the Born-Oppenheimer approximation is no longer valid).

Calculations of the energetics of molecular interactions are just the first step in simulating important properties of molecules and materials. Calculations that average over many configurations or follow the dynamics of systems require large numbers (thousands to many millions) of evaluations of the interaction potential. The traditional approach has been to use the results of accurate electronic structure calculations to fit analytical representations of the interaction potential. Current approaches to the fitting of a PES require large investments of scientists' effort and time. Alternatively, direct approaches evaluate the interaction potential or its derivatives using electronic structure methods without the intermediate fitting step, which is computationally intensive. The area of direct simulations will benefit from development of efficient interfaces between electronic structure and dynamics simulation methods. Although the scaling of computer time is only linear with the number of direct evaluations of the PES, the fact that large numbers of evaluations are needed (up to many millions) significantly increases the demand on computational resources. Current methods, such as Car-Parrinello dynamics, have provided unique insights into dynamic properties on short time scales, yet the electronic structure on which the dynamics are based is only approximate and the time scale is very short. Such methods are needed to explore reactivity in solution and biochemical systems.

Structural, Spectroscopic, and Thermodynamic Properties

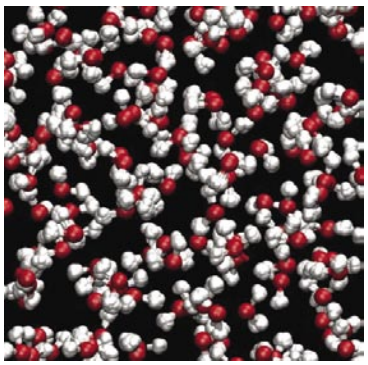
One of the first properties determined in computational studies of molecular systems is the arrangement of atoms—molecular structure. Equilibrium structures that define reactants and products in chemical reactions and transition state structures are critical in understanding reaction mechanisms and kinetics. The structure of a protein determines its functionality and activity, and the structure of a nanoparticle controls its properties. The optimal determination of global minima, especially transition states, of systems is not a solved problem and currently requires repeated sampling of the PES. The computational effort scales nonlinearly with dimensionality of the system; in the most simpleminded approach of sampling over n points in each dimension D, the effort scales as n^D, where D is the number of degrees of freedom in the system. Even for more intelligent algorithms, it is often difficult to know if the global minimum has been located for modest-sized systems (tens of atoms) even after millions to billions of evaluations of the PES have been performed—especially for systems dominated by weak interactions where the directionality conferred by a covalent chemical bond is not present.

Spectroscopy, particularly vibrational (nuclear motion) spectroscopy, is a critical tool for determining molecular structure and is a necessary component of EMSL's spectra signatures and trace detection distinctive science signature. The prediction of the vibrational spectrum of a point on the PES is necessary to determine if the point is a minimum or transition state and to provide spectral properties (e.g., IR and Raman) for comparison with experiment. Accurate calculations of vibrational spectra require not only accurate energetics, but also methods to calculate bound-state energy levels for multidimensional systems. Typically, the energy levels are approximated by harmonic representations of the PES around its minimum. The harmonic approximation is insufficient for accurate determination of vibrational spectra and zero-point vibrational energies required for thermochemical quantities, and a need exists for methods that treat multidimensional anharmonic vibrations. Current approaches to the accurate prediction of the anharmonic contributions are prohibitively expensive for systems with more than several degrees of freedom. However, new, higher-order derivative methods that are computationally expensive but do not require PES sampling are being developed to address this problem. As an example of the critical importance of the vibrational problem, consider the size of the zero-point energy for C_8H_{18}. The zero-point energy for this molecule is calculated to be approximately 150 kcal/mol. To obtain a chemically accurate heat of formation within ±1 kcal/mol requires the zero-point energy to be known with an accuracy of substantially less than one percent.

Experimental observations of thermodynamic properties, such as enthalpies and free energies, are averaged over displacements created by thermal or quantum effects. For molecules, this is not a problem given the low-lying molecular conformations and vibrational energies. However, for clusters dominated by weak interactions with many conformations, this is far more complicated. Calculations of thermally averaged properties require statistical sampling of the regions of configuration space that contribute to the observation. This is accomplished by using methods

such as Monte Carlo and molecular dynamics for classical simulations and methods such as path-integral Monte Carlo for quantum simulations. The level of effort in these calculations depends strongly on the topology of the PES. Properties that require averaging over many local minima can be difficult to converge for complex (rugged) energy landscapes. As in the case for locating local minima, convergence of thermodynamic properties can require millions of evaluations of the PES. In addition, the level of effort in quantum simulations increases as temperature decreases, making quantum effects more important.

Dynamic Properties



Researchers at PNNL use the MSCF to construct accurate centroid dynamics models describing the interaction between water molecules for obtaining accurate macroscopic structural and thermodynamic properties for its liquid phase (image courtesy of Fanourgakis, PNNL).

The calculation of dynamic properties, such as diffusion, energy transfer, and chemical reaction rates, requires following the motions of atoms and molecules as a function of time. Treating the atoms classically, atomic motion is followed by integrating the classical equations of motion (Newton's or Hamilton's equations) to determine the classical trajectory of the system. At each time step, the forces on the atoms are evaluated on the PES. The size of the time step is determined by the inherent time scales for motions in molecular systems, which are on the order of 10^{-15} seconds (fs). Time scales for dynamical properties of interest, especially rare events such as passing over a barrier in a chemical reaction, can be much larger, requiring large numbers of steps along the trajectory. Although the computational effort scales linearly with the number of time steps in a calculation, for many properties it is difficult to predict *a priori* the number of steps that will be needed to obtain a converged value for the property of interest. Prime examples include infrequent events involving many time scales, such as activated reactions, in which the trajectory rarely surmounts a large barrier; protein-folding, in which many regions of configuration space corresponding to different local minima may be explored before the global minima is located; or gas-to-liquid nucleation, which can require simulation sizes of 10^{15} and simulation times of microseconds to have a reasonable probability of observing a nucleation event in a realistic system. Direct simulation currently is not possible for such processes because the number of time steps required in the calculations would be on the order of 10^{10} or larger.

Advanced computational simulation techniques can mitigate some of the problems associated with infrequent events. More efficient methods of integrating the classical equations of motion can help by allowing optimum sizes of the time steps to be determined (as in variable step predictor-corrector algorithms) for a desired accuracy of the trajectory. These improvements can provide savings of one to two orders of magnitude, but are not sufficient on their own to address large gains needed for some of the infrequent event problems. For these problems, methods that accelerate the rate at which trajectories surmount barriers (e.g., hyperdynamics) can provide greater decreases in the number of time steps in a simulation.

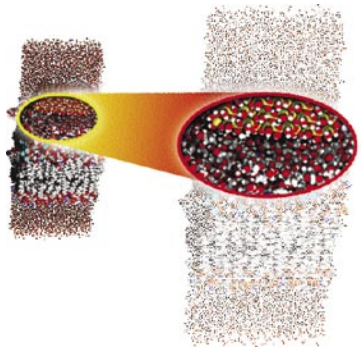
Alternatives to direct simulations of the dynamics use statistical dynamical approximations, such as those used in transition state theory and RRKM (the theory of unimolecular reaction rates), to evaluate chemical reaction rates for the process of interest. The level of computational effort in these types of calculations

is comparable to the level of effort in the thermodynamic simulations described previously. Application to gas-phase processes requires locating the transition state and reaction pathways (e.g., minimum energy path) and calculating the free-energy of activation. Improvement of these methods to deal with path curvature and improved tunneling corrections, such as those incorporated in variational transition state theory, requires about one to two orders of magnitude improvement as compared to a thermodynamic prediction. Determination of reaction pathways is a greater challenge in condensed phases where *a priori* knowledge is lacking and for which the many degrees of freedom are important in controlling the reaction. For these systems, the computational effort is much greater because of the need to evaluate statistical averages (as in potential of mean force calculations) as well as perform searches for critical geometries and minimum energy paths. For example, the electron transfer processes, so important at interfaces and in biological systems, involve transfer of an electron where the dynamics need to be treated quantum mechanically.

Classical mechanics provides a good first-order approximation for the energetics and dynamics of many chemical and biological processes. However, quantum mechanical effects can be important and, in those cases, must be included to accurately represent the process. For example, electronically nonadiabatic processes important in photochemistry are inherently quantum mechanical and cannot be treated using standard classical mechanics. Currently, accurate quantum dynamics methods are limited to small systems, typically with only two to three degrees of freedom, because they scale exponentially in the dimensionality of the system. These methods provide valuable benchmarks for more approximate methods, and there is the continual need to push the limits of accurate calculations to provide benchmarks for more complex systems. Quantum effects can often be included approximately into classical dynamics calculations. Examples include semiclassical methods (e.g., surface-hopping techniques and methods based on the initial value representation) or mixed classical/quantum direct dynamics. However, these methods involve additional computational effort at each time step, large numbers of time steps, and many trajectories resulting in large computational costs.

Multiscale/Multiresolution Methods

Many of the challenges described previously have significant multiscale components and will require new multiscale and multiresolution methods for successful resolution. The extension of *ab initio* electronic structure methods to larger systems and improvements to the scaling of high-accuracy methods likely will be driven by better understanding and exploitation of the multiscale properties of the wave function. Multiresolution methods being developed allow researchers to specify the desired accuracy of the solution at the outset of the calculation, instead of performing complicated studies of the convergence of molecular properties as a function of basis set size (the current approach). Better partitioning of calculations based on spatial and time (energy) scales could lead to more efficient solution methods for *ab initio* theories and lower scaling laws.



Biogeochemical modeling of an interface of a cell membrane and a mineral surface. EMSL's MPP2 can use molecular mechanics to model this interaction (above, left), whereas the future goal is to model the interface itself quantum mechanically (above, right). Image courtesy of Straatsma, PNNL.

Unlike the gas phase, reactions that take place in condensed phases involve large numbers of degrees of freedom, even if many of them are only weakly coupled to the reaction coordinate. However, not all degrees of freedom are equally interesting, and usually there is no benefit in modeling the entire system at a high level of accuracy just to determine the energetics of a small portion. Thus, better hybrid methods that can model components of systems at different levels of accuracy are required to allow the reaction site to be modeled with very high fidelity, but still permit sufficiently detailed modeling of surrounding atoms to provide an accurate determination of the reaction energetics. Examples can be found in the quantum mechanics/molecular mechanics methods that couple quantum mechanical and classical descriptions of molecules, which have been developed primarily to model enzymatic catalysis. These methods are currently restricted to systems where the components remain in relatively fixed relationships with each other, such as in enzymes and surface catalysis. Better methods are needed for more dynamic environments, such as those found in liquids. Methods that allow coupling between different levels of theory, such as semi-empirical, Hartree-Fock/self-consistent field, and CCSD(T), also are needed. Other methods that include the environment explicitly, such as Effective Fragment Potential, must continue to be developed.

Beyond the multiresolution issues involved in the atomic-scale description of matter, a host of multiscale issues is involved in shifting from atomistic to continuum systems—at which juncture sit many of DOE's interests. Nanoscale systems are big enough to contain prohibitively large numbers of atoms for current molecular electronic structure approaches, but are still small enough to be heavily influenced by atomic-scale features. The interactions of nanoparticles with surfaces also have important multiscale features. The adhesion of the particle is dominated by atomic-scale interactions at the interface, but this can lead to material distortions that can extend over significant distances around the nanoparticle. These distortions then couple back to the particle-surface interaction and can influence the organization and cluster of nanoparticles on surfaces. Quasi-continuum methods that couple atomistic and continuum (elastic) descriptions of materials have shown considerable promise in bridging the range of scales involved. More work is needed to extend these methods to model dynamics, finite temperature systems, and solutions.

In addition to handling multiscale systems in the spatial domain, a huge number of problems exists that are multiscale in both time *and* space. Many systems that self-assemble fall into this category—including nucleating aerosols, surfactant solutions, self-assembling monolayers, quantum dots and wires, and proteins. These systems are often characterized by a broad continuum of time scales, ranging from the frequency of individual atomic vibrations to hydrodynamic relaxation times over distances of nanometers to micrometers. This continuum can represent a dynamic range on the order of millions or more. The problem is further complicated by the lack of distinctly separated space and time scales. Traditional coarse graining and averaging procedures break down, and no method exists to accurately integrate out fast degrees of freedom from the system to create coarse grained models that quantitatively capture the dynamics of the system over long periods of time. Rate equation approaches such as transition state theory also are often difficult to apply because many of these

systems have reaction coordinates embedded in complex multidimensional spaces that are difficult to identify or have multiple pathways between different states. New approaches are needed for integrating out fast degrees of freedom to produce lower dimensional models that accurately reflect the dynamics and can be simulated over long periods of time.

Benchmark Calculations

Benchmark calculations are crucial to validating models of atomic and molecular systems and methods used in calculations. They can also be used to establish regions over which the approximate models and methods can be used with specified levels of uncertainty. Benchmark calculations also provide the databases of information needed to shift the paradigm of simulations from one-off projects toward one of automated modeling with "dialable" accuracy. With this shift, the expectations of the broader community will change from viewing calculations as models of uncertain accuracy to viewing them as reliable predictors. One of the roles of benchmarking is to quantify uncertainty and model limits. Often the methods that are used as the standard in benchmarking studies are compared, themselves, to experimental results, emphasizing the continued need for interactions between EMSL's computational and experimental resources.

High-Performance Computing Requirements

Most computational chemistry algorithms are built on a cache-blocked architecture; therefore, cache latency, bandwidth, and size are important to performance and scalability. Also key to performance is the ability to overlap computation, communication, and I/O operations (e.g., use of asynchronous I/O, use of nonblocking communication operations). Communication bandwidth and latency also are critical to calculations involving large matrix operations. In addition, large local memory capabilities benefit the calculations, as does access to large amounts of fast storage capabilities for temporary storage of intermediate results. A key issue is the ability to handle distributed data structures on a fully distributed memory system. It is possible to take great advantage of vector architectures in electronic structure theory based on EMSL's cumulative experience with the original Cray machines up through model C90. Thus, the role of vectors needs to be further explored. In addition, field-programmable gate arrays (FPGAs) could be used to advantage if, for example, they could be programmed in a reduced instruction set with significantly enhanced speed to calculate the required integrals. This has been tried unsuccessfully in the past with custom chips, but it may be appropriate to revisit this concept with modern FPGAs. Finally, it may be appropriate to investigate new architectures for use in electronic structure calculations.

Scaling to hundreds of teraflops or even to petaflops will require changes in algorithms. These algorithms will need to be cognizant of the underlying architecture of the system. For example, replicating some of the data may provide more scalability than fully distributing the data. Also, different parts of the computation may need to be delegated to subgroups of processors and then brought back together to proceed to

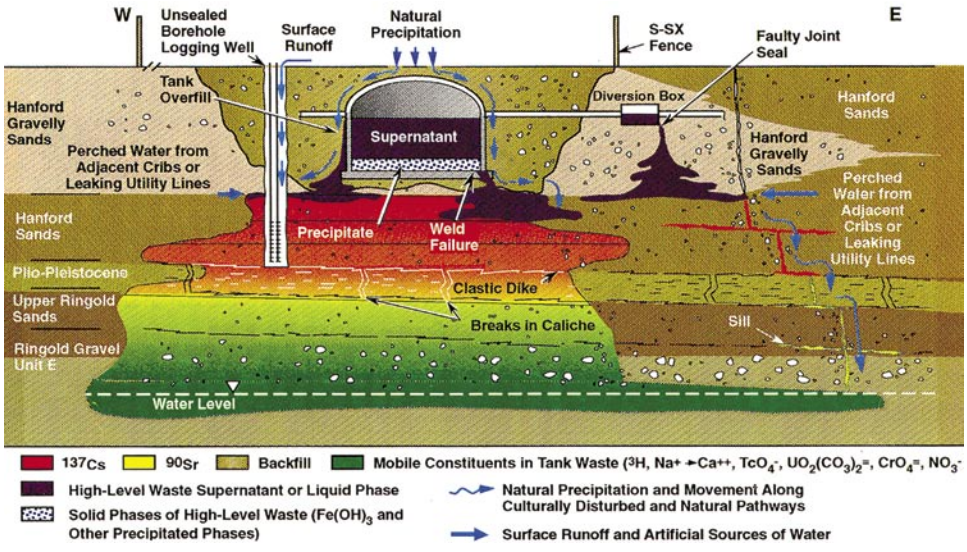
the next step. An analysis of the algorithms used in the computational chemistry software package, NWChem, developed at PNNL (and, indeed, in most quantum chemistry codes), shows that a very balanced architecture in terms of processor speed, memory, cache latency and bandwidth, processor interconnect bandwidth and latency, and latency and bandwidth to I/O is needed. These factors played a determining role in the purchase of MPP2.

Environmental Systems Sciences

Knowledge of the chemical and biological reactions of environmentally important processes can be used to resolve critical DOE environmental challenges such as (bio)remediation and carbon sequestration. Thus, large multiscale subsurface and climate models must be developed to simulate long-term events that may influence policy decisions concerning natural and human impacts on the environment. Challenges representing subsurface and climate modeling are often intertwined—for example, atmospheric modeling may feed into watershed and surface water models that link to the subsurface. (While most of the requirements for the next-generation machine will be developed from chemistry and biology, it is anticipated that the system will perform well for the environmental systems sciences described here.)

Multiscale Subsurface Modeling

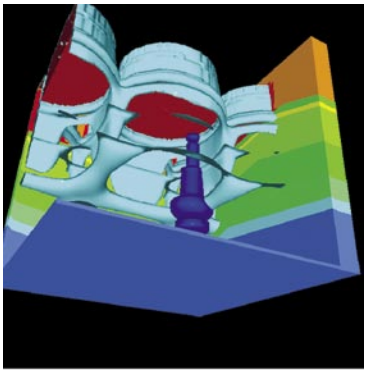
Exposure to contaminated soil and groundwater is a real concern for many communities in the United States. An accurate assessment of these threats and the design of effective and efficient alternatives for the cleanup and closure of DOE



Subsurface simulations for the environment mean accelerated delivery of scientifically defensible simulation and prediction of field-scale contaminant behavior for critical DOE cleanup decisions at the most difficult DOE sites (Hanford, WA; Savannah River, GA; INL, ID; and ORNL, TN).

waste sites is critical to protecting human health and the environment. However, the complexity of the subsurface, coupled with the limited ability to observe processes and interactions in the subsurface as they occur, have proven to be formidable challenges. Scientific model development and computational advances to predict field-scale subsurface contaminant transport and fate facilitate long-term predictions of waste management baselines and alternatives to meet DOE cleanup and closure goals. In addition, biogeochemistry and subsurface science is one of EMSL's distinctive science signatures.

The importance of subsurface simulation has been recognized and documented in several important reviews[4], assessments[5], and guidance documents[6]. These documents place a high priority on exploiting advanced simulation technology to accelerate progress for understanding fundamental subsurface processes. In addition, they further underscore the need for dedicated, advanced computing resources to effectively address issues of uncertainty and scaling.



Researchers at PNNL are using the supercomputer at EMSL to simulate the flow of leaking waste tanks at the Hanford Site (image courtesy of White and Yabusaki, PNNL).

A recent call for proposals by BER ERSD[7] outlined the critical need to improve the understanding of processes and properties controlling the transport and fate of reactive chemical species in the subsurface. One principal need is a mechanistically predictive understanding of the scale-up of contaminant behavior from the idealized and controlled conditions in small-scale laboratory studies to the field scale where multiple scales of heterogeneity in subsurface properties and pollutants exist. Numerical simulation is the principal vehicle for assembling fundamental process models and addressing spatially varying material properties required for predictions, and to obtain comprehensively detailed depictions of complex subsurface systems. Furthermore, simulation is critical to the development, testing, and characterization of field-scale subsurface process and property models, since these field-scale flow and transport models are based on averaged bulk behavior and the conditions that are scaled up from details of the tortuous pathways through individual pore throats.

Another area of subsurface research that will have critical implications to DOE and its Office of Science missions in energy security and waste legacy issues is the need to provide a better understanding of complex three-dimensional subsurface processes using high-resolution geophysical imaging. Advances in subsurface imaging will improve baseline technologies for exploring hydrocarbons and geothermal resources, as well as provide a better understanding of contaminant

4 National Research Council. 2000. *Long-Term Institutional Management of U.S. Department of Energy Legacy Waste Sites.* National Academy Press, Washington, D.C.

5 National Research Council. 2000. *Research Needs in Subsurface Science.* National Academy Press, Washington, D.C.

6 U.S. Department of Energy (DOE). 2000. *The DOE Complex-Wide Vadose Zone Science and Technology Roadmap: Characterization, Monitoring, and Simulation of Subsurface Fate and Transport.* DOE, Washington, D.C.

7 Environment Management Science Program (EMSP): Transport of Contaminants in Subsurface Environments at DOE Sites. Program announcement to DOE National Laboratories, LAB 05-12.

transport within geologically chaotic media, such as the vadose zone. Moreover, advancements in such imaging technology will help ensure the safety of DOE nuclear waste repositories at the Waste Isolation Pilot Plant in New Mexico and at Yucca Mountain in Nevada through better geologic site characterization and provide new tools for safe sequestration of carbon dioxide (CO_2) in underground gas-depleted reservoirs.

Subsurface injection of CO_2 produced from power-generating facilities is now being considered by DOE as a means to alleviate the adverse effects of global warming by reducing the concentration of this greenhouse gas in the atmosphere. Model simulations of CO_2 sequestration in a deep aquifer involving both two- and three-dimensional simulations with at least two distinct phases (liquid water and supercritical CO_2), and potentially a third phase (gas), are being and will continue to be used to assess the escape of CO_2 to the surface. These simulations require coupling flow and transport models with multispecies geochemical interactions involving dissolution (complexing reactions) of CO_2 in the aqueous solution, precipitation and dissolution of solids involving alteration of the aquifer minerals, and potentially the formation of new carbonate-bearing minerals. Both single- and dual-continuum models will be required for analysis. Future computing requirements to account for these multiphase-multispecies interactions and multiple interacting continua will require several orders-of-magnitude increase in performance over present capabilities.

Comprehensively detailed subsurface simulations based on multiphase-multicomponent and multiple domain/continuum models are being used to systematically and holistically integrate multiscale information from the laboratory and field into mechanistic process model representations and capture multiple interacting subsurface processes in many complex subsurface settings. A recent two-dimensional reactive transport simulation on MPP2 generating approximately 70 million "unknowns" during each time step required 256 processors for 72 hours to generate 80 simulated years. The need to add a third dimension and increase to the regulatory time frame of 10,000 years could conceivably increase future computational requirements by 100,000 times.

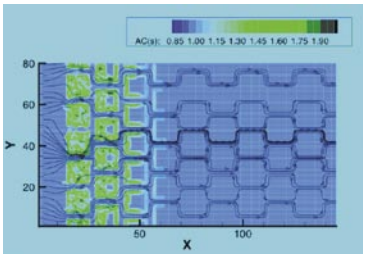
Difficulties associated with subsurface complexity are not unique to environmental remediation. As previously discussed, the engineered repository at Yucca Mountain for storage of commercial nuclear waste from civilian energy production, the geologic sequestration of carbon to reduce the buildup of atmospheric CO_2, and the characterization and extraction of oil and gas deposits illustrate subsurface applications of interest to DOE that require a detailed understanding of multiple physical and chemical processes and will benefit from improved surface imaging of complex mixtures in strongly heterogeneous subsurface materials. The need exists for reliable prediction of field-scale subsurface behavior to help form decisions related to environmental stewardship and the protection of human health, with long-term implications for national environmental and energy security.

Multiscale Theories and Models

Development of new computationally efficient multiscale theories and process models are required for reliable predictions of field-scale behavior. Increases in computational capability and performance are facilitating model development since larger, higher-resolution data sets can be used as test beds. As a result, new physics models of fundamental behaviors at multiple scales are being developed that include:

- ***Molecular-scale models of mineral surface reactions***
- ***In silico models of metal-reducing bacteria***
- ***Imaging of complex three-dimensional geological systems***
- ***Pore-scale models of multiphase flow and multicomponent reactive transport***
- ***Simulations of multiphase flow and multicomponent reactive transport at the scale of the fundamental Darcy representative elementary volume, greater than approximately mm^3.***

The new physics models, in conjunction with recent advances in measurement and characterization technology (e.g., nuclear magnetic resonance, X-ray synchrotron, neutron scattering, electromagnetic and seismic geophysical methods), can provide unprecedented detail of structures and processes in the subsurface environment. These advancements pave the way for comprehensive simulations with high spatial and process detail that can begin to link and integrate research at different conceptual length scale domains (e.g., molecular to microscopic, pore to meso) to resolve long-standing subsurface "upscaling" issues. In addition, new research must be conducted at the pore scale to determine the effect of evolving biogeochemical microenvironments associated with multiregion (multiporosity and multipermeability) domains on the mobility of reactive components. This will require development of new capabilities for existing pore-scale simulators to address spatially detailed alternatives to the simulation of bulk surface processes (e.g., adsorption and redox reactions). Hybrid modeling approaches (e.g., embedding pore-scale models in macro-scale simulators) are potentially powerful tools for merging seemingly disparate process scales.



Lattice Boltzmann simulation of fluid flowing through a network made up of an idealized staggered set of mineral grains AB and reacting chemically. Shown is the volume fraction (normalized to vary between 1 and 2) of mineral AC (light green to red) that precipitates as mineral AB (not shown) dissolves. The black lines indicate the direction of flow (image courtesy of Lichter, LANL).

Robust multiscale subsurface models enable a sound defensible prediction of long-term risk, as well as the engineering of more innovative remediation schemes based on the simultaneous manipulation of physical, chemical, and biological conditions in the subsurface to enhance contaminant immobilization and/or destruction. The ultimate coupling of process models over multiple scales must address:

- ***Complex mixtures of multiple fluids with multiple phases whose properties are dependent on temperature and composition***
- ***Physically, chemically, and biologically heterogeneous subsurface materials with orders of magnitude variability in key parameters***

- *Complex interactions of processes with different time scales*
- *Limited direct measurements of properties and modeling parameters that are typically inconsistent with the modeling scales.*

In addition to the multiscale issues associated with subsurface science, significant improvements in computational simulation capability could also be used for scientific advances at a single scale. These include:

- *Simulating regional ecological impacts of climate change by coupling groundwater, vadose zone, watershed, river, meteorological, and ecological process models*
- *Simulating long-term, large-scale, three-dimensional, high-resolution, three-phase, multifluid flow, and multicomponent reactive transport*
- *Estimating parameters for use in large-scale, long-term, three-dimensional, high-resolution, multicomponent, multiphase, and multiphysics simulations*
- *Stochastic simulation of conceptual models and realizations to quantify uncertainty in model predictions*
- *Quantifiable inversion of three-dimensional, real-time, and multisensor data*
- *Pore-scale simulation of multiple domains in a 10-centimeter cube.*

Whether single scale or multiscale, parameter estimation techniques and the treatment of uncertainty will always be important to field-scale simulation. Many computations must be performed using different parameters to understand the uncertainty in the results. Advanced stochastic simulation technologies are being used to identify sensitivity and uncertainty of model predictions, as well as resulting ecological and human health risks. These include computationally intensive Monte Carlo methods for evaluating the uncertainty of coupled nonlinear processes, upscaling of high-resolution process and property descriptions for use at field scales, and geostatistical simulations conditioned to multiple scales of soft and hard data.

Future analyses of the Hanford Site groundwater models could require a 1,000 time increase in computing power.

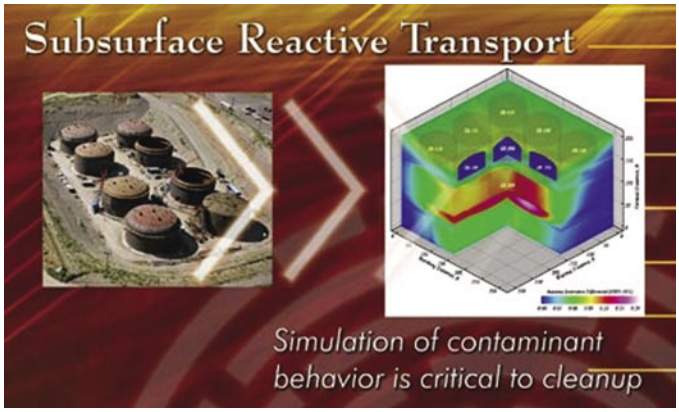
In addition, robust inverse modeling and parameter estimation technologies are being used to characterize complex natural systems by incorporating information from disparate sources (e.g., laboratory, field, static, dynamic, physical, chemical, biological, statistical, insight/prior information) and to assess model assumptions, data consistency, and data quality. The three-dimensional Hanford Sitewide Groundwater Model currently uses inverse simulations to identify geologic zonation and hydrologic parameters, including boundary conditions. Future pilot-scale analyses for the Hanford Site will require significantly more (10 to 100 times) grid resolution, will include more detailed transport processes (e.g., chemistry), and will

address 10 to 100 times more estimated parameters. These additional capabilities could require more than a 1,000 times increase in computational performance.

High-Performance Computing Requirements for Multiscale Subsurface Modeling

Transformational improvements are needed to understand and predict subsurface behaviors, enabling alternative cleanup strategies to be developed that can feasibly achieve the ambitious schedule and budget outlined by DOE. High-performance computing is necessary for developing reliable predictions of engineering performance and risk that will be used to assess alternative waste management approaches. Risk is typically assessed over very long time frames, which puts an additional burden on the mechanistic description of key subsurface processes. Advanced computing resources provide the high-performance and large memory capabilities to simultaneously address long simulation periods, comprehensive treatment of coupled processes, and the resolution of spatial and process-level details in the context of 1) multiscale variability in material properties and 2) uncertainties in conceptual process models and parameters.

The types of multiscale and hybrid approaches mentioned previously, coupled with long-time simulations, are expected to increase computational requirements by several orders of magnitude. In addition, parameter estimation and uncertainty computations will add several more orders of magnitude to obtain the reliability necessary for policy decisions. For example, the Hanford Site composite analysis encompasses several different models and solution techniques that address multiple environmental pathways on the Hanford Site. The risk analysis is based on 1,000-year simulations of 14 contaminants originating at 1,053 waste sites using 100 realizations to address variability and uncertainty in the parameter space. Current analyses using 150 processors of MPP2 takes three to four months. A need exists to 1) increase dimensionality and resolution in the operational areas to better represent the existing and future plumes, 2) incorporate additional process models (e.g., multicomponent geochemistry to reflect spatially and temporally variable contaminant concentrations and bank storage to identify release and migration of contaminants through seepage faces), 3) add more sophisticated time-dependent boundary conditions, and 4) increase the number of realizations to account for more of the parameter space. These additional capabilities could increase the computational requirements by 100 to 1000 times the current use.



Thus, to address the complexities of subsurface modeling, the required computational resources must include large amounts of floating-point operations, low latency and high-bandwidth memory, large cache that will enable reuse of data, overlap of communication and computation, and availability of large disk storage capabilities to store the large amount of data being produced.

Multiscale Climate Modeling

Climate models are essential tools for synthesizing observations, theory, and experimental results to investigate how the Earth system works and how it is affected by human activities. The modeling strategy of the U.S. Climate Change Science Plan (CCSP) envisions two complementary streams of climate modeling activities (see http://www.climate.science.gov). The first is principally a research activity, which will maintain strong ties to the global change and computational science research communities to rapidly incorporate new knowledge into a comprehensive climate and Earth system modeling capability. Closely associated with the research activity, but distinct from it, will be the sustained and timely delivery of predictive model products that are required for assessments and other decision support needs. There are three main climate modeling goals:

- ***Goal 1:** Improve the scientific basis of climate and climate impact models*
- ***Goal 2:** Provide the infrastructure and capacity necessary to support a scientifically rigorous and responsive U.S. climate modeling activity*
- ***Goal 3:** Coordinate and accelerate climate modeling activities and provide relevant decision support information on a timely basis*

Comprehensive climate system models provide the primary quantitative means to integrate scientific understanding of the many components of the climate system and, thus, are the principal tools available for making quantitative projections. However, projections of the details about the magnitude, timing, and specific regional impacts and consequences are variable[8,9]. CCSP thus places the highest priority on research aimed at addressing known modeling deficiencies.

Multiscale Modeling Framework

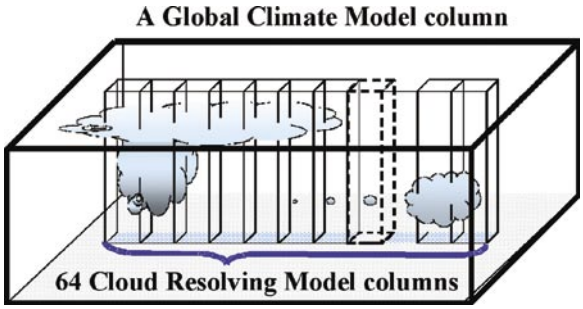
Relative to the goals above, two long-term goals of BER with respect to climate prediction are to:

- *Deliver improved climate data and models for policy makers to determine safe levels of greenhouse gases for Earth*
- *Substantially reduce differences between observed temperature and model simulations at subcontinental scales by 2013, using several decades of recent data.*

8 Intergovernmental Panel on Climate Change. 2001. Climatic Change 2001: The Scientific Basis. Eds. TJ Houghton, Y Ding, DJ Griggs, and M Noguer. Cambridge Univ. Press. Cambridge, U.K.

9 Climate Change 2001 Synthesis Report. 2001. Contribution of Working Groups I, II and II to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Edited by Robert T. Watson and the Core Writing Team. Cambridge Univ Press.

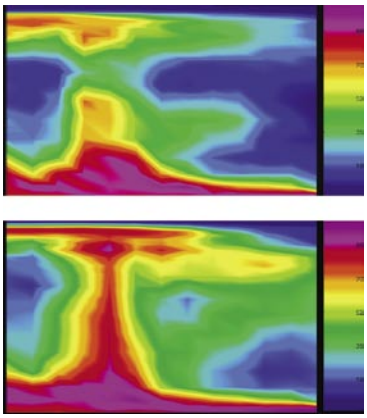
Relative to these goals, the single largest uncertainty in climate models is the role of clouds in controlling solar and thermal radiation onto and away from the earth (see Footnotes 8 and 9). Most current global climate models use a coarse (200-kilometer) resolution to permit multicentury simulations. However, even finer resolution simulations (10 kilometers) do not explicitly resolve cumulus clouds, which are important agents for the vertical transport of heat, moisture, and momentum into the atmosphere; scavenging moisture and pollutants from the atmosphere; and scattering and absorbing solar and infrared radiation. The influence of cumulus clouds on the resolved scales in climate models is parameterized in terms of a handful of resolved variables, but these parameterizations have weak physical bases. This introduces a large source of uncertainty in the simulated response of clouds to climate change, the associated feedback on climate, and climate sensitivity and projection of climate change.



CRM embedded within one grid cell of a global climate model. The cloud model simulates cloud circulation and its influence on microphysics processes for a subset of the global model grid cell.

In the last few years, a new approach to the treatment of clouds has replaced conventional cloud parameterizations in a few models[10,11]. This new approach consists of embedding a cloud resolving model (CRM) in each climate model grid cell. Each CRM explicitly resolves the cloud circulation and its influence on microphysical and radiative processes. The resulting coupled system is often called the Multiscale Modeling Framework (MMF). Preliminary results showing quantitative improvements in climate simulations suggest that the MMF methodology will become the approach of choice for estimating the radiative forcing by anthropogenic greenhouse gases and aerosols, investigating physical feedbacks in the climate system, and understanding the relationship between global climate forcing and regional hydrological changes. It is a computationally expensive solution, increasing the run time of a simulation by a factor of 200 in coarse resolution models. However, because the CRMs dominate the computations and interact with each other only through the outer spatial (coarsest resolution) grid, the MMF scales much better than conventional climate models, permitting efficient parallelization on thousands of processors on some computing systems. The MSCF, with its concentration on chemical and physical processes such as aerosol chemistry, is used to improve and develop the physics and boundary conditions of the model clouds in the CRM.

The current version of the most well-known MMF (which uses the National Center for Atmospheric Research Community Atmosphere Model [CAM] as the global climate model) is a prototype that demonstrates the proof of principle but,



CAM (top) and MMF (bottom) simulations of average relative humidity. The MMF model clearly shows the observed behavior that water vapor is lofted into the atmosphere by deep convection in the Intertropical Convergence Zone (images courtesy of Koontz, PNNL).

10 Khairoutdinov, MF, and DA Randall. 2001. A cloud resolving model as a cloud parameterization in the NCAR Community Climate Model: Preliminary Results. *Geophys. Res. Lett.*, 28, 3617-3620.

11 Randall, DA, MF Khairoutdinov, A Arakawa, and W Grabowski. 2003. Breaking the Cloud-Parameterization Deadlock. *Bull. Amer. Meteorolog. Soc.*, 84, 1547-1564.

for reasons of computational expedience, has five potential weaknesses that are discussed below. Correcting and validating each of these weaknesses will incur a computational cost requiring more resources than are currently available.

- ***One weakness is the coarse resolution of the CRMs themselves. The current four-kilometer horizontal grid size is marginally fine enough to explicitly resolve deep convective clouds but is completely inadequate for shallow convective clouds. During the next few years, simulations with the MMF will involve using a four-fold finer resolution (to one kilometer) for the CRMs, which will run 16 times slower because the CRM time step will also have to be reduced to ensure numerical stability. This alone will not provide adequate resolution of the turbulence that drives boundary layer clouds. Improved treatments of turbulence that can reduce the sensitivity of the boundary layer cloud simulation to the used resolution are available, but add a significant computational burden.***

- ***A second weakness is the two-dimensional geometry of the current CRMs. In the current model, the orientation of the slabs is arbitrary so that the feedback of the clouds on the large-scale momentum budget must be neglected. Preliminary experiments with three-dimensional geometry suggest that this feedback is important in correcting a significant model bias that arises when the momentum flux is neglected. MMF experiments in the next three years are therefore likely to employ three-dimensional geometry in the CRM, albeit for a domain much smaller than that of the global model grid cells. Treatment of the cross terms renders simulations with three-dimensional geometry about 50 percent slower than simulations using two-dimensional geometry and the same number of columns. Enlarging the CRM domain would increase run time proportionately.***

- ***A third weakness is the use of cyclical lateral boundary conditions in the CRMs, which prevents convective disturbances from propagating from one grid cell to the next, but greatly simplifies the coupling between the CRM and the larger scale. By 2007, the MMF will have to treat lateral transport directly between adjacent CAM grid cells during every CRM time step. Although the volume of data communicated between grid cells will be quite modest (10 kilobytes per grid cell), the high frequency of communication will increase costs considerably and thus reduce the scalability of the code.***

- ***A fourth weakness is the crude treatment of cloud-aerosol interactions in the MMF. The influence of aerosols on clouds, which affects both cloud reflectivity and precipitation from clouds, and the influence of clouds on aerosols, which affects their vertical distribution, chemical properties, and lifetime in the atmosphere, are both completely neglected in the MMF. Efforts currently are underway to develop an efficient treatment of these processes for the MMF. Aerosol chemistry is one of EMSL's distinctive science signatures and offers the promise to have significant contributions to the development of these processes within the MMF.***

- ***A fifth weakness is the coarse resolution of the global grid. Current MMF simulations are run using a 300-kilometer grid size for the global model. Simulations of climate change without the MMF are currently run using a 150-kilometer grid size to improve the resolution of mid-latitude storms and topography. Running MMF simulations at that resolution would increase the MMF run time by a factor of two for two-dimensional cloud geometry but a factor of four for three-dimensional cloud geometry.***

High-Performance Computing Requirements for Multiscale Climate Modeling

Climate modeling with computationally intensive models began in the 1960s and has been limited by computing resources during its history. Even development of the improved model, which is the focus here, requires large computational resources for correcting and validating changes in the models. Current global climate models have been parallelized to run on hundreds of processors at coarse (200-kilometer) resolution and thousands of processors at fine (10-kilometer) resolution. Most climate simulations are run at coarse resolution to permit multicentury simulations and to store the resulting data. Fine resolution simulations are severely limited by the lack of data storage ability. Coarse resolution simulations are limited by the cost of communication, which in turn limits scalability to less than 1,000 processors.

Future simulations with advanced and improved versions of the MMF will require enhanced computational resources. Refining the resolution of the CRMs will require at least a 16-fold increase in computing power to achieve the same simulation throughput. Three-dimensional cloud geometry would increase the computational needs by another factor of 1.5. Refining the global model grid size by a factor of two will require another factor-of-four increase in computing power. Thus, future improved MMF simulations will require a 100-fold increase to maintain the same simulation throughput.

Regional climate models are developed and applied to investigate the potential impacts of climate change on snow pack and water resources in the Northwest.

Some improvement in the throughput can be achieved by using a larger number of processors. At the current 2°x2.5°xL26 resolution of CAM, the number of processors is limited to 240. Increasing the number of processors can be accomplished either by increasing the resolution of CAM (which would render the MMF even slower) or by using a different domain decomposition that permits up to 8,000 processors for the same grid resolution. Using a different domain decomposition could thus provide a 30-fold increase in time-to-solution of the simulation. The processor count also could be increased by increasing the number of processors sharing memory on a single node. A factor of 16 is unrealistic, but a factor of 2 to 4 might be achievable. Another factor of two is likely from increased processor clock speed. Thus, a 100-fold increase in throughput is achievable using next-generation microprocessor-based machines.

The amount of memory, however, is not really an issue for climate models. Even the MMF, which is big by climate model standards, requires a total of only 100 gigabytes of memory. Memory latency for the CAM is an issue, limiting the performance of climate models on high-performance computing systems to only five to ten percent of peak theoretical performance. The MMF, on the other hand, was developed for cache-based machines and consumes approximately 90 percent of the computational time for the simulations. Although the standard CAM has been vectorized to run efficiently on vector machines, the MMF version of CAM has not. Further work would be required to develop a vector version of the MMF that can run efficiently on vector systems.

Replacing the cyclical lateral boundary conditions in the CRMs with one that allows propagation between CRMs presents a significant communication challenge to the computing system. The frequent communication requires a low latency interconnect. A significant restructuring of the MMF code will also be required to ensure that message passing between CRMs are only required if the CRMs reside on different processors or nodes (the latter can consist of multiple processors).

Volume data for model history will become an issue at high resolution. Currently, climate modeling history is comprised of 12 gigabytes per simulated day, or 4.4 terabytes per simulated year. It would increase to 50 gigabytes per day (18 terabytes per simulated year) at 1 kilometer CRM resolution. Thus, 20 years of simulations will produce a total of 360 terabytes of data.

3. Recommendations

This section describes the recommendations provided by the MSCF's user community for updating the MSCF's high-performance computing architecture and support infrastructure, based on the science drivers described earlier. More explicit recommendations associated with architecture requirements for the next-generation computer will be provided separately as part of the request for information portion of the procurement process.

Recommendations for the High-Performance Computing Architecture

The future MSCF scientific drivers indicate a paradigm shift from reductionist approaches to system sciences and multiscale models and simulations of systems that accurately represent real scientific problems in the DOE core mission areas of environmental and subsurface chemistry, computational and systems biology, and climate science. System sciences in the three areas discussed in this document could potentially consume all computational and algorithmic resources of the MSCF in the foreseeable future. To advance the biological, chemical, and environmental systems sciences, a *balanced* architecture is needed with respect to processor, memory hierarchy, interprocessor communication, and disk access and storage. A single architecture can satisfy the needs of all of the science areas, although some areas may take greater advantage of certain aspects of the architecture. The aspects of the architecture are discussed briefly below in relation to the associated science areas. Note that many of the requirements for the biological sciences, with the exception of molecular dynamics simulations, are still emerging. Each of the bulleted lists in the sections regarding expert staff, the software stack, and the collaborative environment are arranged in approximate descending priority.

A balanced architecture considers processor architecture and memory hierarchy, interprocessor communication, and disk storage access.

Processor Architecture and Memory Hierarchy

Each of the MSCF science areas requires a CPU with high-performing *floating-point* capabilities. Several of the computational areas will also require significant integer capabilities for addressing memory—but it is anticipated that most of the commercially available CPUs can supply the necessary integer operations. The percentage of peak performance for the chemical sciences, biological modeling with molecular dynamics methods, and environmental systems is strongly driven by the hierarchical memory access, memory bandwidth, and latency. Large amounts of low latency and high-bandwidth memory are needed to account for the 10- to 100-times increase in data set size envisioned by the scientific community in the next five years. Fast memory in combination with large cache and pre-fetching mechanisms will improve scientific application performance. The amount of memory per processor is tightly coupled to the need for local fast disk space and I/O, with memory being the

preferred way to store run-time data. Environmental systems researchers can take full advantage of hardware optimized for the long vectors that are characteristic of the grid-based solution schemes. Several chemistry applications will benefit from vector capabilities, as well. In summary, the MSCF user group provided the following recommendations related to processor and memory requirements:

- ***A CPU with high sustained floating-point performance***
- ***Large memory (with low latency and high bandwidth perceived as important to high sustained floating-point performance)***
- ***Fast hierarchical memory access (i.e., large cache and pre-fetching, again, perceived as important to high sustained performance).***

Interprocessor Communication

High-performance computing today is synonymous with large numbers of processors. Most algorithms used in the MSCF science areas require data to be shared among processors at regular or irregular intervals—requiring efficient interprocessor communication. To obtain efficient communication among a cluster of processors (e.g., a distributed memory machine), an interconnect equipped with high bandwidth and low latency is required. Latency is one of the determining factors for biological molecular dynamics simulations, which require synchronization during every time step. For other computations (e.g., electronic structure computations), high bandwidth is required. For example, NWChem's standard DFT benchmark improved time to solution by approximately 30 percent when the Quadrics interconnect on MPP2 was upgraded to QsNetII. The chemical and biological sciences areas will benefit from interconnect capabilities that enable one-sided communication, facilitating the overlap of computational operations with communications. Environmental systems models tend to communicate in a "nearest neighbor" communication pattern and will benefit from one- to three-dimensional interconnect topologies. Large shared memory processing systems can currently provide low latency and high bandwidth for memory access to small- to medium-sized computing problems. However, existing science problems have demonstrated shortcomings in using current shared memory architectures in terms of the scalability of I/O on single-image operating systems, increasing the complexity and cost of maintaining cache coherency and the physical limitations of the number of simultaneous memory accesses to the same address. In summary, the MSCF user group provided the following recommendations related to interprocessor communication capabilities:

- ***High bidirectional bandwidth and low latency***
- ***One-sided communication and overlap of communication with computation***
- ***Exploration of clustered shared-memory architectures with commensurate I/O capability and operating system architecture.***

Disk Storage and Access

Most of the science areas discussed in this document—specifically those based on high-accuracy chemical sciences methods—require large, temporary, high-speed local disk storage and rapid I/O capability. The amount of local disk storage is tightly coupled with available memory, the latter considered the preferred method for storing data. The environmental systems science area and biological molecular dynamics methodologies require up to hundreds of terabytes of disk space to store associated simulation data. Code performance in the environmental systems science area would improve through the use of efficient parallel I/O to a shared global file system. For example, the CRMs presently require 12 gigabytes of disk space per simulated day, or more than four terabytes per simulated year. To improve model performance, a higher-resolution grid (one kilometer or less) is needed, which would increase the data-storage requirements to 50 gigabytes and 18 terabytes per day and year, respectively. Thus, a 20-year simulation would produce 360 terabytes of data. The full amount of disk space would not be required at run time, although a multi-terabyte shared global file system would be required to perform these simulations. Subsurface modeling also has similar data requirements as those discussed above. In summary, the MSCF user group provided the following recommendations for disk storage and access requirements:

Active storage, the ability of a system to process files as they are being stored, allows the user to extract more out of the data and harness the 75 terabyte capacity of the computer.

- *Large high-speed disk storage*
- *A large shared global file system with reasonably fast parallel I/O.*

Other System Considerations

The level of production of scientific data produced by the three MSCF science areas can reach petabytes. This large quantity of data is driving the need for sufficient long-term storage and backup capabilities. A data archival/storage system, possibly with database support, is required for such high-performance computing research, as the transfer of data to and from storage should be relatively fast.

In addition to the recommended general hardware requirements outlined above, the MSCF workshop and white papers identified additional areas that should be examined when procuring the MSCF's next-generation, high-performance computing system:

- *The balance of the computer should depend on the demands of the algorithms that will be used, and the possible use of special-purpose processors (e.g., FPGA, MD GRAPE, specialized hardware accelerators) should be addressed*
- *Migration to a new platform should be as painless as possible, and code should be able to run on the new architecture in production mode in a short period of time*

The MSCF's unique capabilities—advanced massively parallel computers, sophisticated software, and scientific and technical support experts—are necessary to the future accomplishments of the science critical to addressing DOE's, and the nation's, most challenging environmental problems.

- ***The software environment of the new computer should be stable and equipped with a consistent operating system and compilers, and have a full suite of tools including portable parallel debuggers, performance and profiling tools, and data management and analyses tools.***

The first of these three recommendations already is being addressed in part through monitoring use of the current MSCF high-performance computing system. All user jobs are analyzed "on the fly" by an in-house-developed performance monitoring software tool, NWPerf, that makes use of system hardware counters. This tool monitors memory access, processor use, and disk and network access and use. Results are then used by MSCF staff to analyze and improve user codes and obtain statistics on machine use and any requirements for a future balanced supercomputer. The original intention of NWPerf was to provide a center profiling tool. However, NWPerf has proven very useful in providing system users with performance data, enabling them to increase their software efficiency. More information about NWPerf is provided in the accompanying Supporting Documents.

Recommendations for the MSCF Support Infrastructure

The MSCF is unique since it combines advanced massively parallel computers with sophisticated software specifically designed to run on such computers, and also boasts scientific and technical experts to help use the hardware and software most efficiently. The unique capabilities of the MSCF are necessary to the future accomplishment of the environmental science critical to addressing DOE's—and the nation's—most challenging problems. Therefore, in addition to the previously discussed computer hardware, the MSCF will require:

- ***Expert staff with experience operating large hardware resources, working with scientists with a variety of expertise, and developing and debugging large software libraries and programs***
- ***A robust software stack that includes the operating system, programming tools, and simulation software***
- ***A supportive environment for collaborative use that includes data management, visualization, and fast network capabilities.***

Expert Staff

One area that clearly distinguishes the MSCF from other user facilities is the expert staff who possess knowledge and experience in many areas of high-performance computing, and also are specialists in the MSCF's scientific areas of need. The

MSCF team is dedicated to providing a complete production environment that efficiently and effectively allows researchers to solve large scientific challenges.

The solutions for many of the issues discussed in this document require real collaborations that involve theoretical, computational, and applied scientists, mathematicians, and computer scientists. While the MSCF already provides a plethora of capable expert staff members, the MSCF user group provided the following recommendations for additional and continued support:

- **Operating System**
  - ***Support staff who provide a stable hardware and operating system environment***
  - ***Support staff who provide a stable and robust development environment.***

- **Computational Science, Algorithmic, and Programming Expertise**
  - ***More expert computational science staffing for the Visualization and User Services and Molecular Sciences Software groups to support expansion of the MSCF's user base and software needs***
  - ***Support to provide and facilitate innovative programming approaches***
  - ***Support to provide and facilitate theoretical and algorithmic development.***

- **Software Development**
  - ***Support for existing capabilities and porting and optimizing NWChem and Ecce to maintain the current level of chemistry support***
  - ***Support to incorporate user needs into NWChem (see Appendix D for more information)***
  - ***Support to expand Ecce capabilities in databases, molecular dynamics, electronic structure, and kinetics (see Appendix D for additional information).***

- **Training for High-Performance Computing**
  - ***Support to train software developers in engineering skill areas, such as profiling tools, debugging, testing methodologies, version control, and development of robust scalable applications with the capability to interface with other applications***
  - ***Support to expand users' high-performance computing skills involved in parallel and distributed computing, including the proper use of compiler options, local disk space, parallel I/O options, and memory hierarchies.***

Software Stack

While new hardware innovations are useful, such advances are unlikely to sufficiently impact novel and important science and engineering. On the other hand, new software innovations can potentially have a revolutionary—as opposed to evolutionary—impact on the way science is performed. While much of the research associated with software innovations is, by nature, accomplished outside of the MSCF, the MSCF must provide the environment and expertise that enable and facilitate these innovations, while at the same time provide the production computing environment that leads to new understanding and solutions of scientific challenges.

The MSCF user group provided the following recommendations related to software:

- **Stable and Robust Environment, including Compilers, Debuggers, and Profilers**
  - *Provide an environment and infrastructure conducive to collaborative use of resources*
  - *Validate and provide upkeep of compilers on the platform to ensure accurate results for newly developed and legacy codes*
  - *Provide portable parallel debuggers and performance analysis tools*
  - *Provide a tight coupling between algorithm development and computer/facility upgrades*
  - *Provide portable, parallel programming tools with higher levels of abstraction that insulate software developers from low-level details of message passing, memory sharing, and data coherency, as well as from changes in architectures, operating systems, and compilers.*
- **Full Suite of Mathematical and Computational Libraries (e.g., portable extensible toolkit for scientific computation [PETSc] and BLAS)**
- **State-of-the-Art, Scalable Simulation Software**
  - *Provide state-of-the-art simulation software that is difficult to obtain, maintain, and/or is effective on large resources; software must be optimized not only for efficiency but also for use.*
- **Innovative Programming Models and Approaches**
  - *Provide innovative programming approaches*
  - *Provide application and system software that allow for asynchronous data movement, providing overlap between remote data access communications with computation on local data*

- *Provide frameworks that can accommodate accurate, robust, efficient, portable, scalable, and interoperable component technologies (e.g., adaptive meshing and discretization, nonlinear sparse matrix solvers, transport algorithms, minimization/optimization algorithms, data partitioning, and visualization)*

- *Provide environments that allow loose or tight coupling between simulators (e.g., multiphase flow coupled to reactive transport, or compositional models coupled to geomechanics models)*

- *Provide a software infrastructure that can handle intrinsic complexity (e.g., coupling between length scales in fine steps) rather than attempting to use tools that are designed for systems where coarse graining is effective and appropriate.*

■ **Fault-Tolerant Support at Hardware, System, and Software Application Levels to Counter Effects of the Product "Mean Time Between Failure" for Hardware Components in Large Parallel Supercomputing Systems**

The innovative approaches and programming models are of special concern, as application software continues to adapt to massive computing resources with an increasing number of processors. In particular, the need for simultaneous, multilevel, interactive computing will become even more critical as multiscale simulations become the norm. Multilevel computing essentially is the process of running different tasks on different groups of processors, which may be independent or interacting, where each task may need to be split into subtasks that again run on different subgroups. This processor subgroup approach provides a much more dynamic environment in which to provide simulation capability, but can add significant complexity to the simulations. Solving problems in this type of simulation environment requires extensive knowledge and experience in programming approaches.

Several algorithms would greatly benefit from the scalability achieved through the use of subgroups. For example, full characterization of the potential energy landscape of large systems (e.g., biological systems and nanoclusters) would greatly benefit from computations based on a combination of parallelism with replica models (i.e., similar independent computations with different input geometries). Parallelism is necessary to address systems large enough to realistically represent pieces of material, and replica models are necessary to efficiently scan the huge amount of competing minima that typically characterize the extremely rugged potential energy surfaces common to these (and other) systems.

Another example is the parallelization of both the modeling domain and data volume in new three-dimensional imaging algorithms, which will lead to a significant upscaling in the size of imaging problems that can be addressed. This can be accomplished by distributing the data to different sets or groups of processors. Further, a distributed copy of the modeling problem will reside within

each processor group. Such a scheme is highly parallel, where global communication among the various data processor groups occurs several times per inversion iteration in order to complete several dot products.

Finally, the coupling of multiple physics models, such as the linkages among groundwater, surface water, and the atmosphere in the climate analyses, will require very adaptive programming approaches to not only couple the physics but couple the different data and communication environments of the software involved.

Collaborative Environment – Data Management and Visualization

Ultimately, the computing environment that is provided to MSCF users must be robust, user friendly, and facilitate scientific discovery. This involves not only the computational resources for simulation, but also the collaborative environment necessary to manage, visualize, and share data. All areas of simulation are at a point where simulations are either producing or using vast quantities of data. For example, atmospheric and subsurface model developers are anticipating generation of hundreds of terabytes, if not petabytes, of data in the next five years. Writing this data efficiently requires not only an efficient high-bandwidth file system, but also the necessary resources to mine and visualize the results.

Atmospheric and subsurface model developers are anticipating generation of hundreds of terabytes, if not petabytes, of data in the next five years.

To curate the vast data and analysis resulting from increased experimental and computational capabilities, integrated *data storage and retrieval* infrastructure and algorithms will need to be developed. This includes novel high-performance hardware-, software-, and operating system-level approaches to storage and data access. Standards for data representation and quality control would have to be developed to ensure that multiple investigators generating the data, and the myriad of users querying the data, know what reasonable inferences can be made from the stored data. It is also possible in principle to use the results of query algorithms to direct future or adjunct queries in the same session, improving performance and database usability.

Continued effort needs to be placed in developing tools and algorithms for analysis and visualization of these large data sets, with special emphasis on integration into the overall process of scientific discovery. For instance, discovering the answer to important biological questions requires a researcher to retrieve new sequence information from a given species and then compare that new data set to multiple species data sets in a whole genome or proteome comparison, yielding an enormous amount of information. Understanding the underlying chemical processes also will require the identification or simulation of the chemical pathways and trajectories involved in the systems of interest. This multilayered information will have to be represented to the researcher in such a way that meaningful results can be obtained. The ability to perform this comparison and distill the results into something meaningful and manageable in a coherent and integrated process will significantly enhance the ability of researchers to identify and address novel hypotheses and contribute to leadership-class science.

In addition, large amounts of information are best grasped visually, particularly when one is trying to discover patterns in expression data, a complex interaction network, or three-dimensional model features. For example, when expression data from multiple experiments is overlaid on even a moderately sized molecular interaction network, the resulting image can become cluttered and difficult to manage. Better tools for graph layout and visualization will help users navigate and annotate these networks. Tools to perform pattern discovery could help simplify these graphs by giving the user the option of "collapsing" groups of molecules that act in concert.

Users often will be working at locations distant from high-performance computing facilities. Therefore, fast network connections and efficient protocols for transmitting data as objects will be important.

The MSCF user group provided the following additional recommendations that would improve the collaborative environment of the MSCF:

- **Data Management**
  - *Open repository for computational results.*
- **Visualization**
- **Network Bandwidth**
- **Software Interfaces to Increase Productivity**
  - *Availability of user interfaces, making the software accessible to non-expert users*
  - *User interfaces for remote access (e.g., quick visualization to check ongoing computations through an Internet interface)*
  - *Better methods for disseminating scientific knowledge*
  - *Better interfaces for using software tools to broaden the community that has access to and uses them. The most powerful interface would be one that has a repository of information concerning previous simulations that would be available to assist in designing simulations*
  - *Tools for computational steering of jobs.*

Summary of Recommendations

The following table captures the essence of the recommendations discussed in the preceding pages. The reader is strongly encouraged to review the entire Recommendations section to more fully understand this summary.

Hardware Needs[1]	B	C	E
Memory hierarchy (bandwidth, size, and latency)		X	X
Peak flops	X	X	X
Overlap computation, communication, and I/O	X	X	
Low communication latency	X	X	X
High communication bandwidth		X	
Large memory	X	X	
High I/O bandwidth		X	
Increasing disk storage needs (size)	X		X
Postprocessing	X		X
Expert Staff Needs			
Operating system			
Computational science, algorithmic, and programming expertise			
Software development			
Training for high-performance computing			
Software Needs			
Stable and robust environment, including: compilers, debuggers, and profilers			
Full suite of mathematical and computational libraries			
State-of-the-art, scalable simulation software			
Innovative programming models and approaches			
Fault tolerance			
Collaborative Environment Needs			
Data management			
Visualization			
Network bandwidth			
Software interfaces to increase productivity			

1 The Xs denote areas of critical importance to the algorithms associated with the three science areas discussed in Section 2.

B = Biology

C = Chemistry

E = Environmental Systems Science

Appendix A: List of White Paper Authors

Thomas Ackerman
Atmospheric Science and Global Change Division
Pacific Northwest National Laboratory
ackerman@pnl.gov

Damian Allis
Department of Chemistry and
Keck Center for Molecular Electronics
Syracuse University
dgallis@syr.edu

Gordon Anderson
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
gordon@pnl.gov

Edoardo Aprà
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
edoardo.apra@pnl.gov

Doug Baxter
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
douglas.baxter@pnl.gov

François Boucher
Département de Chimie-Biologie
Université du Québec à Trois-Rivières, Qc. Canada
francois_boucher@uqtr.ca

René Corrales
Chemical Sciences Division
Pacific Northwest National Laboratory
rene.corrales@pnl.gov

Deborah Diamond
Department of Microbiology
University of Washington
ddiamond@u.washington.edu

Dave Dixon
Department of Chemistry
University of Alabama
dadixon@bama.ua.edu

Matt Fitzgibbon
Department of Microbiology
University of Washington
ma2t@u.washington.edu

Alessandro Fortunelli
IPCF-CNR (Italy)
fortunelli@ipcf.cnr.it

Bruce Garrett
Chemical Sciences Division
Pacific Northwest National Laboratory
bruce.garrett@pnl.gov

Steve Ghan
Atmospheric Science and Global Change Division
Pacific Northwest National Laboratory
steve.ghan@pnl.gov

Mark S. Gordon
Ames Laboratory and
Department of Chemistry
Iowa State University
mark@si.fi.ameslab.gov

Michael Green
Department of Chemistry
City College of New York
green@scisun.sci.ccny.cuny.edu

Wei Gu
Center for Bioinformatics
Saarland University (Germany)
w.gu@bioinformatik.uni-saarland.de

Bill Hase
Department of Chemistry and Biochemistry
Texas Tech University
bill.hase@ttu.edu

Volkhard Helms
Center for Bioinformatics
Saarland University (Germany)
volkhard.helms@bioinformatik.uni-saarland.de

Pavel Hobza
Institute of Organic Chemistry and Biochemistry
Academy of Sciences of the Czech Republic
pavel.hobza@uochb.cas.cz

Bruce S. Hudson
Department of Chemistry and
Keck Center for Molecular Electronics
Syracuse University
bshudson@syr.edu

Toshiko Ichiye
Department of Chemistry
Georgetown University
ti9@georgetown.edu

Ahren Jasper
Department of Chemistry
and Supercomputing Institute
University of Minnesota
jasp0029@umn.edu

Don Jones
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
dr.jones@pnl.gov

Michael Katze
Department of Microbiology
University of Washington
honey@u.washington.edu

Annette Koontz
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
annette.koontz@pnl.gov

Danny Letourneau
Département de Chimie-Biologie
Université du Québec à Trois-Rivières, Qc. Canada
danny_letourneau@uqtr.ca

Peter Lichtner
Earth and Environmental Sciences
Los Alamos National Laboratory
lichtner@lanl.gov

Benjamin Lynch
Department of Chemistry and
Supercomputing Institute
University of Minnesota
lync0059@ umn.edu

Roger Marchand
Atmospheric Science and Global Change Division
Pacific Northwest National Laboratory
rog@pnl.gov

Greg Newman
Earth Sciences Division
Lawrence Berkeley National Laboratory
ganewman@lbl.gov

Shuqiang Niu
Department of Chemistry
Georgetown University
sn72@georgetown.edu

Chris Oehmen
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
christopher.oehmen@pnl.gov

Gayla Orr
Chemical Sciences Division
Pacific Northwest National Laboratory
gayla.orr@pnl.gov

Mikhail Ovtchinnikov
Atmospheric Science and Global Change Division
Pacific Northwest National Laboratory
mikhail.ovtchinnikov@pnl.gov

Hrvoje Petek
Department of Physics and Astronomy
University of Pittsburgh
petek@pitt.edu

Monty Pettitt
Department of Chemistry
University of Houston
pettitt@uh.edu

Tjerk Straatsma
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
tps@pnl.gov

Don Truhlar
Department of Chemistry
and Supercomputing Institute
University of Minnesota
truhlar@umn.edu

Lai-Sheng Wang
Department of Physics
Washington State University
ls.wang@pnl.gov

Bobbie-Jo Webb-Robertson
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
bobbie-jo.webb-robertson@pnl.gov

Bob Williams
Department of Biomedical Informatics
Uniformed Services University of the Health Sciences
bob@bob.usuhs.mil

Angela Wilson
Department of Chemistry
University of North Texas
akwilson@unt.edu

Appendix B: List of Workshop Participants

Invited speakers and white paper presenters

Edoardo Aprà
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
edoardo.apra@pnl.gov

Dave Dixon
Department of Chemistry
University of Alabama
dadixon@bama.ua.edu

Bruce Garrett
Chemical Sciences Division
Pacific Northwest National Laboratory
bruce.garrett@pnl.gov

Steve Ghan
Atmospheric Science and Global Change Division
Pacific Northwest National Laboratory
steve.ghan@pnl.gov

Michael Green
Department of Chemistry
City College of New York
green@scisun.sci.ccny.cuny.edu

Wei Gu
Center for Bioinformatics
Saarland University (Germany)
w.gu@bioinformatik.uni-saarland.de

Bill Hase
Department of Chemistry and Biochemistry
Texas Tech University
bill.hase@ttu.edu

Peter Lichtner
Earth and Environmental Sciences Los Alamos
National Laboratory
lichtner@lanl.gov

Chris Oehmen
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
christopher.oehmen@pnl.gov

Steve Yabusaki
Natural Resources Division, Hydrology Department
Pacific Northwest National Laboratory
yabusaki@pnl.gov

Jin Zhao
Department of Physics and Astronomy
University of Pittsburgh
jiz38@pitt.edu

Participants

Alex Aceves-Gaona
Washington State University, Tri-Cities
alex.aceves@pnl.gov

Doug Baxter
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
douglas.baxter@pnl.gov

Gary Black
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
gary.black@pnl.gov

Bill Cannon
Media and External Communications Department
Pacific Northwest National Laboratory
cannon@pnl.gov

Wibe de Jong
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
wibe.dejong@pnl.gov

Ram Devanathan
Chemical Sciences Division
Pacific Northwest National Laboratory
ram.devanathan@pnl.gov

Jincheng Du
Chemical Sciences Division
Pacific Northwest National Laboratory
jincheng.du@pnl.gov

Rogene Eichler West
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
rogene.eichler.west@pnl.gov

Steve Elbert
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
steve.elbert@pnl.gov

Andrew Felmy
Chemical Sciences Division
Pacific Northwest National Laboratory
ar.felmy@pnl.gov

Maciej Gutowski
Chemical Sciences Division
Pacific Northwest National Laboratory
maciej.gutowski@pnl.gov

Todd Halter
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
todd.halter@pnl.gov

Susan Havre
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
susan.havre@pnl.gov

Shawn Kathmann
Chemical Sciences Division
Pacific Northwest National Laboratory
shawn.kathmann@pnl.gov

Kevin Kautzky
Media and External Communications Department
Pacific Northwest National Laboratory
kevin.kautzky@pnl.gov

Karol Kowalski
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
karol.kowalski@pnl.gov

Manojkumar Krishnan
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
manoj@pnl.gov

Jun Li
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
jun.li@pnl.gov

Patricia Medvick
Information Sciences and Engineering Division
Pacific Northwest National Laboratory
patricia.medvick@pnl.gov

Jarek Nieplocha
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
jarek.nieplocha@pnl.gov

Sallie Ortiz
Research Directorate Communications
Pacific Northwest National Laboratory
sallie.ortiz@pnl.gov

Mikhail Ovtchinnikov
Atmospheric Science and Global Change Division
Pacific Northwest National Laboratory
mikhail.ovtchinnikov@pnl.gov

Bruce Palmer
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
bruce.palmer@pnl.gov

Kevin Regimbal
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
kevin.regimbal@pnl.gov

Bob Rittenhouse
Chemistry Department
Walla Walla College
rittro@wwc.edu

Annanaomi Sams
Scientific and Technical Information Department
Pacific Northwest National Laboratory
annanaomi.sams@pnl.gov

Greg Schenter
Chemical Sciences Division
Pacific Northwest National Laboratory
greg.schenter@pnl.gov

Karen Schuchardt
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
karen.schuchardt@pnl.gov

Steve Shoemaker
Information Sciences and Engineering Division
Pacific Northwest National Laboratory
steve.shoemaker@pnl.gov

Gary Skouson
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
gary.skouson@pnl.gov

Joanne Stover
Scientific and Technical Information Department
Pacific Northwest National Laboratory
joanne.stover@pnl.gov

Tjerk Straatsma
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
tps@pnl.gov

Kamakshi Sundaram
Process Science and Engineering Resources Division
Pacific Northwest National Laboratory
sk.sundaram@pnl.gov

Marat Valiev
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
marat.valiev@pnl.gov

Erich Vorpagel
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
erich.vorpagel@pnl.gov

Bobbie-Jo Webb-Robertson
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
bobbie-jo.webb-robertson@pnl.gov

Theresa Windus
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
theresa.windus@pnl.gov

Sotiris Xantheas
Chemical Sciences Division
Pacific Northwest National Laboratory
sotiris.xantheas@pnl.gov

Ping Yan
Biological Sciences Division
Pacific Northwest National Laboratory
ping.yan@pnl.gov

Vasiliy Znamenskiy
Department of Chemistry
City College of New York
znamensk@sci.ccny.cuny.edu

Department of Energy Representation

Gary Johnson
Department of Energy
garyj@er.doe.gov

Drew Tait
Department of Energy
Drew.Tait@science.doe.gov

Appendix C: Molecular Science Computing Facility 2004 Annual Report

Molecular Science Computing Facility

The Molecular Science Computing Facility (MSCF) at the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL) in Richland, Washington, supports a wide range of computational activities in environmental molecular research, from benchmark calculations on small molecules to reliable calculations on large molecules, from solids to simulations of large biomolecules, and from reactive chemical transport modeling to multiscale climate modeling. The MSCF provides an integrated production computing environment with links to external facilities and laboratories within the U.S. Department of Energy (DOE) system, collaborating universities, and industry.

Capabilities

The MSCF provides computational resources to support Computational Grand Challenge projects in environmental molecular science and basic and applied research areas that address the environmental problems and research needs facing DOE and the nation. These projects typically involve multiple investigators from universities, national laboratories, and industry working collaboratively in teams and are usually allocated computer time for three-year periods. In 2004, the MSCF supported 15 three-year Computational Grand Challenge projects, providing an average annual computational allocation time of 739,080 central processing unit (CPU) hours. This represents an increase of nearly 320 percent over the previous year and is attributed largely to the new and fully functional High-Performance Computing System-2 (MPP2) computing system.

The MSCF also supports smaller, shorter-term Pilot Projects, which are limited to a maximum of 75,000 CPU hours and a one-year duration, with short extensions occasionally granted for project completion. The MSCF supported 44 Pilot Projects during 2004, with an average allocation of 40,000 CPU hours per project. Pilot Projects are typically directed at developing the basis for submitting a Computational Grand Challenge proposal in the future (e.g., a combination of theory/method or code development activities, or calculations that provide the initial scientific basis of a Computational Grand Challenge proposal).

A total of 424 users used MSCF high-performance computing systems during 2004. Seventy-eight percent of these users were external (not from the Pacific Northwest

Instrumentation and Capabilities

- **MPP2 –** Production cluster of 978 Hewlett Packard rx2600 nodes, 1956 1.5-gigahertz IA64 processors, 450-terabyte local disk, 6.8-terabyte memory, and 11.8-teraflop theoretical peak performance.
- **Lustre –** Shared cluster file system, 53 terabytes.
- **NWfs –** EMSL's long-term data storage, 85 terabytes.
- **Network –** OC12 (600 MBit/sec) internet connection, Gigabit Ethernet MSCF backbone.
- **NWVisus –** Visualization server, Silicon Graphics Incorporated Onyx 3400 Graphics, eight processors, 16 gigabytes of random access memory, two Infinite Reality3 pipes, 144-gigabyte disk, with a PanoramTech three-screen monitor.
- Digital video editing suite
- Access Grid internet node
- **Molecular Science Software Suite –** NWChem, Ecce, GA Tools (formerly known as ParSoft).

National Laboratory [PNNL], on whose campus EMSL resides), and the remaining 22 percent were comprised of PNNL staff, postdoctoral fellows, and students. Figure 1 shows a breakdown of MSCF resource allocation by user affiliation.

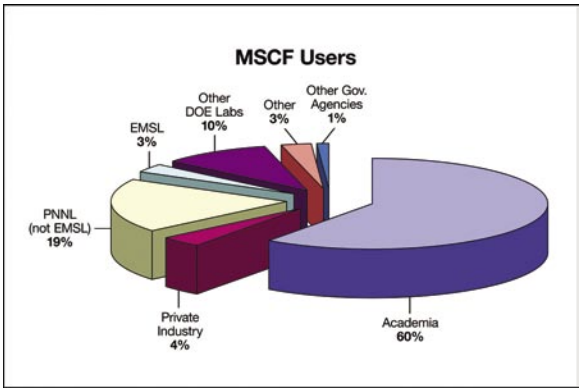


Figure 1. MSCF resource allocation by user affiliation.

MSCF User Computing. To optimally address the complex environmental problems facing DOE and the nation and to best use limited staff resources, EMSL follows the recommendation of its Science Advisory Committee to use a Computational Grand Challenge approach for providing large blocks of computing resources to its user community. A call for proposals is issued annually, and teams of computational scientists respond with peer-reviewable proposals for system time allocations. As stated previously, selected teams are awarded access to MSCF computational capabilities for one to three years, with award based on:

- *Scientific merit*
- *Appropriateness of the proposed method or approach*
- *Relevance to the environmental problems and research needs of DOE and the nation*
- *Technical competence of the investigators*
- *Reasonableness and appropriateness of the proposed computer resources.*

The process used to review Computational Grand Challenge proposals and allocate MSCF computing resources is shown in Figure 2. The request for proposals is open to all researchers, regardless of their institution or source of funding. For reference, the recent call for Computational Grand Challenge proposals involved 22 external scientific reviewers from leading universities and research institutions from around the world. Proposals received two reviews on average, and three reviews in certain cases.

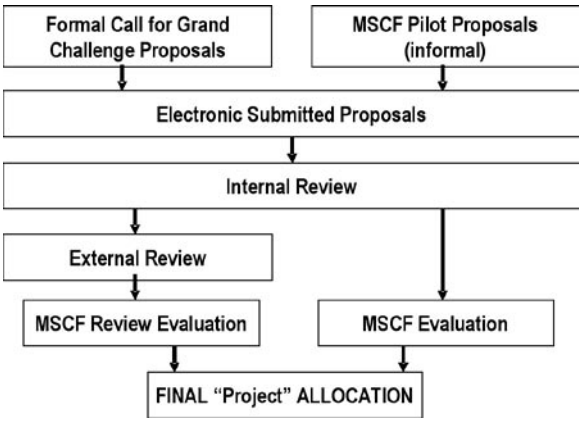


Figure 2. Review process and allocation proposal flow chart.

MSCF Resources. The MSCF provides a combination of production computing hardware and software resources and visualization tools to support the scientific research activities of the Computational Grand Challenges and Pilot Projects. Hardware and visualization resources include the MPP2, the 85-terabyte data storage system (NWfs), the Graphics and Visualization Laboratory (GVL), and the Molecular Science Software Suite (MS3). These resources are discussed below.

MPP2. Since becoming operational in July 2003 with a theoretical peak performance of 11.8 teraflops, 6.8 terabytes of random access memory, and 450 terabytes of disk space, the Hewlett-Packard–designed MPP2 (Figure 3) has been tailored to meet the operational needs of EMSL users.

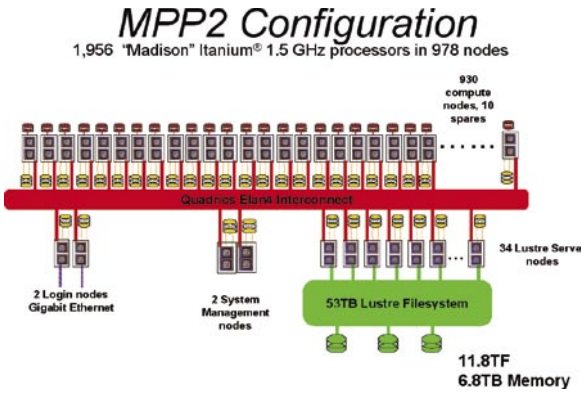


Figure 3. MPP2 configuration.

NWfs. This MSCF archive uses a groundbreaking approach to disk storage involving clusters of many low-cost commodity disks to provide fault-tolerant, high-performance storage. The current EMSL archive is equipped with more than 85 terabytes of storage available to users, with the ability to grow to more than a petabyte of space.

GVL. The GVL provides production graphics and visualization facilities for the display and analysis of complex datasets from both experiments and simulations. It contains four high-performance graphics stations based on Silicon Graphics Incorporated (SGI) technologies with high-speed Gigabit Ethernet connections to the production supercomputers and to NWfs. It is also equipped with a digital video system integrated with the workstations to facilitate the display and capture of scientific data, and digital video editing equipment for preparing scientific presentations. The video system is also connected to EMSL's auditorium and the Internet to facilitate online conferencing. An IBM Scalable Graphics Engine is connected to the MSCF Dell Linux Cluster to provide the highest-performance visualization capabilities.

MS3. Software resources include MS3, a comprehensive, integrated set of tools that enables scientists to understand complex chemical systems at the molecular level. MS3 couples the power of advanced computational chemistry techniques with existing and rapidly evolving high-performance massively parallel computing systems equipped with extensible problem-solving capabilities. The suite consists of three components: (1) Northwest Computational Chemistry Software (NWChem), (2) Extensible Computational Chemistry Environment (Ecce), and (3) the Global Array Tools (GA Tools). These three components are briefly described below.

- ***NWChem.*** Version 4.6 of NWChem was released in June 2004. This software provides many methods for computing the properties of molecular and periodic systems using standard Gaussian and planewave-based quantum mechanical descriptions of the electronic wave function or density. In addition, NWChem can perform classical molecular-dynamics and free-energy simulations. These approaches may be combined to perform mixed quantum-mechanics and molecular-mechanics simulations. NWChem is available on almost all high-performance computing platforms, workstations, personal computers, and clusters of desktop or workgroup servers. Development of the software provides maximum efficiency on massively parallel processors. Documentation and information are available on the NWChem website (http://www.emsl.pnl.gov/docs/nwchem).

- ***Ecce.*** Ecce, which is composed of a suite of client/server UNIX-based applications, is a domain-encompassing, problem-solving environment for computational chemistry. Applications for setting up, running, and analyzing the results of computational chemistry studies are built on top of a web-based data management and inter-application messaging server framework. A computational code registration capability supports several underlying chemistry codes and the ability to integrate new codes without reworking core Ecce applications. Running jobs through industry-standard remote communications, like secure shell, and a batch queue management system registration capability allow transparent access to high-performance computing resources from users' desktop workstations. A simple installation procedure and extensive online help combine to make Ecce a preeminent user environment for computational chemistry. The current production release of Ecce is version 3.2.1. There are six major application components of Ecce:

1. Calculation Manager aids in the organization and manipulation of computational chemistry studies. This tool allows an at-a-glance overview of the status of every calculation and easy access to key setup parameters and run statistics.
2. Molecule Builder is an intuitive point-and-click tool that enables the building, visualization, modification, and manipulation of three-dimensional visualizations of chemical systems.
3. Basis Set Tool enables the choice from more than 245 predefined Gaussian basis sets or the ability to create new basis sets for use in ab initio electronic structure calculations.
4. Calculation Editor allows the user to choose input options using point-and-click interfaces for different chemistry codes, and then generates the code-specific input.
5. Job Launcher is used for submitting a calculation to a computer for processing. The user may submit a calculation to any computer that has been registered with Ecce where the user has an account.
6. Calculation Viewer provides convenient access to current information for a single calculation during execution or after completion. It has many features for viewing and visualizing chemical system properties.

- ***GA Tools.*** GA Tools (also known as ParSoft) includes high-performance computing libraries and tools for applied parallel computing focused on interprocessor communications through the aggregate remote memory copy interface, high-performance input/output (I/O) through the Parallel I/O tools, and programming models for hierarchical memory systems through the Global Arrays and Memory Allocator libraries. The development of these tools is driven by needs of real scientific application codes on high-end parallel systems. Development of Aggregate Remote Memory Copy is supported by EMSL operations and by the DOE Center for Programming Models for Scalable Parallel Computing.

MSCF Organization

The MSCF is organized into three groups: (1) the Visualization and User Services Group (VisUS), (2) the Molecular Science Software Group, and (3) the Computer Operations Group.

VisUS. VisUS provides an extremely diverse set of services for all users of the MSCF high-performance computers and the GVL. Scientists who need access to high-end computing equipment frequently have difficulty getting started, ranging from logging in to getting user codes to run efficiently. VisUS handles user proposal applications, follows user progress during computational projects, manages proposal reviews for both Computational Grand Challenge Projects and Pilot Projects, helps with user accounts, provides general consulting support for MSCF software packages, supports and maintains software, manages the GVL, conducts training and user workshops, develops visualization software and high-quality visualizations, and produces websites.

The Computational Grand Challenge and Pilot Project proposal processes are managed for the MSCF by VisUS, to include proposal receipt and preliminary review to determine applicability to EMSL missions; preparing proposal packets for external peer review; evaluating peer review responses; granting project computational allocation time; and managing the allocation via the MSCF QBank, an open source dynamic reservation-based allocations management system.

VisUS consultants assume various roles, including administrator, tutor, programmer, and research scientist, and field a variety of requests for support. In 2004, five scientific consultants responded to more than 700 email requests, and about 500 additional requests were handled over the telephone or during office visits. VisUS consultants also work directly with MS3 development teams to provide customer feedback and test functionality.

Information about the use and configuration of MSCF computational resources is critical to the user base and is provided efficiently to users via the Internet through the MSCF website (http://mscf.emsl.pnl.gov/). This website contains all necessary information about how to establish accounts and get started, and about computer configurations as well as documentation and web-based tutorials for MS3. Scientists generate enormous amounts of complex data, either from computational resources or through EMSL scientific instruments, using the GVL's high-performance graphics computing servers and state-of-the-art multimedia equipment. The real-time digital video capture capability from the graphics computing servers allows high-quality video production rapidly. Users can generate presentation media in any form—from video (including all international video standards) to web-based animation. VisUS also provides basic video production services.

Molecular Science Software Group. This group is primarily responsible for developing and supporting MS3. This effort includes developing high-performance versions of the software and new high-performance algorithms, responding to user requests for additional features, supporting and maintaining the software, diagnosing problems associated with computer vendor hardware and software, consulting on specific problems related to MS3, distributing MS3 to remote sites, porting software to new architectures, and conducting training and user workshops.

Staff in the Molecular Science Software Group are focused on developing next-generation molecular modeling software for newly evolving computer technologies, especially massively parallel computers. The group includes other staff matrixed from PNNL and is composed of computational chemists and computer scientists (with external collaborations with mathematicians) who work together to develop the software. MS3 is used by many of the Computational Grand Challenge Projects and has been distributed to over a thousand sites worldwide. To facilitate user training needs, this group developed several resources for user interactions, to include MS3 websites containing user and reference manuals, information downloads, release notes, frequently asked questions, a list of known bugs, tutorials, and benchmark information, as well as a support queue for answering direct questions and user email lists. In 2004, the Molecular Science Software Group responded to queue requests, and ultimately the group moved the NWChem support mechanism to the user majordomo list to create a larger community of well-informed users.

Computer Operations Group. This group operates, maintains, and advances the capabilities of MSCF scientific computing systems and is responsible for operating and implementing the various MSCF production supercomputers, with primary focus on providing high-quality, reliable production computing cycles to support extremely large parallel calculations for Computational Grand Challenge Projects. This group has also developed numerous unique system management, monitoring, allocation management, and scheduling capabilities.

Upgrades

NWfs Hardware. In August 2004, the MSCF received new hardware for the second-generation NWfs archive storage system. This hardware provides 380 terabytes (about 389,120 gigabytes) for storage of data primarily generated by EMSL scientific instruments and the MPP2.

RAID5 and RAID6 Arrays. The MSCF contributed code to the Linux 2.6 kernel, enabling creation of Software RAID5 and RAID6 arrays that are larger than two terabytes. Testing of these changes resulted in creation of a six-terabyte array, which uses low-cost Serial ATA disks; it is believed to be the largest Software Raid5/6 array created under Linux. This upgrade will enable low-cost, high-volume data storage to support proteomics and other high-volume projects with great storage needs.

Advanced Storage Technologies. EMSL and its users benefit from a research alliance between PNNL and SGI that is aimed at enabling a new generation of fast and efficient storage technologies for data-intensive computing. The long-term collaboration involves researching options needed for obtaining more than 2.5 petabytes of storage in the next two

years. SGI delivered a single 380-terabyte file system to EMSL as part of the alliance's first phase. More information is available on the PNNL website (http://www.pnl.gov/main/highlights/sgi.html).

MSCF Expansion and Elan4 Upgrade. Preliminary design work on expansion of the MSCF continued through January and into early February 2004. Additionally, the parallel MPP2 supercomputer file system was upgraded to Elan4 to provide users with much faster access to the global file system. Early results of the upgrade demonstrated a sustained "write" rate of more than 600 megabytes per second (previously 200 megabytes per second) by each client to the global file system. This three-fold increase profoundly impacts the ability of the system's clusters to perform disk-intensive operations, such as bioinformatics and indirect electron structure methods.

NWChem Version 4.6. With the upgrade to Elan4 and significant efficiency improvements to NWChem, the MSCF provided users with even more improved computing capability and higher-capacity computing. Figure 4 shows a logarithmic plot of the wall clock time versus the number of processors used to execute the density-functional module of NWChem for the $Si_{75}O_{48}H_{66}$ molecule (3554 basis functions) on various platforms.

The figure illustrates that use of the new Elan4 achieved significant improvement (25 to 45 percent using a moderate numbers of processors) in time-to-solution by upgrading the network and also clearly show the excellent performance of the MPP2. In addition to hardware changes, significant modifications were made to the grid partitioning capability of the density functional code and in pre-fetching in the molecular dynamics software to obtain better performance on the MPP2.

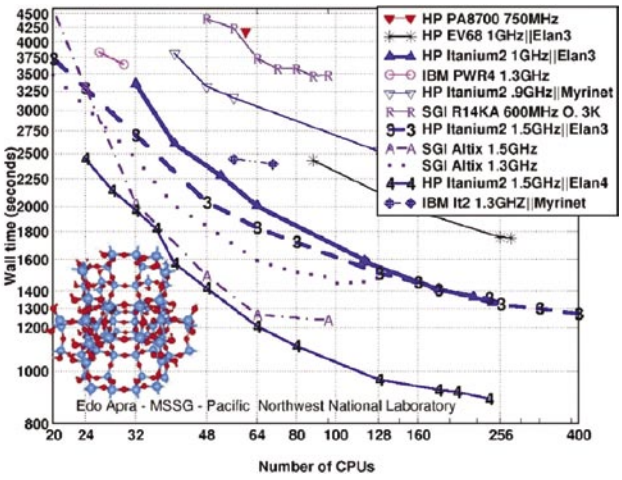


Figure 4. Wall clock time as a function of the number of processors used on various platforms for local density approximation calculations of 3,554 basis functions.

In addition to the changes made to the code to enable efficient, large scientific computations, new many-body methods using the tensor contraction engine (TCE) have been added with the new release of NWChem: CCSD(T) and CCSD[T] for closed- and open-shell systems with Abelian symmetry; EOM-CCSD, EOM-CCSDT, and EOM-CCSDTQ for excitation energies, transition moments, and excited-state dipole moments of closed- and open-shell systems; and CCSD, CCSDT, and CCSDTQ for dipole moments of closed- and open-shell systems.

TCE is a code developed in Python that produces a set of tensor equations from mathematical equations for many-body wave functions represented in normal-ordered second quantization. The tensor equations are further manipulated to produce a parallel implementation in software. In other words, TCE is an automated code-generating mechanism to produce highly scalable software in an expedited manner. The TCE effort is collaborative with other computational chemists and computer scientists located at Ohio State University, Oak Ridge National Laboratory, the University of Waterloo, and the University of Florida. The PNNL portion of the TCE effort is supported both through EMSL and the DOE Office of Science's Office of Basic Energy Sciences.

Significantly, large simulations were performed this year using the MPP2, and modifications were required to the software to enable high performance. For example, using NWChem and the MPP2, MSCF staff were able to compute a CCSD(T) energy of octane to achieve improved computational thermodynamics of the system. The (T) part of the

computation required 23 hours on 1,400 processors and achieved 75 percent of the peak CPU capability–an unusual efficiency improvement for large simulations.

In addition, a capability development project was initiated to provide the EMSL and NWChem user community with a new property module that greatly enhances the ability to calculate newly available experimentally observable properties: multipoles up to octupoles, electron and spin density, electrostatic potential, electric field, electric field gradient, spin-dipolar and Fermi-contact for hyperfine splitting, nuclear magnetic resonance shielding, indirect spin-spin coupling, and Mulliken population analysis.

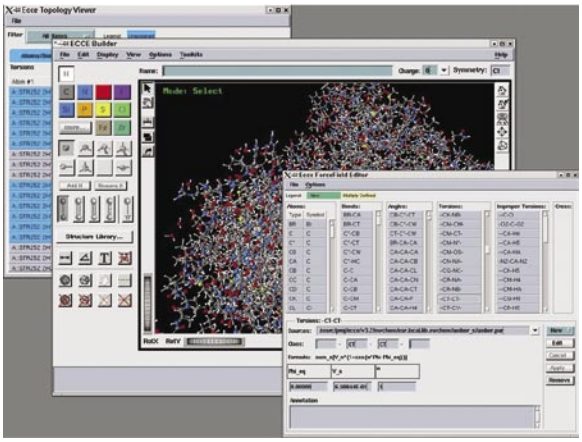


Figure 5. Snapshots of the Ecce Force Field Editor, Builder, and Topology Viewer (foreground to background, respectively).

Ecce. Ecce underwent two production releases in 2004: version 3.2, released in May, and version 3.2.1, released in July.

One hundred and twenty-two sites downloaded Ecce version 3.2 during its three months of availability, and another 185 sites downloaded version 3.2.1 between July and the end of the year. Since version 3.1 was released in mid-August 2003, more than 440 different sites downloaded Ecce production releases. Featured highlights of Ecce version 3.2 are highlighted in Figure 5 and include:

- Completion of the Builder molecular-dynamics toolkit Force Field Editor, which allows users to edit and combine files that use the NWChem force-field format
- Completion of the Builder molecular-dynamics toolkit Topology Viewer, which allows users to assess whether available force fields are sufficient to cover systems being developed for molecular-dynamics simulations
- Builder DNA toolkit for creating segments of double-stranded DNA
- Gaussian 03 support (Gaussian 98 also remains supported)
- Packaging of NWChem binary distribution with Ecce
- Support for sharing calculation data by allowing users to grant user "read-write" or "read-only" access to project folders
- A simplified and more robust installation procedure.

Highlights of Ecce version 3.2.1 include:

- A Builder Constraint/Restraint toolkit to freeze the value of bonds, angles, and torsions at specified values during geometry optimizations
- Builder isotope mass editing through the geometry table for overriding the default most-abundant isotope

- A Builder quantum mechanical/molecular mechanical toolkit for setting up electronic structure calculations in a field of fixed-point charges
- NWChem 4.6 support
- Gaussian cube file support for the automatic detection and analysis of files created as part of a calculation run
- Macintosh OS X computer server support for running Ecce jobs
- Online help updated with all new version 3.2 and version 3.2.1 features.

Specific details about versions 3.2 and 3.2.1 are available at http://ecce.emsl.pnl.gov/docs/release/2864B-RV-32.pdf/ , while general information concerning Ecce can be found at the Ecce homepage (http://ecce.emsl.pnl.gov).

In August 2004, Ecce was used for the first time ever in an undergraduate classroom. Professor Matthew Asplund, Brigham Young University, Salt Lake City, Utah, incorporated Ecce and NWChem into the curriculum of his physical chemistry class. This milestone is significant in the development history of Ecce; the software has reached such a high level of robustness, reliability, and ease of use that it was chosen to replace an established commercial product. A number of the new features introduced in Ecce version 3.2.1 were developed in collaboration with Professor Asplund. More details about these upgrades to Ecce can be accessed at http://www.emsl.pnl.gov/docs/inbriefs/ecce20040618.pdf.

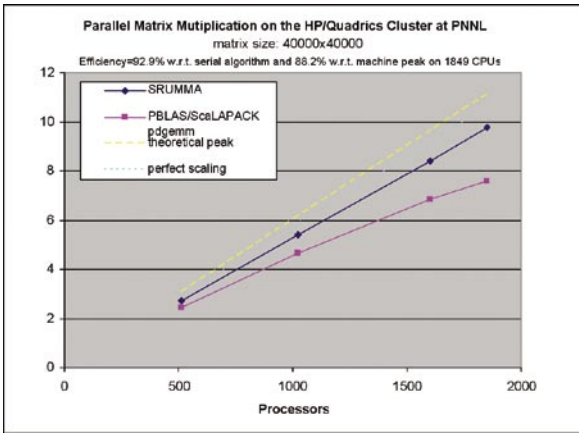


Figure 6. Optimized matrix multiplication (SRUMMA) performance.

GA Tools. Much of EMSL's work on GA Tools has concentrated on porting, tuning, and performance issues for the MPP2 and on new algorithms required for improved performance. Additional efforts were directed toward bringing prototyped research capabilities into a production code.

Version 3.3 of the GA Tools was released in May 2004 and was made available for download from the GA Tools website (http://www.emsl.pnl.gov/docs/global/). Additionally, in July, a beta version of GA Tools 3.4 was created that includes support for processor groups. Processor groups will provide an opportunity for exploiting multilevel parallelism in applications and improving overall scaling on large processor configurations. In addition to maintenance, testing, and user support, specific new capabilities developed for GA Tools include:

- Non-blocking, one-sided operations on the Linux/Elan4 systems (MPP2)
- Optimized matrix multiplication (SRUMMA), which supports rectangular and transposed matrices (Figure 6)
- Re-implementation of the Aggregate Remote Memory Copy protocol stack for Elan4 to improve handling of noncontiguous data types and maximize overlap of communication with computations
- Optimized port for the Mellanox Infiniband Network.

Future Direction

In 2005, the MSCF will be a stable production facility focused on accomplishing large, impactful science, including support for MSCF Computational Grand Challenges and Pilot Projects, as well as EMSL Grand Challenge Projects and continuing support and development of the MS3 software capabilities.

MSCF staff will continue to build advocacy and awareness of the capabilities available to the scientific user community through:

- Creation of informational products (e.g., research animations, brochures, fliers)
- Presentations at national meetings
- Organization of symposia in national meetings
- Development of tutorials
- Establishment of an EMSL Distinguished Fellow who is initially associated with the MSCF.

Staff

Visualization and User Services

Theresa L. Windus, Staff Scientist, Acting Technical Lead
(509) 376-4529, theresa.windus@pnl.gov

Donald R. Jones, Staff Scientist, Technical Lead (January – May 2004)
(509) 376-3013, dr.jones@pnl.gov

Bettina M. Foley, Administrator
(509) 376-2767, tina.foley@pnl.gov

Toni P. Quackenbush, Administrator (January – February)
(509) 376-2767, Toni@pnl.gov

Doug J. Baxter, Senior Research Scientist
(509) 376-3751, douglas.baxter@pnl.gov

Wibe A. de Jong, Senior Research Scientist
(509) 376-5290, wibe.dejong@pnl.gov

Jun Li, Senior Research Scientist
(509) 376-4354, jun.li@pnl.gov

Chris S. Oehmen, Senior Research Scientist
(509) 376-1481, christopher.oehmen@pnl.gov

Erich R. Vorpagel, Chief Scientist
(509) 376-0751, erich.vorpagel@pnl.gov

Operations

Kevin M. Regimbal, Acting Technical Lead
(509) 376-4598, kevin.regimbal@pnl.gov

R. Scott Studham, Technical Lead (January – September 2004)
(509) 376-8430, scott.studham@pnl.gov

Lisa G. Hobson, Administrator
(509) 376-2744, lisa.hobson@pnl.gov

David E. Cowley, Senior Research Scientist
(509) 376-9181, david.cowley@pnl.gov

Evan J. Felix, Senior Research Scientist
(509) 376-1491, evan.felix@pnl.gov

Kevin M. Fox, LTE Post Bachelors
(509) 376-4465, kevin.fox@pnl.gov

Brandon H. Hayden, LTE College Student
(509) 376-1493, brandon.hayden@pnl.gov

Scott M. Jackson, Senior Research Scientist
(509) 376-2205, scott.jackson@pnl.gov

Cindy Marasco, Senior Research Scientist
(509) 376-1241, cindy.marasco@pnl.gov

Ryan W. Mooney, Senior Research Scientist
(509) 376-3590, ryan.mooney@pnl.gov

Kenneth P. Schmidt, Technician
(509) 376-4178, kenneth.schmidt@pnl.gov

Gary B. Skouson, Senior Research Scientist
(509) 376-5401, gary.skouson@pnl.gov

Nathan D. Tenney, Research Scientist
(509) 376-1493, nathan.tenney@pnl.gov

Tim A. Witteveen, Senior Research Scientist
(509) 372-1363, timw@pnl.gov

Ryan P. Wright, Research Scientist
(509) 376-3502, ryan.wright@pnl.gov

Molecular Science Software

Theresa L. Windus, Staff Scientist, Technical Lead
(509) 376-4529, theresa.windus@pnl.gov

Jessica M. Foreman, Administrator
(509) 376-3412, jessica.foreman@pnl.gov

Yuri Alexeev, Postdoctoral Research Fellow
(509) 376-5152, yuri.alexeev@pnl.gov

Edoardo Aprà, Senior Research Scientist
(509) 376-1280, edoardo.apra@pnl.gov

Gary D. Black, Senior Research Scientist
(509) 375-2316, gary.black@pnl.gov

Wibe A. de Jong, Senior Research Scientist
(509) 376-5290, wibe.dejong@pnl.gov

Todd O. Elsethagen, Senior Research Scientist
(509) 375-4431, todd.elsethagen@pnl.gov

Mahin T. Hackler, Scientist
(509) 376-2746, mahin.hackler@pnl.gov

So Hirata, Senior Research Scientist
(509) 376-6751, so.hirata@pnl.gov

Jarek Nieplocha, Staff Scientist
(509) 372-4469, jarek.nieplocha@pnl.gov

Bruce J. Palmer, Senior Research Scientist
(509) 375-3899, bruce.palmer@pnl.gov

Lisa A. Pollack, Postdoctoral Research Fellow
(509) 376-2023, lisa.pollack@pnl.gov

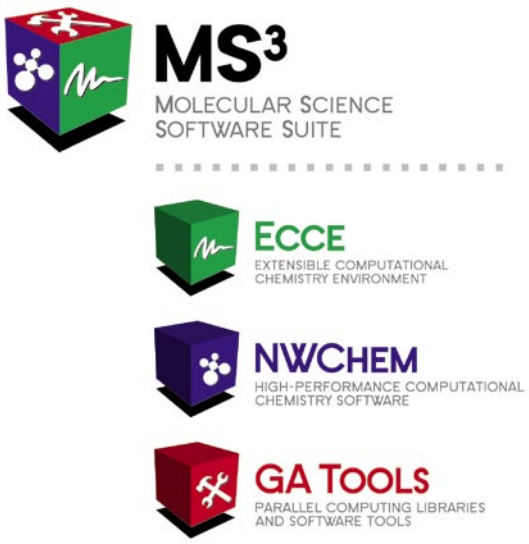
Marat Valiev, Senior Research Scientist
(509) 376-2514, marat.valiev@pnl.gov

The MSCF would also like to acknowledge the contributions of Eric J. Bylaska, Karol Kowalski, Manojkumar Krishnan, Steven W. Matsumoto, Michael C. Perkins, Kenneth A. Perrine, Michael R. Peterson, Karen L. Schuchardt, T. P. Straatsma, Lisong Sun, and Colleen Winters.

Appendix D:

User Needs for Molecular Sciences Software Suite Development/Improvement

During the scientific computing workshop held in December 2004, the Molecular Science Computing Facility user group who defined the future science directions driving chemical research recommended that new functionality be added and improvements made to elements of the Molecular Science Software Suite (MS3). These recommendations are listed (in no particular order) for the in-house-developed Northwest Computational Chemistry (NWChem) and Extensible Computational Chemistry Environment (Ecce) software. While no specific recommendations were made for the Global Array Tools software, improvements and capability development to that software will be required to support upgrades of NWChem.



The following are recommendations for NWChem:

- Ensure that core components of NWChem perform at optimal speeds on the next-generation MSCF architecture
- Optimize the open-shell coupled-cluster code
- Incorporate analytic first and second derivatives for higher-order ab initio methods
- Develop improved methods for converging high-spin system wave functions, if available
- Incorporate high-performance semi-empirical methods
- Make spin-orbit coupling generally available, starting with density functional theory (DFT) and time-dependent DFT methods
- Include new and improved exchange-correlation functionals in DFT
- Include "linear scaling" (e.g., local correlation) for high-level correlation methods
- Implement improved continuum solvation approaches beyond "conductor-like screening models"
- Incorporate higher-order derivatives to obtain accurate zero-point energies
- Improve approaches for excited states
- Implement multireference methods
- Expand plane-wave DFT capabilities
- Develop a module for checking errors in input files and suggesting plausible corrections
- Develop direct dynamics methodologies
- Automatically create potential energy surfaces and force fields from direct dynamics or other sampling approaches

- Develop surface hopping methods
- Seamlessly integrate rate theories with electronic structure
- Provide access to potential mean force methods for condensed-phase systems
- Develop the ability to extend molecular dynamics with new force fields
- Improve and simplify the installation process.

The following are recommendations for Ecce:

- Implement stereo capability for viewing and manipulating structures
- Incorporate support for molecular dynamics and plane wave capability
- Integrate Khimera capability, a unique multipurpose program for chemical kinetics and thermodynamics analysis
- Improve and simplify the installation process
- Create an open repository to store computational results and combine it with data mining tools
- Develop an interactive, three-dimensional manipulation of interior structures (possibly immersion graphical visualization or virtual reality).

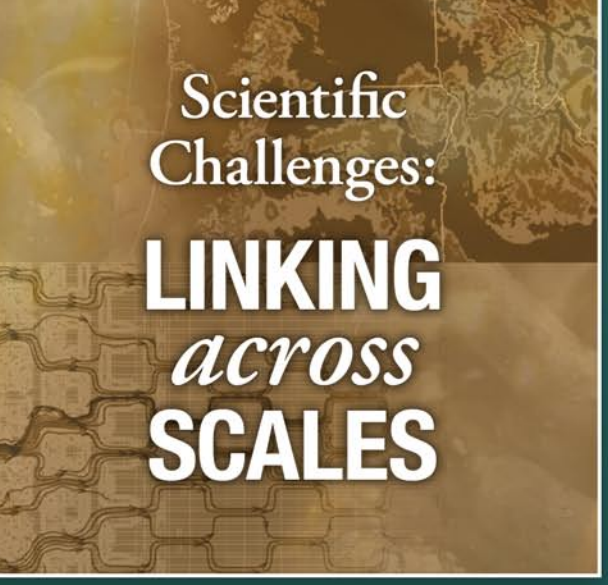
Resources are needed so that the Visualization and User Services and the Molecular Science Software Groups can continue to provide expert support to the expanding user base and added capabilities of NWChem and Ecce. The current level of support provided to chemistry researchers cannot decrease. In addition, adequate resources must be provided for porting and optimizing NWChem on new architectures and to support the incorporation of user needs into the NWChem software.

Appendix E:

List of Supporting Documents

The following documents are contained in the accompanying CD:

- White Papers
- Raw Materials from the December 2004 Scientific Computing Workshop (at EMSL)
- NWPerf Documentation



Pacific Northwest
National Laboratory
Operated by Battelle for the
U.S. Department of Energy

Office of
Science
U.S. DEPARTMENT OF ENERGY

William R. Wiley Environmental Molecular Sciences Laboratory

MOLECULAR SCIENCE COMPUTING FACILITY

Pacific Northwest National Laboratory