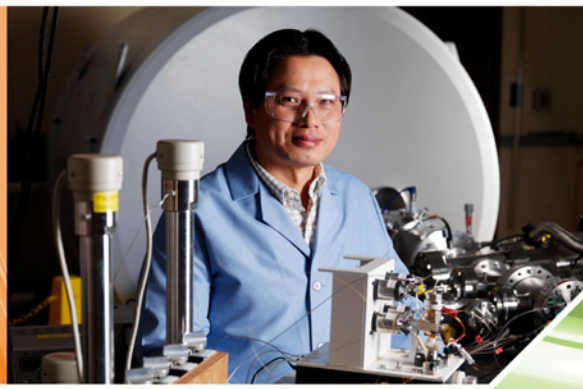
Environmental Molecular Sciences Laboratory

Molecular Science Computing: 2010 Greenbook



EMSL

DISCLAIMER

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

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 not printed on paper, rather it was distributed on CDROM and as a PDF file available on the EMSL web page.

(3/2010)

PNNL-15739 Rev.2

Molecular Science Computing: *2010 Greenbook*

WA de Jong
DE Cowley
TH Dunning, Jr.
ER Vorpagel

March 2010

Prepared for the U.S. Department of Energy's Office of Biological and Environmental Research under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Executive Summary

Advanced high-performance computing (HPC) resources are crucial to enable world-class fundamental research with the increasingly predictive, system-level simulation tools necessary for breakthrough discoveries addressing complex environmental science challenges facing the U.S. Department of Energy (DOE) and the nation. It is widely recognized that complex scientific problems are solved using multidisciplinary science, which demands the close integration of cutting-edge experiments and simulations, as well as instantaneous access to data coupled with essential analysis and visualization capabilities.

As a national user facility, the William R. Wiley Environmental Molecular Sciences Laboratory, or EMSL, provides its users with a tightly integrated scientific environment in the environmental molecular sciences. To accelerate scientific discovery and technological innovation in environmental molecular sciences, EMSL features essential transformational experimental and HPC capabilities that enable unique cross-discipline approaches to complex scientific problems. EMSL's research is focused around its three Science Themes: Biological Interactions and Dynamics, Geochemistry/Biogeochemistry and Subsurface Science, and Science of Interfacial Phenomena.

The Molecular Science Computing (MSC) capability provides EMSL users with HPC capabilities, EMSL-wide data storage, and expert staff focused on large computational scientific challenges in environmental molecular sciences. With its focused mission and equipment, MSC is tightly integrated with the experimental capabilities in EMSL.

This *2010 Greenbook* outlines the science drivers for performing integrated computational environmental molecular research at EMSL and defines the next-generation HPC capabilities that must be developed in the MSC to address this critical research. The EMSL MSC Science Panel used EMSL's vision and science focus and white papers from current and potential future EMSL scientific user communities to define the scientific direction and resulting HPC resource requirements presented in this *2010 Greenbook*. The Computational Grand Challenge to accomplish predictive, systems-level science is achieving realistic modeling of the multiple length scales (nanometers to kilometers) and time scales (picoseconds to millennia) involved. The continued development of efficient and accurate methods that couple phenomena from one size and time domain and pass relevant parameters to an adjacent size and time domain is paramount to successful multiscale modeling. This is germane to all the sciences areas outlined in this document.

In order for EMSL's scientific user community to continue addressing complex system-level environmental science challenges, the MSC capability must maintain fully integrated, leading-edge HPC resources. The availability of environmentally focused and optimized HPC capabilities means significant scientific discovery supporting DOE's missions can be accomplished. Next-generation parallel computing capabilities (with an order of magnitude increase in capability), fast data transfer and storage, scalable codes, algorithm development, and consulting all are crucial components of the scientific discovery process.

The science drivers outlined in this *2010 Greenbook* show that computing at EMSL will need to undergo a qualitative shift, providing high-performance capabilities for both model-driven and data-driven computation. The MSC must continue to support model-driven computing and should strive for, at least, an order of magnitude increase in model-driven computing capability. It is clear MSC will need to support analysis of large and complex data sets as an aid to both computation and experiment. This data-driven approach is a different type of computation altogether from MSC's "traditional" model-driven approach. It will require new architectural approaches and, likely, different hardware. However, it will need to be well integrated with and complementary to model-driven science.

In consideration of HPC needs, the Science Panel pondered the advent of new disruptive computing technologies that could deliver one to two orders of magnitude speedup in performance over conventional technologies and have the potential to affect scientific research fundamentally by removing time-to-discovery barriers. New hardware technologies, such as complex multi-core and General Purpose Graphical Processing Unit (GPGPU) processors, are the way to accomplish this. The latter will require a significant investment in simulation software to provide full use of the new technologies. In addition to computing hardware, it is essential that MSC continue to provide a world-class, integrated production environment with dedicated expert staff and leading-edge software that enables the computing capability to be used effectively, with ease and the fastest time-to-solution.

Ultimately, bringing next-generation HPC capabilities to the MSC capability will improve EMSL's scientific ability to deliver predictive, system-level research and accelerate its goals for breakthrough scientific discoveries generated by integrated multidisciplinary research to solve the critical environmental science challenges facing the world today.

Acknowledgements

The *2010 Greenbook* reflects the vision of the EMSL Molecular Science Computing Science Panel and contributions by white paper authors from the universities and national laboratories listed in Appendices A and B. The editors wish to thank all who provided such invaluable input to this document. Special thanks goes to EMSL Science Advisory Committee member Thom Dunning, Jr. for co-chairing the Science Panels and the leads of the three sub-panels for the development and writing of the "Molecular Science Computing Science Drivers" chapters of this document: Jeffry Madura (Duquesne University) and Tjerk Straatsma (PNNL) for Biological Sciences, Larry Curtiss (Argonne National Laboratory) and Ram Devanathan (PNNL) for Chemical Sciences, and David Dixon (The University of Alabama) and Glenn Hammond (PNNL) for Environmental Sciences.

The editors also wish to thank the PNNL staff who provided support in the production of the document: Charity Plata for technical editing, Mike Winter for developing the template layout, and Nathan Johnson for providing the cover. The following staff members were instrumental in the organization and success of the Science Panel held at EMSL: Chris Montgomery, Stacey Henderson, Debbie Krisher, and Donna Eberhart.

Acronyms and Abbreviations

2-D	two-dimensional
3-D	three-dimensional
AMBER	Assisted Model Building with Energy Refinement: a code for modeling the dynamics of biomolecular systems
AMR	adaptive mesh refinement: a mathematical technique for refining the grid in regions where the solution is varying dramatically
ARRA	American Recovery and Reinvestment Act of 2009
ATP	adenosine triphosphate: an organic molecule prevalent in living cells with energy stored in phosphate bonds
BER	DOE Office of Biological and Environmental Research
BLAST	Basic Local Alignment Search Tool: a computer program used to identify regions of similarity between biological protein or nucleic acid sequences
BS	broken-symmetry
BSE	Bethe-Salpeter equation: a relativistic equation for bound-state problems
CCSD(T)	Coupled-Cluster with Single and Double and Perturbative Triple excitations: a highly accurate post-Hartree-Fock computational method for adding back multi-electron wavefunctions to account for electron correlation
CESD	BER Climate and Environmental Sciences Division
CHARMM	Chemistry at HARvard Macromolecular Mechanics: a code for modeling the dynamics of biomolecular systems
CIR	Computationally Intensive Research
CO_2	carbon dioxide
CPU	central processing unit(s)
Cyt	cytochrome: a general membrane-bound protein containing one or more heme groups that carry out electron transport or catalyze redox reactions
DFT	density functional theory: a quantum mechanical theory used in chemistry to investigate the electronic structure of molecules
DME	dimethyl ether
DOE	U.S. Department of Energy
Ecce	Extensible Computational Chemistry Environment: a graphical user interface for quantum chemistry codes such as NWChem
EDTA	ethylenediaminetetraacetic acid: a good chelator of many metal ions
EMSL	William R. Wiley Environmental Molecular Science Laboratory
EOMCC	equation of motion coupled-cluster: a variant of coupled-cluster theory for calculating excited states of molecules
EOMCCSD	equation of motion coupled-cluster singles and doubles: a variant of coupled-cluster theory for calculating excited states of molecules with single and double excitations

EOMCCSDT	equation of motion coupled-cluster triply: a variant of coupled-cluster theory for calculating excited states of molecules with single, double, and triple excitations
EOS	Equations of State: a semi-empirical functional relationship between pressure, volume, and temperature of a pure substance
eV	electron volt: a unit of energy that must be obtained experimentally
FLOPs	floating point operations per second: used to measure computer speed
Gbyte	gigabyte: one billion bytes
GPGPU	General Purpose Graphics Processing Unit(s)
GPU	graphics processing unit(s)
HPC	high-performance computing
I/O	input/output
IR	infrared
JGI	Joint Genome Institute
kcal/mol	kilocalorie per mole
KMC	kinetic Monte Carlo: a computer simulation method that relies on the repeated random sampling of molecular conformations
MBPT	many-body perturbation theory: a technique for evaluating an N-body problem where N goes to infinity
MD	molecular dynamics: the time evolution of a molecular system
MM	molecular mechanical (mechanics): a classical method for simulating a molecule using Newton's laws of motion and treating atoms as balls with mass and bonds as springs
MP2	second order Møller-Plesset theory: a post-Hartree-Fock computational method for adding electron correlation by means of perturbation theory
MRCI	multi-reference configuration interaction: a computational method for including electron correlation by adding multiple eigenstates to represent the excitations of the ground state's electronic configuration
MSC	Molecular Science Computing: a capability in the EMSL
nm	nanometer
NMR	nuclear magnetic resonance
NWChem	Northwest Computational Chemistry Software
ORNL	Oak Ridge National Laboratory
PB	Poisson-Boltzmann: a differential equation that describes electrostatic interactions between molecules in ionic solutions
PDE	partial differential equation(s): a mathematical relation that contains functions of only one independent variable and one or more of its derivatives with respect to that variable
PEMFC	polymer electrolyte membrane fuel cell(s): an electrochemical cell that converts a fuel into an electrical current and where a polymer is used for the membrane separating the anode and cathode compartments
PES	potential energy surface(s): a multidimensional surface used in the theory of electronic states of polyatomic molecules and chemical reactions

PNNL	Pacific Northwest National Laboratory
QM	quantum mechanics: a theory of matter based on the concept of elementary particles possessing both wave and corpuscular properties
RC	reaction centers: (this document) the place within an enzyme catalyst where the chemical reaction takes place
REV	Representative Elementary Volume: (in hydrogeology) the smallest volume over which a measurement can be made that will yield a value
$scCO_2$	supercritical carbon dioxide: refers to CO_2 in a fluid state while also being above both its critical temperature and pressure
STOMP	Subsurface Transport Over Multiple Phases: a computer model for simulating subsurface flow and transport
TST	transition state theory: a theory of reaction rates where reactants pass through a high-energy, short-lived transition state before forming products
TW	terawatt: one trillion watts
U.S.	United States
U(VI)	uranium(VI): the oxidation state of uranium with six electrons formally removed

Contents

1.0 Introduction

Molecular Science Computing (MSC) is a key capability of the William R. Wiley Environmental Molecular Science Laboratory (EMSL), a national user facility located at Pacific Northwest National Laboratory (PNNL). EMSL's mission is to support the needs of the U.S. Department of Energy (DOE) and the nation by providing integrated experimental and computational resources for discovery and technological innovation in the environmental molecular sciences. Today's most challenging scientific problems can only be solved using multidisciplinary science, which demands the integration of transformational experiments and high-performance computing (HPC) simulations, as well as instantaneous access to data with unique analysis and visualization capabilities. EMSL's distinctive focus on tight integration of research capabilities, as well as collaboration among disciplines, yields a strong, synergistic scientific environment that enables unique approaches to scientific problems, leading to high-impact scientific discovery.

The MSC capability houses a high-performance supercomputer, data storage and archives for all capabilities at EMSL, additional computational resources, and expert staff—all tailored to tackle large computational scientific challenges in environmental molecular sciences. With its focused mission and equipment, MSC forms a unique DOE computing facility for environmental chemistry and biology research. MSC is tightly integrated with the experimental capabilities in EMSL. Integration of theory, modeling, and simulation with experiment provides multidisciplinary science teams the essential suite of tools for fundamental transformational research on the physical, chemical, and biological processes that underpin scientific issues of interest to DOE and the nation.

Figure 1.1. The Molecular Science Computing capability is tightly integrated with all other capabilities in EMSL.

EMSL and MSC capability operations are funded by the DOE Office of Biological and Environmental Research (BER). BER's program focuses on world-class fundamental research aimed at achieving a predictive, systems-level understanding of complex subsurface contaminant fate and transport; climate; and biological systems across many spatial and temporal scales, from sub-micron to global, individual molecules to ecosystems, and nanoseconds to millennia. BER-funded environmental research 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 environments. 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 (CO_2) in the atmosphere. BER's current priorities are: 1) the development of biofuels as a major, secure national energy resource; 2) to understand relationships between climate change and Earth's ecosystems and assess options for carbon sequestration; 3) to predict the fate and transport of subsurface contaminants; and 4) to develop new tools to explore the interface of biological and physical sciences

BER's Climate and Environmental Sciences Division (CESD) sponsors EMSL and has a core mission to advance the fundamental science that will lead to solutions for complex environmental problems, such as climate change and subsurface contaminant fate and transport—both keys to supporting the DOE mission. CESD supports an integrated

portfolio of research, ranging from molecular to field-scale studies, with emphasis on the use of advanced computer models and multidisciplinary experimentation.

Research at EMSL is well aligned with DOE's mission and focused on gaining a more thorough, predictive, and system-level understanding of the physics, chemistry, and biology governing environmental processes starting at the molecular scale and propagating to larger scales. EMSL's research focus has been defined into three Science Themes:

Biological Interactions and Dynamics

- Developing a quantitative, systems-level understanding of the dynamic network of proteins and molecules that drive cell responses and how groups of different cells interact to give rise to functional cell communities.

Geochemistry/Biogeochemistry and Subsurface Science

- Studying molecular scale reaction mechanisms at the mineral-water, microbe-mineral, and fluid-fluid interfaces and understanding the effect of these mechanisms on the fate and transport of contaminants.

Science of Interfacial Phenomena

- Developing an understanding and gaining control of atomic- and molecular-level structure–function relationships at interfaces that enable the optimization of interfacial properties, such as the control of catalytic activity and selectivity.

MSC provides its users the computational capabilities to contribute extensively to scientific advances in these Science Themes. It supports a wide range of computational modeling activities, from small model systems to reliable calculations on real-world systems, solids to simulations of large biomolecules, and modeling reactive chemical transport to environmental systems. Results of this research and EMSL's experimental capabilities have combined to serve as a foundation for new science-based solutions to environmental challenges that are critical to DOE and the nation.

The *2010 Greenbook*'s purpose is to define the evolving science drivers and needs for performing integrated computational environmental molecular research at EMSL and provide guidance associated with the next-generation MSC HPC center that must be developed to address this critical research. The *2010 Greenbook* first describes scientific challenges in the areas of biological, chemical, and environmental sciences, then the role that MSC computing and expert staff resources will play, and how the MSC capability and its upgraded computational resources will positively impact both current and future DOE missions.

1.1 Molecular Science Computing Capability

The MSC capability in EMSL is an integrated production computing environment with specifically designed hardware architecture, software resources, and visualization tools to support EMSL Science Themes and accomplish DOE environmental mission goals. To support these goals, MSC has dedicated Operations, Scientific Consulting, and Software Development teams.

To meet the needs posed by the EMSL Science Themes, MSC's current hardware resources are comprised of an HPC system (Chinook), the EMSL data archive (Aurora), and a graphics and visualization resource laboratory.

Chinook is a balanced Hewlett-Packard supercomputer composed of 2310 compute nodes with 18,480 AMD "Barcelona" Opteron processor cores, a theoretical peak performance of 163 teraflops, 74 terabytes of random access memory, 1.25 petabytes of disk, a DDR InfiniBand interconnect, and a Linux operating system. Application software includes the

internally developed Molecular Science Software Suite—consisting of Northwest Computational Chemistry Software (NWChem), Extensible Computational Chemistry Environment (Ecce), and Global Array Tools—as well as ScalaBLAST and various external software packages.

Chinook was tailored to provide balanced performance for chemistry, biology, and environmental science. These scientific fields greatly benefit from data stored on a machine with large memory capability, a high-bandwidth interconnect to scale the modeling calculations, fast local scratch disk on all nodes, and a large (1/4 petabyte) global file system to stage simulation results.

Figure 1.2. The 163-TFlop Chinook supercomputer is housed in the MSC.

Aurora was procured in fiscal year 2009 with American Recovery and Reinvestment Act (ARRA) funding and provides long-term storage and data services for EMSL experimental data and computational results. It has hierarchical storage management (HSM) features, allowing EMSL abundant flexibility to choose appropriate storage technologies (i.e., solid state disk, conventional mechanical disk, or tape) to maintain a good balance of performance, expense, and energy consumption as the archive grows to meet EMSL's data management needs.

In addition, EMSL is connected to PNNL's high-bandwidth optical fiber link from its campus in Richland, Washington to Seattle, Washington, which connects to DOE's Energy Sciences Network (ESnet) and other networks across the nation.

A more comprehensive overview of MSC capabilities is available on the MSC website at http://mscf.emsl.pnl.gov.

1.2 Scientific Impact

The MSC has leveraged the concept of Computational Grand Challenges—projects that address complex, large-scale scientific problems with broad scientific and environmental or economic impacts whose solution can only be advanced by

Figure 1.3. Research performed at the MSC capability has been highlighted on a variety of scientific journal covers.

applying high-performance scientific techniques using large computational resources. A three-year, externally peer-reviewed Computational Grand Challenge project involves a team of researchers from universities, national laboratories, and industry who work closely together. Over the last two years, Computational Grand Challenges have evolved to become Computationally Intensive Research (CIR) projects and are associated with EMSL Science Themes. Today, CIR projects still require teams work together to solve environmentally related problems facing the DOE and the nation, with the modification that teams are also encouraged to use other capabilities at EMSL in those pursuits. Since publication of the *2005 Greenbook* (July 2005), more than 427 articles have appeared in peer-reviewed scientific publications from research derived using a together n MSC computer—more than 25 percent of all EMSL-focused publications in the same time frame. Of those 427 publications, 45 combined MSC computation with EMSL experiments. Clearly, the MSC capability is making a significant scientific impact in the nation's research community. To exemplify the scientific impact generated by research using MSC's HPC resources, various scientific accomplishments from EMSL user teams are highlighted within this section.

Enhanced Remedial Amendment Delivery through Fluid Viscosity Modifications: Experiments and Numerical Simulations

Remediation efforts are often incomplete and laborious because of subsurface contaminants located in hard-to-reach places, such as areas of low permeability in aquifer systems. Using the resources at EMSL, a PNNL and EMSL research team found a way to enhance cleanup effectiveness and efficiency by incorporating the inexpensive and readily available polymer, xanthan gum, into the remediation process. Using flow cell experiments, the PNNL and EMSL research team simulated the remediation of contaminated areas. The team found adding xanthan gum to remediating solutions increased their viscosity, helping deliver the remediating agent to areas in which the contaminant may otherwise have been left behind and increasing the portion of the contaminated volume touched by the remediating agent. Further, a version of the Subsurface Transport Over Multiple Phases (STOMP) simulator—a general-purpose tool developed by PNNL scientists for simulating subsurface flow and transport—that was modified to account for using the polymer in the system accurately predicted the results of the experiments. The modified version of STOMP may be used to predict subsurface remediation performance in similar systems at larger scales, and the xanthan gum additive may prove useful in real-world scenarios, such as cleanup of uranium-contaminated areas. Such advances in remediation research demonstrate the feasibility of an inexpensive alternative remediation technique and of using computational tools to predict the effectiveness of such techniques under real-world conditions. This is an excellent example of **linking theory with experiment at EMSL**. This research was published in the top-10 *Journal of Contaminant Hydrology*.[1]

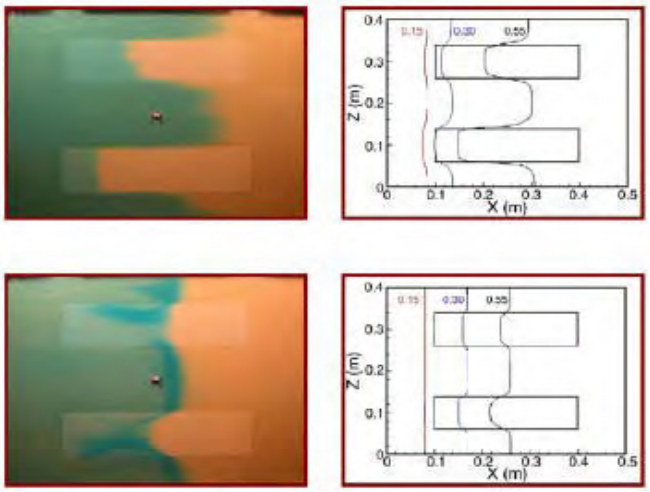


Figure 1.4. Xanthan gum (lower left) helps remediating agents reach areas of low permeability as compared to a control (upper left) in flow cell experiments (permeated volume shown in blue). STOMP predicts the experimental results with impressive accuracy (right).

[1] Zhong L, M Oostrom, TW Wietsma, and MA Covert. 2008. "Enhanced remedial amendment delivery through fluid viscosity modifications: Experiments and numerical simulations." *Journal of Contaminant Hydrology* 101:29-41.

The Catalytic Oxidation of Dimethyl Ether over Pt Surfaces

Dimethyl ether (DME) is of interest because it resembles liquefied petroleum gas and burns with low emissions of NO_x, CO, and volatile organic compounds (VOCs). DME is easily produced from methanol or synthesis gas and has potential use in power generation and radiant heating devices. This study provided detailed kinetic, isotopic, and theoretical evidence for the relevant elementary steps involved in DME combustion on platinum nanoclusters supported on aluminum and zirconium oxide surfaces. Using EMSL's supercomputer to run *ab initio* density functional theory (DFT) quantum chemical calculations, the team from the University of California, Berkeley, and the University of Virginia was able to provide evidence for the relevance of C-H bond activation steps in molecularly adsorbed DME and the role of vacancies within chemisorbed oxygen layers in the catalysis. DFT calculations and experimental measurements led to a consistent mechanistic picture of DME combustion reactions. The research, supported by British Petroleum as part of the Methane Conversion Cooperative Research Program, was published in the top-10 *Journal of the American Chemical Society*.[2]

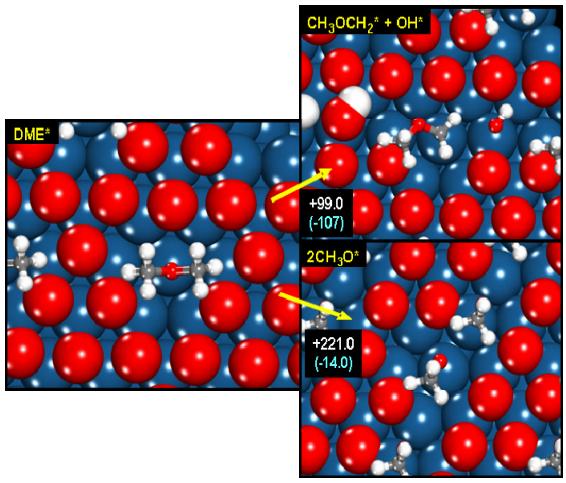


Figure 1.5. Pathways for the activation of DME over an oxygen-covered (red) Pt (111) surface (blue).

Photoelectron Spectroscopic and Theoretical Study: An All-Boron Naphthalene

Engineering technologies to solve energy or security issues benefit from knowledge of the atomic-level structure of materials, such as highly reactive boron. Researchers from Utah State University, Washington State University, and PNNL have discovered the atomic structure of two boron clusters—B_{16}^- and B_{16}^{2-}. Using EMSL's laser vaporization and time-of-flight mass spectrometry capabilities, the research team produced B_{16} clusters and examined them using photoelectron spectroscopy. They conducted theoretical calculations to compare with the experimental data and determined the cluster's structure and chemical bonding. Molecular orbital analysis indicated that B_{16}^{2-} possesses 10 π electrons and a π bonding pattern similar to naphthalene, and it can be viewed as an all-boron version of the aromatic organic molecule used in mothballs. This research provides detailed knowledge about the structure and chemical bonding in B_{16} clusters—information that previously was unknown. This work furthers the scientific basis for the development of novel boron nanostructures. In addition, fundamental insights into the structure of highly reactive boron clusters provide foundational information that other researchers can build on. This information could help the design of new boron-based nanomaterials and exemplifies how EMSL advances our fundamental understanding of matter. The research, supported by the National Science Foundation, was published in the top-10 *Journal of the American Chemical Society*.[3]

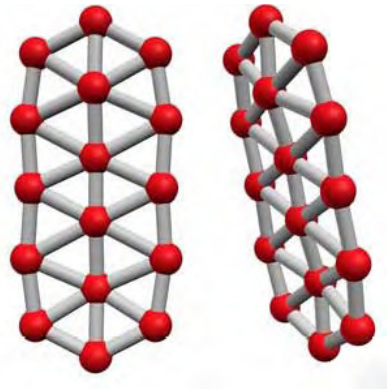


Figure 1.6. Recently, EMSL users discovered structures of two 16-atom boron clusters. The B_{16}^{2-} structure has 10 π electrons and a π bonding pattern similar to naphthalene.

[2] Ishikawa A, M Neurock, and E Iglesia. 2007. "Structural Requirements and Reaction Pathways in Dimethyl Ether Combustion Catalyzed by Supported Pt Clusters." *Journal of the American Chemical Society* 129(43):13201-13212.

[3] Sergeeva AP, DY Zubarev, HJ Zhai, AI Boldyrev, and LS Wang. 2008. "Photoelectron Spectroscopic and Theoretical Study of B_{16}^- and B_{16}^{2-}: An All-Boron Naphthalene." *Journal of the American Chemical Society* 130(23):7244-7246.

Metals with a Complex: Better Bioremediation Through Metal-ligand Complex Studies

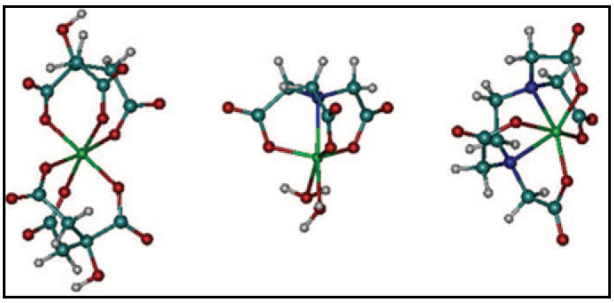


Figure 1.7. Computed structures of Fe-(citrate)$_2^{3-}$ (left), FeOH-NTA- (middle), and Fe-EDTA- (right).

New details about how bacteria and metals interact highlight the importance of considering metal-ligand complexes as part of bioremediation strategies. Bacteria such as Shewanella oneidensis MR-1 hold promise as a bioremediation tool because they exchange electrons with metals, affecting their solubility and level of danger to the environment and human health. Scientists have made significant progress toward understanding electron exchange between bacteria and metals. Led by PNNL, the research team used spectroscopic experiments and computational tools at EMSL to determine the kinetics of electron exchange when the metal (in this case, iron) is coupled to ligands of geological and environmental significance. The research team determined how Fe(III) complexes with the ligands citrate, nitrilotriacetic acid (NTA), and ethylenediaminetetraacetic acid (EDTA) were reduced by two Shewanella surface proteins known to be involved in electron transfer—MtrC and OmcA. The team's results were surprising. Although electron transfer from the surface proteins to the Fe(III) EDTA complex is thermodynamically unfavorable compared to reactions involving Fe(III)-citrate and Fe(III)-NTA, the reduction happened quickly. The team's work demonstrates the importance of metal complexation to bioremediation. For contaminated sediments where radioactive metals are co-disposed with organic chelating agents, any effective bioremediation strategy should take into consideration the ligand complexation effect. These types of experimental and computational studies refine the understanding of the fundamental biological process of bacterial electron transfer and contribute to EMSL's goal to rapidly link theory and experiment. The team's work, published in the No. 1 journal on microbiology, *Applied and Environmental Microbiology*, may lead to enhanced bioremediation strategies to remedy contaminated environments, such as the DOE's Hanford Site in Richland, Washington.[4]

Engineering an Ultra-Stable Affinity Reagent Based on Top7

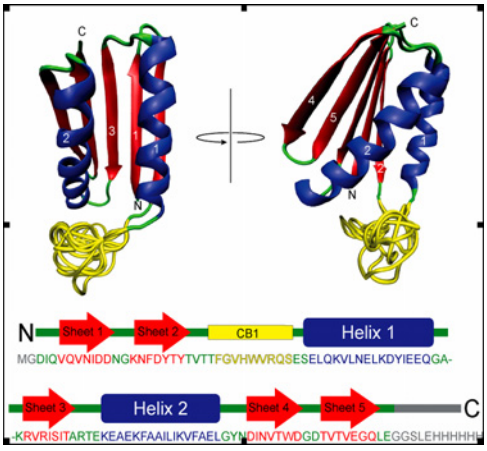


Figure 1.8. Mutant structures were generated by insertion of an 8-residue loop between residues Thr25 and Glu26. The engineered structure of Top7 is displayed as a cartoon.

Antibodies are one of the weapons used to fight disease or detect harmful substances. These proteins are one of the most commonly used reagents in laboratories because they can bind, recognize, and quantify specific targets, such as toxins or proteins that specify different disease states. Antibodies can be generated against virtually any target by immunization or by *in vitro* selection. However, they are large and frequently unstable, which makes them difficult to use for many practical applications. PNNL scientists used a combination of nuclear magnetic resonance (NMR) and computer modeling to design a highly stable antibody alternative. Starting with the synthetic protein Top7 (designed by University of Washington scientists), PNNL scientists added a short peptide fragment from an antibody, PDP-CB1. After a series of modifications, they succeeded in synthesizing a variant that was stable at boiling water temperatures. And, it was able to bind to the

[4] Wang Z, L Shi, C Liu, X Wang, MJ Marshall, JM Zachara, KM Rosso, M Dupuis, JK Fredrickson, and SM Heald. 2008. "Kinetics of Reduction of Fe(III) Complexes by Outer Membrane Cytochromes MtrC and OmcA of Shewanella oneidensis MR-1." *Applied and Environmental Microbiology* 74(21):6746–6755.

protein surface of immune cells to which the full antibody bound. Thus, a smaller molecule alternative to a large antibody with much greater stability was produced. This is another example of EMSL's unique research capabilities advancing scientific discovery and was featured on the cover of the May 2009 issue of *Protein Engineering Design & Selection*.[5]

Scaling Up for Large Metagenomic Computations with ScalaBLAST

The ability to recognize similarity in the DNA code defining particular genes or proteins is a fundamental tool in molecular biology. When similar DNA or polypeptide sequences are observed, a biologist can infer homology, which is a relationship-based similarity. Proteins that are homologous share a common three-dimensional (3-D) structure that often allows their functions to be predicted. This type of evidence underlies much of modern biology. However, there currently are approximately 1,000 microbial genomes completed, which will grow to 10,000 in a few years. Used for decades, the Basic Local Alignment Search Tool (BLAST) algorithm cannot keep up with the rapidly expanding data. Scientists from PNNL and EMSL teamed up with the Joint Genome Institute (JGI) and the Oregon Graduate Institute (OGI) to apply EMSL's supercomputer to do BLAST searches. The result, ScalaBLAST, is a high-performance extension to the National Center for Biotechnology Information's BLAST code. Four significant projects have benefitted from ScalaBLAST:

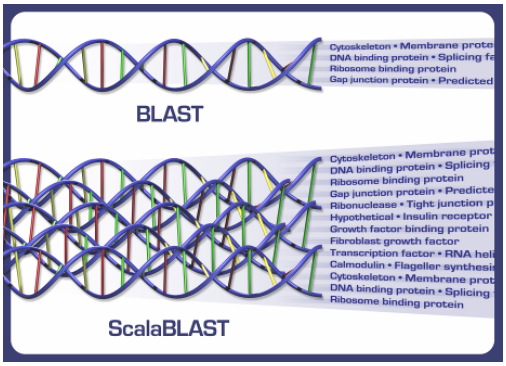


Figure 1.9. BLAST matches DNA or RNA patterns. ScalaBLAST can run the search on thousands of processors.

- **Global Ocean Samples.** Understanding the natural cycling of atmospheric carbon and the interaction of the deep-water microbial community with dissolved organic carbon.
- **Eel River Basin Sea Sediment Community.** This analysis of highly redundant proteins resulted in significant scientific insight into the Eel River Basin community in terms of metabolic pathways the community can engage in.
- **Host-Pathogen Interactions.** Combinations of microbial and mammalian genomes from the mouth, gut, and respiratory membranes were analyzed to study interactions between pathogens and mammals.
- **Microbial Population Structure.** Seasonal variation in an estuary includes salinity gradients, changing temperature, light, river flow, and nutrient availability. The red tide in the Columbia River estuary is an example of complex community response to these conditions. By sampling sites across the salinity gradient, a relationship between microbes present and the nutrient profile led to characterization of seasonal patterns.

Using conventional computing platforms, these analyses would have taken years to complete. Running on the EMSL supercomputer, ScalaBLAST was able to complete the analyses in a matter of days. Details on ScalaBLAST were published in *IEEE Transactions on Parallel and Distributed Systems*.[6]

1.3 Current Science Projects and Use

In the current fiscal year (2010), more than 150 million core-hours have been allocated to 83 research projects on Chinook. These projects span 30-day rapid access proposals with 80,000 core-hours to EMSL Science Theme projects to

[5] Boschek CB, DO Apiyo, TA Soares, HE Engelmann, N Pefaur, TP Straatsma, and CL Baird. 2009. "Engineering an ultra-stable affinity reagent based on Top7." *Protein Engineering, Design & Selection* 22(5):1-8. DOI:10.1093/protein/gzp007.
[6] Oehmen C and J Nieplocha. 2006. "ScalaBLAST: A scalable implementation of BLAST for High Performance Data-Intensive Bioinformatics Analysis." *IEEE Transactions on Parallel and Distributed Systems* 17(8):740-7.

CIR projects with close to 16 million core-hours. More than half of these projects (48) take advantage of EMSL's experimental capabilities as part of their research. In addition, five percent of the resources are allocated for use by recipients of EMSL Intramural projects.

EMSL Science Themes provide strategic direction for critical investments in the development of new technologies to enable innovative research, as well as prioritization of user access. There is overlap between all three Science Themes, and computation is an integral research component in all three. By far, most projects are associated with the Science of Interfacial Phenomena, which encompasses more than half the research projects on Chinook. Biological Interactions and Dynamics and Geochemistry/Biogeochemistry and Surface Science use the other cycles on Chinook.

Several specific examples include:

Biology:

- Interfacial Properties of Biomolecular Systems: Mechanism And Kinetics of Association and Melting
  - In collaboration with an experimentalist at Baylor College of Medicine.
- Quantum Calculations as a Tool in Structural Biology
  - Investigation of ion channels in biological membranes
  - In collaboration with experimentalists at Rush University Medical School and the New York Structural Biology Center.
- A Systematic Molecular Modeling Study of The Effect of Lipid Bilayer Composition on Resistance to Alcohol-Induced Changes
  - In collaboration with experimental research being done at the University of California, Davis and elsewhere.
- Theoretical Modeling of Fluorescence Properties in Biological Systems
  - In collaboration with an experimental group at Temple University.
- Simulating Surface-Mediated Proton Transport At Decorated Surfaces
  - Project is relevant to DOE's research directives, "Materials Sciences" and "Biological and Life Sciences" missions.
- Electron and Proton Transfer Reactions in Photo-Biological H_2 Production
  - Examining the processes that take place in enzyme systems and studying the electron and proton transfers that take place simultaneously during oxygenic photosynthesis.
- Computer Simulations of Protein-Nanomaterial Interactions
  - In support of the experimental data of two groups within EMSL working on the molecular basis of nanoparticle toxicity, recognition by scavenger receptors, and the affect of asbestos (commonly found in nanoparticles) on the epidermal growth factor receptor.
- High-Throughput Sequence Analysis for Communities and Environmental Sample Metagenomics
  - Analyzing experimental data from the JGI and PNNL's proteomics group located in EMSL.

- Development of Accurate Force Fields for Aqueous Biological Environments Using a Massively Parallel Multiscale Approach
  - This project aims to reshape the modeling landscape by providing a large data set based on highest-quality *ab initio* methods.
- Advanced Biomolecular Simulations—Development and Applications
  - A computational collaboration with Jacobs University Bremen, EML Research GmbH, and University of Heidelberg (all in Germany); Universidade Federal de Pernambuco (Brazil); and PNNL.
- Correlation of Structure and Function of Zinc Metalloproteins Via a Combined NMR/Molecular Theory Approach
  - In collaboration between the University of Michigan and PNNL, zinc-containing enzymes are studied using NMR at EMSL and extensive simulations using Chinook.

Chemistry:

- Computational Design of Catalysts: The Control of Chemical Transformation
  - Focus is on developing and using computational methods for accurate thermodynamic properties of catalysts.
  - A combination of seven universities, PNNL, and close collaboration with experimental groups makes this one of the largest and computationally intensive research projects on Chinook.
- Computational Chemical Dynamics of Complex Systems
  - Expands theoretical and computational capabilities for chemical thermodynamics and reactive dynamics, which will benefit future collaboration between experiment and theory in all Science Themes, including the developing theme of atmospheric aerosol chemistry.
- Establishing First-Principle Description of Heavy Fermion Materials
  - Studying rare earth compounds and actinides related to DOE energy initiatives in collaboration with experimental efforts at Lawrence Livermore and Los Alamos national laboratories.
- First Principles Computations of Interfacial Phenomena for Environment-Friendly Catalysis
  - In support of experimental work on catalytic surfaces conducted at PNNL and Argonne National Laboratory.
- Density Functional Theory Studies of Complex Systems: Structure, Dynamics, and Excited States
  - Partially in support of experimental studies on magnesium-based hydrides that have favorable properties for hydrogen storage, as well as experimental and theoretical studies on forces governing metal speciation at interfacial regions.
- Fundamental Studies of Nitrogen Oxide Surface Chemistry: A Model System Approach Combining Experiments and Theory
  - A collaboration between Nanjing Normal University, China; the University of Texas at Austin; and PNNL using multiple capabilities at EMSL to understand the formation mechanisms and stabilities of different NO_x species on metal oxide surfaces.

Environmental:

- Highly Scalable Molecular Scale Software Development for Environmental Sciences
  - A collaboration between Iowa State University and Oak Ridge National Laboratory (ORNL) using computational chemistry codes to develop new algorithms combining *ab initio* methods with statistical methods applicable to formation dynamics of aerosol clusters in the atmosphere; the properties of molecular reactions; and solution structure of ions (actinides and heavy elements) in water and other solutions, including ionic liquids used in "Green Chemistry."
- Mechanistic Modeling of Subsurface Multifluid Flow and Biogeochemical Reactive Transport
  - In support of multiscale experiments in EMSL's subsurface flow and transport experimental laboratory using intact core samples from the Hanford Site.
- Large-Scale Computational Modeling of the Chemical Behavior of Actinide Elements at Interfaces
  - A collaboration between the California Institute of Technology; Vrije Universiteit/Scientific Computing and Modeling, The Netherlands; University of Alabama, Tuscaloosa; Iowa State University; Washington State University; ORNL; and PNNL developing fundamental insight for remediation.
- A Community Modeling Framework to Advance Numerical Treatments of Aerosol Processes
  - Developing numerical representations for including aerosol chemical processes in climate models.
- Complex Processes in Separations, Catalysis, Hybrid Materials, and Nuclear Energy and Waste Management
  - Using experimental results published in the literature, a team from the University of Arizona, Washington State University, and PNNL are examining complicated processes of complex systems related to environmentally cogent issues.

Over the past year, the MSC HPC capability has supported the research of between 260 and 330 users, or about half of the entire EMSL user community. Forty-five of the 104 distinguished EMSL users, as determined by the essential science indicators and endowed chairs, are associated with the MSC capability.

External researchers make up 66 percent of the user community, while the remaining 34 percent is composed of PNNL and EMSL staff, postdoctoral fellows, and students. A distribution of users by affiliation shows that 50 percent of the scientists come from universities, whereas 47 percent of the users are from DOE-sponsored laboratories.

Since acceptance in March 2009, the Chinook HPC system has maintained high availability and utilization, as exemplified by 96.66 percent availability and 80.69 percent utilization in its ***initial 11 months of operation***. Additionally, there has been high usage of the storage capability (NWfs), with approximately 410 terabytes of the more than 1.2 petabytes of disk space available used to store scientific data. This data currently is being migrated to the next-generation archive system, called Aurora, procured with 2009 ARRA funding.

2.0 Molecular Science Computing Science Drivers

Research at EMSL is focused on advancing fundamental understanding regarding biological interactions and dynamics; geochemistry/biogeochemistry, subsurface science, and the science of interfacial phenomena; and the acceleration of scientific discovery and innovation for DOE mission-relevant applications through tight integration of transformational HPC simulations and experiments. The scientific drivers and direction for this research have been established through broad input and support from the scientific community and represent the science directions envisioned by the current and potential future EMSL user community. The EMSL MSC Science Panel, using input from EMSL's vision and science focus, as well as solicited white papers, defined the scientific direction and resulting HPC resource requirements presented in this *2010 Greenbook*.

In October 2009, the MSC issued a call for white papers from a broad scientific audience representing current users, as well as potential users and leaders, in areas of key interest to BER and DOE. In their white papers, scientists were encouraged not only to extrapolate their science toward the future, but to envision new scientific possibilities that have yet to take full advantage of HPC at the EMSL MSC capability. They also were tasked to consider how this would require different computing paradigms, or how these approaches could benefit from new types of architectures. This forward-thinking methodology will help prevent future scientific discoveries from being hampered by inadequate HPC resources, and it will advance the integration of HPC with experimental observations. In total, 26 white papers were received (see Appendix B).

On November 20–21, 2009, the EMSL MSC Science Panel was convened. The 22 panel members (see Appendix A) are scientific experts who comprised three sub-panels in the areas of biological, chemical, and environmental sciences, which are well aligned with the three EMSL Science Themes. Using input from experts, white papers, and stakeholders, the sub-panels developed scientific drivers and computational needs for EMSL's next-generation supercomputing resources and summarized their findings. These summaries form the basis of this document, which describes the science drivers, the associated requirements for enhancing HPC resources at the MSC, and the role a next-generation MSC HPC facility will play within three to five years across the scientific community and for the DOE.

The following subsections address, in more detail, the drivers and challenges encountered in the biological, chemical, and environmental research areas; their impact on DOE environmental missions; their ties to the EMSL focus areas; and the computational resources that will be required to advance the research. The science drivers highlighted by each of the sub-panels show a significant synergy with the EMSL Science Themes and each other.

2.1 Biological Sciences

In investigating solutions to the DOE's challenges in energy, security, and the environment, multiscale computational biology must be both focused in its analyses and elevated in terms of hardware and software capabilities. Specifically, this section discusses scientific areas within multiscale computational biology important to DOE's environmental and energy missions and details the value of exploring complex biological systems using high-performance computational modeling and simulations. It also addresses challenges in developing and translating model-driven applications and the voluminous, data-intensive nature of data-driven applications.

2.1.1 Science Drivers

The biology sub-panel identified two major biological themes—bioenergy and bioremediation—that will drive future complex biological systems studies and influence HPC resources that host such study areas.

Bioenergy

Extensive research is being conducted to understand, exploit, and optimize biological systems capable of providing novel kinds of fuels as an alternative to non-renewable resources such as petroleum, natural gas, and coal. The goal of this research is to reduce greenhouse gas emissions and limit the world's dependence on non-renewable resources. A critical use for such biofuels is for aviation fuel where alternative propulsion systems to combustion are not as optimal as for other forms of transportation. Ethanol, currently extracted from corn grain (starch) and sugar cane (sucrose), is one biofuel considered a primary alternative to non-renewable resources. However, the supply of raw plant materials is limited because optimum growing is prescribed to specific regions or is costly because of the need for water and fertilizer. Stored as carbon as a product of photosynthesis, lignocellulose—a far more abundant renewable natural resource—is a logical alternative. The amount of energy from the sun converted into cellulosic biomass is 10 times the world's consumption. Additionally, an enormous amount of cellulose is produced in municipal and industrial waste, contributing to pollution problems. Therefore, the breakdown to sugars from cellulosic biomass constitutes a logical and environmentally sustainable option as an energy source.

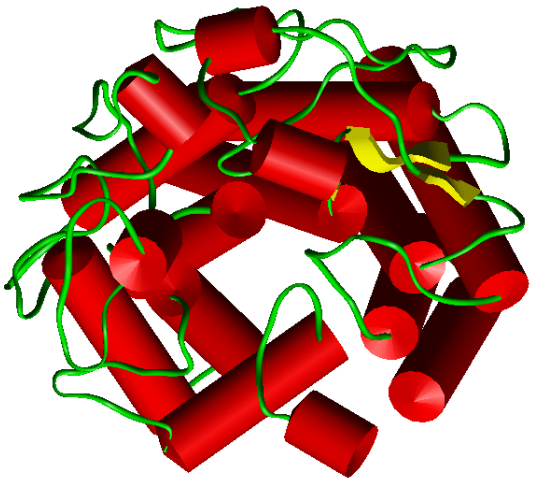


Figure 2.1. A cellulase cartoon from the crystal structure obtained from the Protein Data Bank (1IA6) (*Image courtesy of E. Vorpagel, EMSL*).

Plant cell wall degrading enzymes, or cellulases (Figure 2.1), are produced by a variety of fungi and bacteria. They exist either in complexed systems or cellulosome (anaerobic bacteria and fungi) or non-complexed discrete enzymes (aerobic organisms) secreted into the growth media. Found predominantly in prokaryotes and fungi, these enzymes have been classified on the basis of their mode of action on the substrate into cellobiohydrolases (exoglucanases), which sequentially release sugar units from the reducing and non-reducing ends of the cellulose chain, and endoglucanases, which also attack the chain at internal positions. Other enzymes such as xylanases and β-glucosidases also can attack cellulose without sharp distinction. Among fungi and bacteria, some species, such as *Trichoderma reesei*, *Clostridium thermocellum*, and *Thermobifida fusca*, are capable of secreting complete sets of cellulolytic enzymes that synergistically can completely degrade highly crystalline cellulose. The use of cellulases to convert cellulosic biomass into liquid fuel involves the cost-effective production of efficient enzymes. Large-scale cellulase production no longer seems to be a bottleneck. Through fermentation process development and enzyme engineering, a tenfold cost reduction in the production of cellulases with *Trichoderma reesei* recently was reported, with final enzyme production costs around $0.10 to $0.20 per gallon of ethanol produced. Currently, technical barriers to an efficient enzymatic hydrolysis of cellulosic materials to low molecular weight products (i.e., hexoses and pentoses) include mainly low specific enzyme activity and a general lack of understanding about enzyme biochemistry and mechanistic fundamentals.

Another attractive source of energy from biomass is hydrogen produced by the reactivity of light-harvesting proteins and protein complexes involved in microbial photosynthetic reaction centers (RC). Photosynthesis is a reaction in which light energy is converted into chemical energy. The primary process of photosynthesis is carried out by a pigment-protein

complex embedded in the membrane. In photosynthetic purple bacteria, the cyclic electron transfer reaction is performed by the RC and two other components: 1) the cytochrome (Cyt) bc1 complex and 2) the soluble electron carrier protein. First, the RC accepts light energy from antenna proteins and promotes a light-induced charge separation across the membrane, which results in the oxidation of the special pair and the reduction of quinone to quinol. The quinol molecule then leaves the RC and moves to the Cyt bc1 complex through the quinone-pool in the membrane. Second, the Cyt bc1 complex re-oxidizes quinol to quinone, and the released electrons are transferred to soluble electron carriers. Third, the soluble electron carriers transport the electrons to RC through the periplasmic space. Finally, the photo-oxidized special pair is reduced by the soluble electron carriers, and RC return to the initial state. In the course of the oxidation and reduction of quinone, a trans-membrane electrochemical gradient of protons is formed, which provides the driving force for adenosine triphosphate (ATP) synthesis by ATP synthase and other electrochemical-controlled enzymatic reactions. From the current knowledge of the structure and reactivity of some of the proteins involved in this cycle, it is apparent the elucidation of their molecular-level processes, including the coupled electron and proton transfer steps, requires the application of extensive experimental and computational capabilities. Because these enzymatic reactions combine long-range electron transfer, proton transfer, and conformational protein dynamics for their mechanism of action, new development and applications of advanced computational tools will enable the modeling and simulation of these increasingly complex systems with a reduced level of approximation, an increased level of model sophistication, and an increased level of accuracy.

Bioremediation

Many important processes in the subsurface, such as oxidation/reduction reactions, mineral dissolution, and metal ion precipitation, are microbially mediated and believed to take place at the microbial membrane or the interface between the microbial membrane and mineral surfaces. Because of the complexity and size of these systems, current theoretical understanding of the processes that take place at the interface between biological membranes and geochemical surfaces is limited. With the advent of modern, massively parallel computers and sophisticated, highly efficient software for computational chemistry modeling, the theoretical study of such systems is now feasible. Significant research is focusing on the outer membranes and outer membrane proteins of gram-negative bacteria (e.g., *Pseudomonas aeruginosa* has a gram-negative outer membrane, Figure 2.2). In particular, those found to have the ability to use metal (specifically iron) reduction in the respiratory cycle, take up solvated metal ions from the environment, and show adsorption to mineral surfaces are potentially important target microbes in the design of bioremediation technologies.

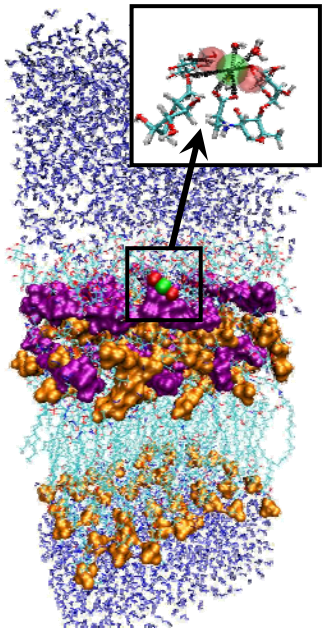


Figure 2.2. Modeling uptake of uranyl cation by the lipopoly-saccharide membrane of *Pseudomonas aeruginosa* (*Image courtesy of R. Lins and E. Vorpagel, PNNL*).

The interaction of microbes with their environment involves a broad range of scales, including the atomic scale of the interaction involving individual functional groups (Å scale), the molecular scale in the formation of long molecular chains (nm scale), the scale of molecular assemblies leading to the formation of membranes (μm scale), and molecular to macroscopic scale of the interactions of membranes with minerals and other surfaces (μm and greater scales). The key geochemical and biochemical interactions and reactions at the interface of microbial membranes transpire across all of these spatial scales. The study of the dynamics and energetics of the interactions between bacterial membranes and mineral surfaces and the process of exchange and uptake by the membrane of charged species from the mineral or from solution is especially complex. Understanding such complex processes at an atomic level of detail requires the integration of theoretical and computational modeling and

simulations with experimental capabilities. These systems consist of a number of complex interacting components: lipid membranes, polysaccharide membranes, trans-membrane proteins, minerals, and solvated ions. Not only do these systems involve many spatial and temporal scales, they also involve complex assemblies and exhibit complex behaviors.

Gram-negative bacteria are characterized by the existence of a cellular envelope consisting of an inner membrane that encloses the cytoplasm and an outer cell wall providing the structural rigidity required for the protection of the bacterium. The abundance of gram-negative bacteria in the environment and their specific interactions with solvated ions and mineral surfaces has prompted many experimental studies of their potential for use in bioremediation. Many metal ion complexes found in contaminated soil and water are potential biohazards because of their toxic and genotoxic effects. For the development of efficient and cost-effective bioremediation strategies, it is crucial to understand the molecular processes occurring at the interface between microbial membranes and their immediate environment. The complexity of bacterial cell walls requires a combination of experimental studies and theoretical and computational modeling investigations to provide insight into the complex interactions and to design additional experiments. The current theoretical understanding of the reactions and adsorption processes mediated by the bacterial cell surface is limited in large part because of the size and complexity of these systems. Novel computational methodologies and implementations are required to increase theoretical understanding of the molecular mechanisms of membrane components that determine metal binding; bacterial attachment to mineral surfaces; and microbially mediated reduction, precipitation, and dissolution processes. Such understanding will be fundamental to increasing knowledge of the bioavailability of reducible soil minerals; microbial transport mechanisms; and the binding, oxidation, reduction, and precipitation of solvated metal ions. This information will be crucial to the design of novel, effective microbial remediation technologies.

2.1.2 Computational Challenges

This section addresses the computational challenges for two distinct application areas, model-driven applications and the data-intensive nature of data-driven applications, as identified by the panel.

Model-driven applications

Broadly defined, biomolecular simulations face numerous computational challenges that include both hardware and software barriers in addition to many theoretical approximations. During the initial era of molecular dynamics (MD) simulations of biomolecules, the barrier was central processing unit (CPU) clock speed and memory size. As the CPU clock speed and memory size increased, the number of particles increased (from 10^2 to 10^6 particles), as did the simulation lengths (from 10^{-12} s to 10^{-9} s). Advances in applied mathematics led to improved simulation details, e.g., Ewald summation rather than large cutoffs. Similar advances (mostly occurring within the past 20 years) have been made in electronic structure calculations involving biomolecular systems.

The picture of protein folding motivating the approach to *ab initio* protein tertiary structure prediction is that sequence-dependent local interactions bias segments of the chain to sample distinct sets of local structures. Nonlocal interactions select the lowest free-energy tertiary structures from the many conformations compatible with these local biases. Different models are used to treat the local and nonlocal interactions. Rather than attempting a physical model for local sequence structure relationships, the protein database can take the distribution of local structures adopted by short sequence segments (fewer than 10 residues in length) in known 3-D structures as an approximation to the distribution of structures sampled by isolated peptides with the corresponding sequences. Hydrophobic burial, electrostatics, main-chain hydrogen bonding, and excluded volume are the primary nonlocal interactions considered. Structures consistent with both local sequence structure biases and nonlocal interactions are generated by minimizing the nonlocal interaction energy using simulated annealing. Advances in the field are limited almost entirely by conformational sampling. Biomolecular

systems have many degrees of freedom, and meaningful calculations must thoroughly explore the relevant conformational space, which can be extremely time-consuming.

Soil is a highly complex heterogeneous system, both in terms of its inhabitant microorganisms and its spatial properties. Soil structure can be thought of as a dynamic hierarchy of aggregates of different sizes. Resident microbial cells may have different metabolic properties. Extracellular environment conditions can have a significant regulatory role on the internal characteristics of the cells, and the intracellular processes dictate the overall cellular responses. Thus, a complete and realistic characterization of multicellular systems requires using models where the intra- and extra-cellular dynamical properties are included and appropriately integrated. With current technology, capturing such details in wet lab experiments is difficult, if not impossible. Therefore, computational simulations form an invaluable complementary alternative to experimental sciences in terms of developing new hypotheses and making predictions about trends in the system dynamics. In addition, developed mathematical models form an excellent platform to integrate the experimental data in a context-dependent manner. The image in Figure 2.3 is from one such simulation and includes multicomponent reactive transport with complex reaction networks involving aqueous complexation, mineral sorption, and biologically mediated reactions. These simulations, though computer intensive, help identify crucial missing data for a more comprehensive understanding of carbon transformation by the biomasses in terrestrial systems.

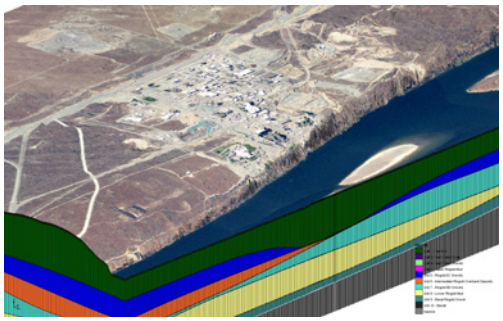


Figure 2.3. Modeling multiscale–multiphase–multicomponent subsurface reactive flows under the Hanford 300 Area (*Image courtesy of G. Hammond, PNNL, and P. Lichtner, LANL*).

In terms of mathematical modeling, modeling of carbon transformation in the soil system by microbial communities translates to having a heterogeneous dynamical network system in which participating objects/agents can have different classes of dynamical properties and the number of the objects can change in time. In addition, as their location and environment are constantly changing, the inner dynamics of the objects/agents also change dynamically. Such changes must be incorporated into the models and taken into consideration during the simulations. As living objects, cells frequently go through a growth-division-death cycle, which changes their numbers as a function of time. Alternatively, through transformations due to exposure to external agents, environmental conditions, or through their regulatory mechanism, the cells may get transformed in such a way that their biological attributes also change considerably. In analogy with physical sciences, the latter phenomenon may be thought of as a phase transformation between the system's possible physical phases. Properties of the biological systems where the number of objects or their characteristics change in time are most suited for the treatment of the system as an agent-based dynamical network. This requires a modeling paradigm shift, and the scientific field recently has started to move toward this goal.

In agent-based computing, dynamical features of multicellular biological systems can be characterized as a dynamical network in which the number of nodes (i.e., the number of cells/objects) varies, while the internal attributes of the nodes also change in time. This leads to a dynamic network-of-networks problem where the interactions between the agents define the network properties. However, as the interaction patterns between the cells change over time, the size and the structure (i.e., topology) of the involved network varies in real time during the simulation of the multicellular systems. This leads to a challenging high-performance computation problem where variations in the topology of the network make it difficult to predict and balance the computational load distributions across nodes. Thus, most successful algorithms should be able to detect how the entire network is evolving through the simulation and adaptively adjust the load assignments accordingly. Estimating from the known properties, a large soil aggregate may have 10^9 microbial cells. As such, the ultimate goal should be to simulate cellular models consisting of 10^{12}–10^{15} cells. Achieving this aim clearly requires new advances in HPC software development and access to the appropriate computing resources to perform the simulations.

Iron-sulfur (Fe-S) proteins are important in bioenergetics, including respiration, photosynthesis, nitrogen fixation, and hydrogen production, and are prime candidates for bioengineering or biomimetic efforts in environmental bioremediation and biofuel production. A wealth of experimental information exists on both Fe-S cluster analogs and proteins. Computational studies have helped to attribute reduction potentials to intrinsic properties of the redox sites, as well as electrostatic effects and hydrogen bonding from the surrounding protein and solvent. Thus, their redox potentials, which provide the driving forces for electron transfer, can be engineered either by substitution of different metals or ligands into the cluster or by site-specific mutations to alter the electrostatic environment of the cluster. Computational methods can guide the engineering since quantum chemical calculations can predict redox energies of iron-containing cluster analogs, and classical calculations combined with bioinformatics analyses can predict the effect of changes in the protein environment (including site-specific mutations) on reduction potentials of these proteins (i.e., assuming the cluster contribution is constant).

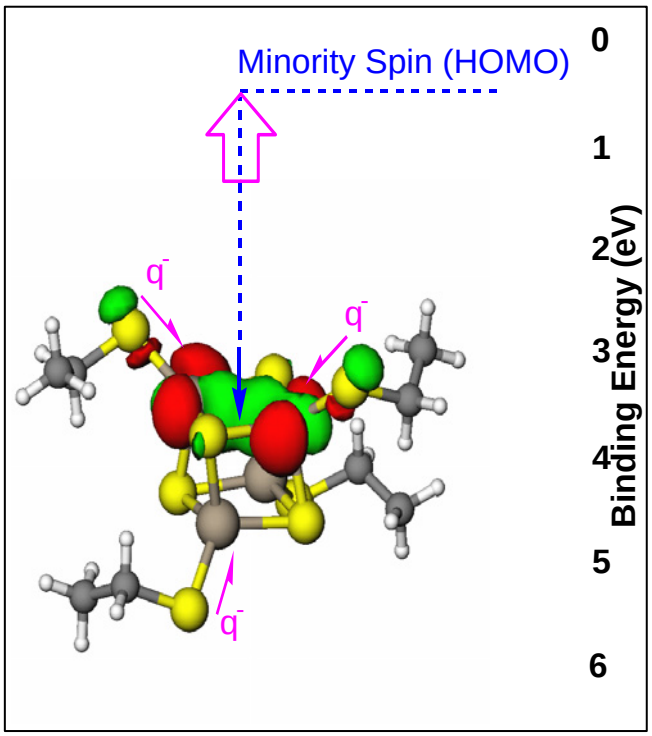


Figure 2.4. Broken-symmetry B3LYP calculations—together with XAS and PES experiments—help unravel complex factors underlying redox energy and chemical bonding of $[Fe_4S_4]$ clusters in Fe-S proteins (*Image courtesy of S. Niu and T. Ichiye, Georgetown University, Washington D.C.*).

However, currently no single calculation can accurately predict the redox energetics of proteins. In particular, combined quantum mechanical/molecular mechanical (QM/MM) methods are limited because the protein force fields cannot provide average structures that are accurate enough, and MD for the MM portion is necessary to get the correct dielectric response, making the calculations prohibitively expensive. Moreover, the QM calculations are nontrivial, i.e., the clusters containing (multiple) transition metals are spin-polarized and require the broken-symmetry (BS) approach in DFT calculations (Figure 2.4). Larger scale multi-reference configuration interaction calculations to eliminate this issue are beyond current computational capabilities.

In the next four years, a goal is to make computational tools that allow a nonspecialized user to predict the redox energetics of proteins containing complex transition metal clusters. For a nonspecialized user, the quantum chemistry calculations of the iron-containing clusters are difficult to perform correctly because of the complex spin problems. They also are computationally quite intensive. Current studies show accurate absolute reduction potentials for proteins can be obtained by simply adding the redox energy of the cluster from QM calculations to the electrostatic interaction energy of the cluster with the surrounding protein and solvent from Poisson-Boltzmann (PB) continuum electrostatic calculations in a "QM+MM" approach. Creating a user tool requires combining a PB solver with a new "transferable" redox energy library of the clusters in the same spirit of transferable force fields such as Chemistry at HARvard Macromolecular Mechanics (CHARMM) and Assisted Model Building with Energy Refinement (AMBER) for biomolecules. Although this seems less sophisticated than a full QM/MM calculation, the input protein structures from the Protein Data Bank can be more accurate than those from MM with current force fields, and results indicate that PB models the dielectric response accurately for this problem.

The development of computational tools for redox properties of Fe-S proteins would be useful for bioengineering these proteins into new or modified functions, as well as designing biomimetics of efficient biological processes. For instance,

respiratory complex "I" contains nine Fe-S clusters, seven of which form a protein "wire" that conducts electrons across the membrane that could be used in bioengineering biocircuits or bio-batteries. Also, they are found in nitrogen fixation electron transport chains and could be used in bioengineering nitrogen fixation under ambient conditions rather than the high pressures and temperatures required by the Haber process. In addition, Fe-S clusters are found in hydrogenases, which catalyze the reversible oxidation of molecular hydrogen. As such, understanding their activity may help in designing clean biological energy sources. Moreover, Fe-S proteins from extremeophile organisms are stable under extreme conditions and may be ideal as starting proteins for bioengineering.

The major stumbling block in developing computational tools for calculating redox properties is the creation of a new "transferable" redox energy library of the clusters. Numerous classical calculations generally show that differences in reduction potential between different proteins with the same cluster are due to the electrostatic environment created by the protein matrix, which indicates the cluster energetics is independent of the specific protein environment. Work with analogs also indicates that environmental factors, such as hydrogen bonds to the cluster, do not affect the cluster properties except by a purely electrostatic contribution, whereas the identity of the ligands and the metals of the cluster, as well as the dihedral angles of the ligands, affect the redox energetics. In particular, the recent discovery of the effect of the ligand dihedral angles is a complication not recognized before, but it provides a new mechanism for redox tuning. Thus, the initial focus will be on calculating the redox energies as a function of the ligand dihedral angles of the standard $[Fe_4S_4Cys_4]$ cluster found in many proteins.

One major issue is to confirm that the library is truly transferable, which requires careful studies of the cluster in various proteins using QM/MM methods. For instance, the $[Fe_4S_4]$ cluster is found in both high potential iron-sulfur proteins (HiPIPs) with a 1-/2- reduction at 100 to 450 mV versus the normal hydrogen electrode (NHE) and the ferredoxins with a 2-/3- reduction at -100 to -645 mV. The QM/MM modules in the NWChem package are ideal for these studies, and close cooperation with EMSL computational staff has been essential in previous work on the cluster analogs. The second issue is that a large number of calculations will need to be performed in order to map out a potential energy surface (PES) of the dihedral angles. The high-performance computers at EMSL will be essential in providing the necessary throughput to perform the multiple calculations.

At present, studies of proton transport are dependent on MD, which uses averaged potentials not adequately accurate in the systems being considered. In addition, standard MD is limited to times <1 μs. This is too short for the biologically and chemically interesting cases that would need to be considered for the purpose of fitting the calculations to either of the Science Themes relating to biological interactions and interfaces that should be addressed. Furthermore, the accuracy of hydrogen bond potentials that omit polarization and variable charge transfer is limited. To fully model the proton transport requires the inclusion of quantum mechanics. *Ab initio* molecular dynamics simulations on large complex proteins at realistic time scales will require much larger computing resources than those currently available. Quantum calculations on parts of proteins have been carried out by several groups. These are more accurate, yet still limited by the size of the system that can be calculated. Only now is it becoming possible to do frequency calculations on systems of the adequate size needed to compare to experimentally observed systems. The frequency calculations needed to obtain room temperature results are limited in part by memory and in part by speed. For several hundred atoms, about the minimum required to do a useful section of a protein, it is not possible to use a large enough basis set to get high accuracy. Currently, roughly 2,000 to 2,500 basis functions seem to be the practical limit (although not the limit in principle). Time (i.e., computer speed) limits not only frequency calculations, but the search for energy minima in systems larger than 500 atoms or much smaller systems for global minima.

However, researchers are within a factor of two to four of being able to do calculations that are sufficient to provide the thermodynamic properties of a large enough section of a protein to describe its behavior with reasonable accuracy. For example, quantum calculations on a section of a protein with 1,000 atoms, with accompanying frequency calculations

even at the second order Møller-Plesset theory (MP2) level, should suffice to allow understanding of a conformational change in response to a change in protonation state. Its importance duly noted, this is the sort of information that can provide detailed understanding of the underlying mechanism of enzyme catalysis. To achieve a calculation of this size, an increase of a factor of roughly 100 in computational speed would be adequate, assuming concomitant memory resources.

In many cases, starting structures without hydrogen atoms are available from X-ray crystallography, and the quantum calculation contributes the orientation of the hydrogen atoms and the hydrogen bonds; the position of water molecules; and reliable energetic, especially for reactions, which are not available from other techniques but are critical for understanding mechanisms. The starting structures help to solve the problem of exploration of conformational space so that the combination of experiment and theory is particularly powerful. NMR structures may also be combined with computations to help understand the protein. Quantum calculations also assist in development of potentials, making MD simulations more effective as well.

Data-driven applications

Due to advanced technologies applied to both experimental and computational efforts, biological research has evolved dramatically in recent years. Biologists previously used labor-intensive methods to study a few genes, proteins, or metabolites at a time. Now, the same experiments can be conducted in a massively parallel fashion with the use of robotics, automation, and microfluidics. These high-throughput experimental pipelines provide many revealing snapshots of dynamic cellular processes as they progress in time and space. However, the use of this information is sharply limited by the data-intensive character of the results. The difficulty is not just that collected data set sources are massive. Other science areas, such as physics and astronomy, have developed successful means for dealing with very large-scale data sets. The key issue is that biological data are both large and complex. This produces significant obstacles in the effort to understand biological systems as a dynamic, integral whole. These challenges are especially apparent in the biological applications described here for 1) large-scale proteomics data, 2) the structure and dynamics of molecular machines, 3) the reconstruction of cellular networks, and 4) the simulation of their dynamic properties. The combination of huge scale and high complexity in systems biology makes this data-intensive science an excellent choice to drive the development of novel computational solutions in both hardware and software.

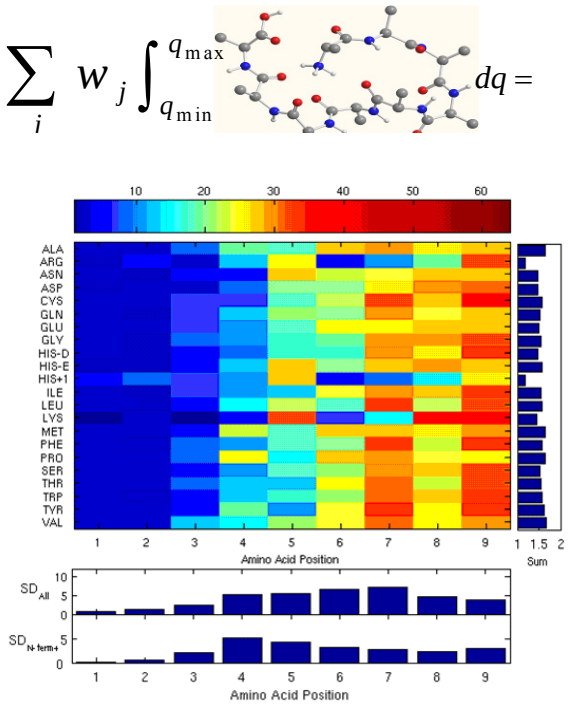


Figure 2.5. Calculating the proteomic state of a cell using statistical inference and statistical thermodynamics for protein identification (*Image courtesy of B. Cannon, PNNL*).

Advanced computational capabilities are emerging that address biological data analysis problems using HPC resources. Some examples include: 1) analysis of mass spectrometry data for whole genome scale proteomics measurements, 2) the structure and dynamics of molecular machines, 3) the discovery of cellular networks and their reconstruction from high-throughput data, and 4) the simulation of the dynamic properties of cellular networks (Figure 2.5). The first two examples, mass spectrometry and dynamic simulation of proteins, are both powerful mechanisms for investigating the molecular machinery of the cell. Mass spectrometry is a primary tool for measuring proteins and protein interactions in the cell. An active area of research is the computational analysis of tandem mass

spectra necessary to identify peptides from these experiments. While the number of spectra generated from a single experiment is small (tens of thousands), searching this spectra against a large protein database presents a significant computational challenge. When searching for post-translational modifications, the search space increases exponentially with the number of modifications. Modeling protein structures and their dynamics provides molecular detail on how they function. Although each MD trajectory for a protein consumes large amounts of disk space, ideally many of them must be compared against each other to extract this structure-function information. These comparisons will require huge data-handling capabilities and will combine the model-driven computational data generation with data analytics to perform time and space trajectory comparisons as the data-intensive aspects of these large sets. Massive amounts of data must be collected, managed, stored, retrieved, queried, and analyzed effectively.

The expanding catalog of completed genome sequences is another example of an exponentially growing data source that will require large-scale computational resources to analyze. The most basic analysis is homology searching by pair-wise sequence comparisons with algorithms such as BLAST, which is heavily used by biologists. These database searches return what can be thought of as a one-dimensional view from the limited perspective of a particular query sequence. However, a new category of bioinformatics analysis capabilities focuses on extracting complex data objects from completed genomes to reveal function using genomic context. To match the rapid growth of the genome databases, the use of these methods will require extensive computational resources. The basis for genomic context analysis methods is the ability to carry out large numbers of BLAST comparisons. The power of these methods is they also can reveal connections between proteins in networks representing metabolic and regulatory pathways. Network topology also can be extracted from other systems biology data sources.

Research needs

Where should research be in the next five years? Recently, there have been several advances in the development of computer technology, such as the introduction of multithreading and GPUs (graphics processing units). Concurrently, there has been development of simulation software to enable the new hardware architecture. Advances in theoretical models are also being realized through these advances (e.g., inclusion of polarization). In the next five years, routine MD simulations will be conducted on systems containing ~10^7 atoms for 1 μs. Routine electronic structure calculations will incorporate 10^3 atoms with 10^4 basis functions for energy and frequency calculations. Routine peptide identification and searching will incorporate searching for post-translational modifications on peptide sequences, as well as search spectra against large collections of proteins from multiple organisms (metagenomic communities).

What kinds of problems will need the types of hardware and software capacities that are of interest to DOE? A few typical scenarios that are of current DOE interest are explored (as follows):

Achieving accurate predictions for the energetics and structural information of proteins containing complex transition metal clusters that are essential in bioenergetics such as photosynthesis, nitrogen fixation, and hydrogen production is beyond the current limits. For example, QM/MM methods and protein force fields cannot provide average structures with enough accuracy. These systems are spin-polarized and require the BS approach in DFT calculations, while the MD for the MM portion is necessary to get the correct dielectric response, resulting in prohibitively expensive calculations. With accurate methods and the essential computing resources, researchers will be able to develop new candidates for bioengineering or biomimetic efforts in environmental bioremediation and biofuel production.

For many bioenergy applications (e.g., fermentation of biomass to alcohols), an understanding of cell membrane interactions with the energy carrier is crucial as these energy carriers (alcohols and others) are toxic to cells and change the membrane structure. There is evidence that different membrane compositions react differently to such toxic stress.

Therefore, it would be beneficial to select or engineer organisms (yeast, *E. coli*, etc.) that have a membrane with a higher tolerance for alcohols (chosen as the example here as it is the best understood).

Increases in the atmospheric concentrations of CO_2 and other greenhouse gases due to the use of fossil fuels are predicted to have major impact on global ecosystems. Sequestration of terrestrial carbon via microbial and chemical reactions in soil can offset some of the fossil fuel carbon emissions. Understanding the flux and residence times of different carbon forms and the biomass will enable greater understanding of the terrestrial carbon transformation processes.

Software considerations

The next generation computer must be able to run standard HPC software programs efficiently, as well as set a direction for their improved performance capabilities by taking advantage of new hardware architecture. Current HPC software packages used in the biological simulation community at the MSC include: NWChem, NAMD, Gromacs, AMBER, CHARMM, and CPMD for model-driven applications and Rosetta, MMC, ScalaBLAST, PolyGraph, and NWLang/NWKsim for data-driven applications. Other important analysis and visualization software used by the bio-simulation community includes Ecce, VMD, Cytoscape, and Starlight.

To further assist in deciding the best state-of-the-art hardware to purchase, bio-simulation software was placed in two broad categories: 1) Model-driven (FLOP centric) and 2) Data-driven (memory/bandwidth driven). Table 2.1 summarizes the various simulation models that fall into these two categories.

Table 2.1. Categories for Bio-Simulation Software

Model-driven (FLOP centric)	Data-driven (memory/bandwidth centric)
Classical MD Simulations: Large, long complex; Free energy	Proteomics
Quantum Chemistry: Energy and frequency calculations; Electron transfer kinetics	Sequence analysis
Hybrid QM/MM: Enzyme catalysis	Network inference
Car-Parinello MD	Image analysis
Quantum Dynamics for electron and proton transfer	
Network Simulations	Statistical methods; Trajectory analysis

2.1.3 High-Performance Computing Requirements

New programming models

Without question, computer architectures have undergone a noteworthy paradigm shift that now delivers multi- and many-core systems with tens to thousands of concurrent hardware processing elements per workstation or supercomputer node. Capitalizing on these new hardware architectures is a significant issue facing scientists in planning future research at EMSL's MSC capability.

The trend toward massive parallelism appears to be inescapable as current high-end commodity processors, the workhorse for most scientific computing applications, now support eight or more simultaneous threads of execution per CPU socket. Most computational nodes and scientific workstations contain several of these multi-core processors. Multithreaded applications using all of this hardware capability can deliver an order of magnitude increase in computational throughput and corresponding decrease in time-to-solution. Future systems will support even more processors and threads per processor.

A 10x performance increase is a significant advance and certainly will be an expected level of performance for high-end supercomputers procured in a few years, but it will not necessarily represent a fundamental change for computation-dependent science. Machines with this level of performance make the computational workflow more interactive because computational tasks that previously took hours now take minutes, and extended computational work that previously took days now can occur overnight.

Scientific and technical literature published during the previous two years demonstrates[1] a proliferation of General Purpose Graphics Processing Unit (GPGPU)-enabled applications and algorithms that deliver one to two orders of magnitude speedup (10x to 100x, respectively) in performance over conventional processors across a broad spectrum of algorithmic and scientific application areas. The applications that achieve high performance are massively-threaded, so they can fully use the many hundreds to thousands of simultaneous hardware threads of execution that become available when one or several GPGPU boards are plugged into a conventional processor motherboard. The very highest performing applications also exhibit relatively high data reuse within the graphics processors. In rare cases, even higher performance can be obtained as some researchers have reported three orders of magnitude, or 1000x, greater performance when performing calculations that heavily use transcendental functions on NVIDIA® GPGPU hardware.

Applications that deliver 100x or faster performance create paradigm shifts and have the potential to affect scientific research fundamentally by removing time-to-discovery barriers. Computational tasks that previously would have required a year to complete can be finished in days. Better scientific insight becomes possible because researchers can work with more data and have the ability to use more accurate, albeit computationally expensive, approximations and numerical methods. For the experimentalist in particular, the results of new high-throughput instruments (or collections of many instruments) can be used to create higher-resolution, more informative pictures of what is occurring in nature potentially in real time.

To use this new computational capability—and the ever more massively parallel systems that generally will be available in the next few years—requires a commitment to rewrite many of the computationally intensive portions of current applications to engage a large number of simultaneous threads effectively. As massively parallel hardware becomes more inexpensive, capable, and ubiquitous in the worldwide scientific community, rewriting certain computational applications appears necessary to keep computation-dependent research at the MSC capability competitive.

[1] See http://www.nvidia.com/object/tesla_computing_solutions.html. Some examples from the literature are: Stone JE, JC Phillips, PL Freddolino, DJ Hardy, LG Trabuco, and K Schulten. 2007. *Journal of Computational Chemistry* 28:2618; Ufimtsev IS and TJ Martinez. 2008. *Computing in Science and Engineering* 10:26; Anderson JA, CD Lorenz, and A Travesset. 2008. *Journal of Computational Physics* 227:5342; Manavski SA and G Valle. 2008. *BMC Bioinformatics* 9(2):S10; and Schatz MC, C Trapnell, AL Delcher, and A Varshney. 2007. *BMC Bioinformatics* 8:474.

Hardware and software requirements

Table 2.2 lists broad hardware and software requirements to permit state-of-the-art computations to meet the challenges in biological computations.

Table 2.2. Hardware and Software Requirements for State-of-the-Art Computations to meet Biological Computations Challenges

Hardware	Software
Fast CPU	Visualization (e.g., Ecce)
Shared versus distributed memory	New programming models (e.g., massive threading; massive scalability; fault resiliency application programming interface [API])
Fast interconnect	
Data storage	
Fast networks	
Visualization	
Fault resiliency	

The challenge in supporting biomolecular calculations is the diverse set of resource requirements for the various methodologies. Table 2.3 lists the various methodologies and highlights specific resource requirements for each of them.

Table 2.3. Various Methodologies and Resource Requirements that Support Biomolecular Calculations

Methodology	Resource Requirements							
	CPU	Massive Threading	Memory Bandwidth	Memory Latency	Interconnect bandwidth	Interconnect Latency	Storage Bandwidth	Storage Latency
	F/I/O						L/G/W	L/G/W
MC	F	+	X	X				
MD	F	+	X	X	X	X		
QM	F/O	+	X	X	X	X	(L)	(L)
Hybrid QM/MM	F/I/O	+	X	X	X	X	L/G	L/G
CPMD	F/O	+	X	X	X	X	L	L
Network Simulations	F/I	+		+				
Proteomics	I	+	X	X			L	L
Sequence Analysis	I	+	X	X				
Network Inference	I	(+)	X	X				
Structure Prediction (*ab initio*)	F/I	+	X	X				
Structure Prediction (Comparative)	F	+	X					
Image Analysis	F/I/O	+	X	X	X		G	
Statistical Methods	F/I/O	+	X	X				
Trajectory Analysis			X		X	X	G/W	G/W

F = floating point operations; I = integer operations; O = matrix transposes (orthogonalization, FFT); L = local; () = algorithm dependency; G = global; W = WAN; + = significant new science can be performed; X = required.

2.2 Chemical Sciences

Advanced computing resources will enable computational chemistry to provide critical insight into chemical processes and material properties necessary for breakthrough discoveries to address the challenges involved in developing new technologies for environmental security and energy production, storage, and use. In order to be successful in making scientific advances through computational chemistry modeling, there are distinct challenges in improving accuracy; using realistic models that include chemical complexity and heterogeneity of the system; generating predictive models of interfacial processes important to energy and the environment; using reliable structure-function relationships (structure, chemical composition, reactivity) in complex systems; and maintaining fundamental understanding of charge transfer, ionic transport, and dynamical processes in nanomaterials. In the following sections, these issues are discussed in terms of key science drivers, computational challenges, and required high-performance capabilities.

2.2.1 Science Drivers

Energy storage and conversion

It is critical to meet national and global energy needs with minimal environmental impact. A growing world population, growing per capita Gross National Product (for every country), and a consequent growing energy demand has led to dramatic growth in atmospheric CO_2 from anthropogenic sources. As of 1998, the world used 12.8 terawatts (TW) of power, and the United States (U.S.) consumed 3.3 TW. Current world energy consumption is 14.5 TW, and, by 2050, world power demand is projected to exceed 28 TW. On the basis of many different analyses, the long-term energy needs of the world could potentially be best met by direct solar energy capture if efficient solar cells can be developed and the captured energy transferred into a suitable storage medium. Solar cells must be able to directly convert sunlight to electricity or directly provide the energy for chemical transformations, e.g., the conversion of H_2O to H_2 (+ O_2) for use in fuel cells. The ability to store energy efficiently is critical to cover periods (nights, cloudy days, and winter) when there is insufficient sunlight. As the efficiency and costs of solar cells and other alternative energy sources improves, there is a definite, concurrent need to improve the ability to store that energy. Transformative technologies in the area of energy storage for solar energy capture systems and other alternative energy technologies, as well as the coupling between energy capture, storage, and production of electricity, must be developed. As stored energy is converted into electricity (via existing infrastructure), it needs to be delivered to the point of use, i.e., a home, business, or for fuel transportation.

Photovoltaics based on inorganic and organic materials hold great promise to provide attractive alternate energy sources. There are numerous fundamental material and applied issues that need to be addressed. Organic photovoltaic (OPV) systems offer the promise of low-cost, readily manufacturable alternatives to traditional inorganic systems for producing solar electricity, and impressive advances recently have been reported. If crucial scientific understanding challenges can be surmounted, power conversion efficiencies (PCEs) as high as 10 percent to 12 percent may be achievable. Computational simulations will provide significant insight into nanostructure and transport at both soft and hard matter interfaces. Photovoltaics based on nanowires offer unprecedented advantages arising from the 1) band-gap tunability with nanowire diameter; 2) increased effective path length; and, most importantly, 3) the ability to increase the optical dipole moment, which translates to an increase in absorption in indirect band-gap materials such as silicon. A theoretical understanding of nanowire performance as a function of materials, diameter, and orientation is required.

Advanced electrochemical energy storage systems are a critical cross-cutting technology for domestic applications that target enhanced environmental friendliness and a reduction in U.S. dependence on foreign oil. The commercialization of new energy storage technologies will accelerate the introduction of high-mileage electric vehicles, lightweight military

equipment, and distributed (though intermittent) wind and solar power generation. Forefront research to address challenges in developing advanced electrochemical energy storage systems is essential. In particular, advanced computational methods have the power to guide the design and development of new classes of materials that could revolutionize energy storage technology. Advances beyond today's technologies are needed to satisfy the energy storage requirements for transportation and use of renewable energy sources. In the area of lithium-ion batteries, there are opportunities for the development of advanced materials, such as new electrolytes, electrodes, and membranes, which will improve the performance, life, economics, and inherent safety of this technology, rendering it more broadly acceptable for use in transportation applications. New approaches based on computation methods with supercomputers, such as screening thousands of candidate materials through the use of descriptor-based models, hold the promise of reducing the cost and greatly accelerating the discovery process.

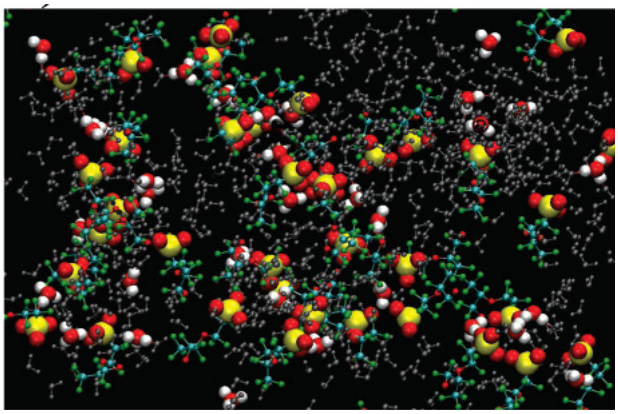


Figure 2.6. From a simulation in water of a sulfonated fluorocarbon polymer electrolyte including hydronium ions (*Image courtesy of A. Venkatnathan, R. Devanathan, and M. Dupuis, PNNL*).

Fuel cells have been attracting considerable attention as energy sources. Polymer electrolyte membrane fuel cells (PEMFCs) convert the chemical energy of fuel into electrical energy with high efficiency and minimal pollution. They have the potential to revolutionize transportation, distributed power systems, and portable power systems. Fundamental scientific understanding from multiscale computer simulations, in conjunction with experiment, is needed to overcome barriers to commercialization of fuel cells, such as high production cost, performance degradation, poor durability, and low service life under challenging operating conditions. Similar challenges also have been encountered in the development of solid oxide fuel cells with a ceramic membrane. A key component of the PEMFC is a polymer membrane (Figure 2.6) designed to separate the reactant gases and selectively conduct protons. Unfortunately, PEM development has mostly followed an Edisonian approach. To enable widespread commercialization of PEMFC technology, there is a pressing need to develop novel membranes based on a fundamental understanding of membrane chemistry, morphology, charge transfer, proton transport, percolation of water molecules, and the nature of the pore network. Modeling of fuel cell membranes that can account for the multiple length and time scales will provide the insight needed to solve these problems.

Nuclear materials

Advanced nuclear energy systems involve the exposure of a broad range of actinide-bearing fuel and structural materials to extreme environments that include ionizing radiation, elevated temperature, stress, and corrosive chemicals. The behavior of materials is governed by the formation, migration, and long time scale evolution of atomic-level defects, defect clusters, and gas bubbles and their interactions with impurities, dislocations, and interfaces. These phenomena in multi-component actinide fuels, structural materials, and nuclear waste forms often are not easily accessed by experiment. Computational studies have to span length scales from nanometer to meter and time scales from picoseconds to years.

A critical need in nuclear materials science is the ability to predict the electronic structure of materials to obtain reliable information about the thermodynamics of the systems and the kinetics of critical reactions and processes. Compounds containing heavy elements require a proper treatment of relativity, which includes both scalar relativistic and spin-orbit components. Materials composed of atoms and molecules with open 4*f* and 5*f* shells exhibit strongly correlated electron behavior, which, thus far, has prevented reliable predictions of how the physical properties of a material system changes

in response to external conditions such as temperature, pressure, and impurities. Furthermore, the development of a predictive understanding of the formation, stability, reactivity, and structures of aggregated actinide complexes under conditions relevant to the reprocessing of spent fuel is needed to design separations processes that target these species. Small aggregates in nuclear waste streams can cause a low-level waste stream to require treatment as a high-level waste thereby increasing treatment costs. This calls for a fundamental understanding of the chemistry underlying nanophase formation, structure, stability, and reactivity.

The presence of the actinides makes the chemistry of nuclear reactor fuel initially complex, and the continuous loss of uranium and plutonium and formation of a broad range of new species due to fission introduces a challenging time-dependence to this chemistry. The fuel ultimately contains multiple *f*-electron elements: uranium, plutonium, americium, neptunium, and curium, as well as many lighter elements. This situation leads to the potential formation of many phases that can influence critical physical properties, such as thermal conductivity. The integration of new *ab initio* electronic structure results with available thermodynamic databases is necessary to enable the prediction of phase equilibria and oxidation states in fuel that contains fission products, which may have been generated *in situ* or mixed into fresh fuel.

The separation of elements either for further use as a fuel or waste is a critical part of the overall nuclear power system, but it is not independent of the system. Separation processes are intermediate in the overall nuclear power generation system and are affected by other steps (e.g., the fuel that is input for the separation), and the extent of separations affect other steps (e.g., the type of waste form or reprocessed fuel). Many of the processes for generating nuclear fuels and separating spent fuel for recycling or as waste take place in solution, nominally aqueous solution, or at interfaces. The dynamics of the process involve collective effects and more complicated reaction coordinates than simple bond breaking or formation. Host-guest 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. Critical issues for the reliable prediction of solution phase processes include being able to predict entropy in complex systems involving the coupling of strong and weak interactions. The large number of potential structures demonstrates the need to develop new phase space sampling methods to deal with this currently computationally intractable problem for realistic systems.

Interfaces and surfaces

Catalysis science is an essential discipline for addressing many of the DOE's missions, including developing new technologies for a secure energy future. Catalysts are molecules or materials that facilitate chemical reactions (e.g., the breaking and making of chemical bonds) to occur at increased rates (i.e., increased activity) or produce desired products (i.e., increased selectivity) while leaving the catalyst unaltered over long times (i.e., stability). Catalysis plays a critical role in the conversion of feedstocks, such as fossil energy, into fuels, such as gasoline and diesel. Catalysis also will be essential for energy conversions of electrical energy generated from nuclear energy or renewable energy sources (e.g., solar, wind, and hydro) to other forms of energy that are more easily stored and transported. The optimum means to store energy is in chemical bonds because chemical energy sources have the highest energy density (Figure 2.7).

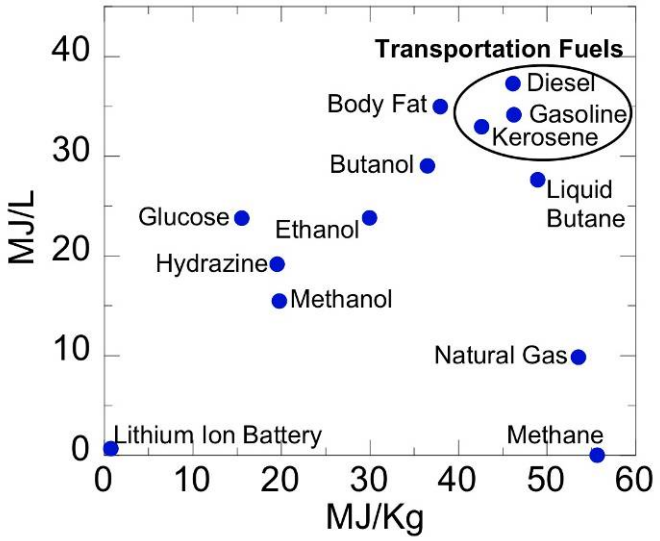


Figure 2.7. Volumetric and gravimetric energy densities of energy sources.

As the availability of traditional fossil energy feedstocks decreases, future energy sources will require treating transformations with increasing complexity. These conversions include complex feedstocks such as biomass and coal, which are large, multifunctional molecular systems. Catalytic conversion of CO_2 in fuels and water to form hydrogen using solar or electric energy will require much more complex processes for driving thermodynamically stable species uphill in energy to more energy-rich fuels. These types of catalytic conversions require the means to control chemical transformations far beyond what is possible today. The drive to exert more control over catalytic processes has led to the development of new catalysts. Nanostructured catalysts are an active field because of their increased activity (Figure 2.8). These catalysts present challenges for computational studies because of the variety of possible "active" sites and the fact that they are large, but finite, so periodicity cannot be exploited. Catalysts with multiple functionalities are required to carry out sequences of reaction steps necessary to transform feedstocks into useful fuels. These systems introduce the complexity of the interaction of reactants and products with multiple active sites in close proximity and the coupling of the dynamics and kinetics of the reaction steps. Catalysts are being developed that incorporate structural motifs and functionality exhibited in biological systems, such as enzymes. These systems require understanding the role of weak interactions on controlling structures and how flexibility influences reaction dynamics.

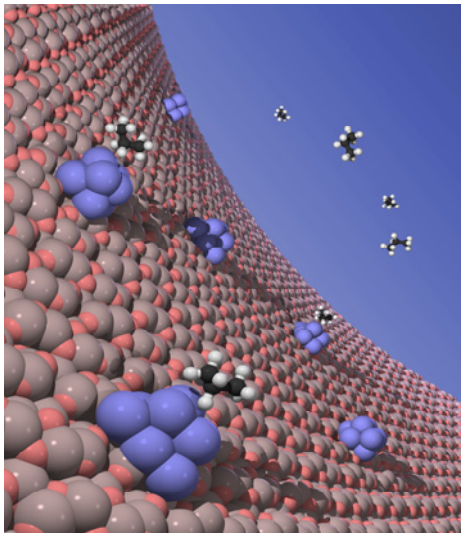


Figure 2.8. Pt_{8-10} clusters stabilized on nanoporous anodized aluminum oxide membranes (*Image courtesy of Curtiss Group, ANL*).

For many catalytic processes, it still is unclear how the catalyst works and which steps control the activity and selectivity. A workshop sponsored by the Basic Energy Sciences Advisory Committee (BESAC) reached the conclusion that "the Grand Challenge for Catalysis Science in the 21st Century is to understand and thereby control the relationship between catalyst structure and catalytic chemistry (both activity and selectivity)."[2] A more recent workshop report emphasized the important role theory and computation will play in advancing catalysis science: "Advances in theory and computation are also required to significantly advance catalysis for energy applications. A major challenge is to understand the mechanisms and dynamics of catalyzed transformations, enabling rational design of catalysts. Molecular-level understanding is essential to 'tune' a catalyst to produce the right products with minimal energy consumption and environmental impact."[3]

Developing a fundamental understanding of reaction mechanisms underlying catalytic conversions requires close coupling of theory and computation with experiment. Advances in experimentation such as scanning probe microscopies to study interfaces at the atomic scale, high-resolution electron microscopies that can study molecular processes at interfaces under realistic catalytic operating conditions, and X-ray spectroscopies that probe the local environment around molecules and ions require computations to interpret experimental observations and guide the design of experiments. Experiments play an essential, complementary role in providing tests of the computational predictions and new data for interpretation. The interaction of experiment with theory and computational modeling is always at its best when they are working together to answer common scientific questions.

[2] White JM and J Bercaw. 2002. "Opportunities for Catalysis in the 21st Century." Report from the U.S. Department of Energy Basic Energy Sciences Workshop, May 12-14, 2002. See: http://www.er.doe.gov/bes/reports/files/OC_rpt.pdf.
[3] Bell AT, BC Gates, and DR Ray. 2007. "Basic Research Needs: Catalysis for Energy." U.S. Department of Energy Basic Energy Sciences Advisory Committee Subpanel Workshop Report, August 6-8, 2007. See: http://www.er.doe.gov/bes/reports/files/CAT_rpt.pdf.

Sensors

The computational study of charge transport in nanostructures is a pathway for both discovery of underlying operation principles and design of nanoscale electronic sensors. Such sensors can detect toxins, chemicals, nuclear materials, and biological molecules. The realization of nanoscale sensors is challenging from an experimental viewpoint because of the need for a high degree of specificity for a large number of target molecules concomitantly present at low concentrations in the sample. Currently, the building of sensor prototypes is driven by empirical observations resulting from experiments coupled with methods of statistical inference. Chemistry- and physics-based modeling to explain the data and design of devices is in the early development stages. The primary needs are: 1) accurate computational modeling based on *ab initio* and semi-empirical methods, 2) coupling of accurate computational modeling with approaches to calculate the electrical conductance of the sensors, and 3) development of computational algorithms to enable fast solvers to calculate the charge and current densities under a diverse set of experimental conditions. The successful development of a framework is challenging because of the length scales involved. One has to deal with an accurate description of the chemistry at the atomic scale using quantum chemistry methods while being able to predict the sensor performance over a length scale of tens to hundreds of nanometers.

2.2.2 Computational Challenges

Multiscale modeling

The major computational challenge for addressing the previously described science drivers is achieving realistic modeling of the multiple length scales (nanometers to meters) and time scales (picoseconds to years) involved. Multiscale modeling is a developing research area aimed at bridging the gaps between the diverse computing communities in the different scales. This general approach is being applied to a wide variety of areas, such as nanoscience, atmospheric sciences, combustion, materials, and biology. Currently, there are numerous computational methods, such as DFT, finite element methods, MD, etc., that already are well developed, applicable, and suitable for a comparatively narrow range of scales and capable of predicting selected aspects of a material with a certain accuracy. Advances in multiscale modeling will enhance the capability to more reliably compute coarse scales while maintaining the required accuracy of smaller scales, opening up exciting opportunities for design and discovery in chemistry.

The development of efficient and accurate methods for information passing is a key challenge for multiscale modeling, which can be done concurrently or hierarchically. In concurrent methods, the simulated system contains multiple models, for instance, a region of interest treated with a high level theory embedded in a matrix that is coarse grained. Embedding schemes and QM/MM methods fall into this category. Different regions of the problem are treated with different resolutions. The drawbacks of this method include difficulties in the correct treatment of the interfaces between different regions and the fact that the time scale is dictated by the slowest model. In order to overcome the time constraint, it is important to consider hierarchical methods where information is passed from lower to higher scales, usually in the form of parameters. The ultimate and elusive goal is to develop a simple constitutive equation that incorporates the essential physics from lower scales. It is important to select the essential information that needs to be passed to the higher level from the large volume of data generated at each level. As an example of this challenge, the development of truly transferable force fields for all atoms in the Periodic Table for use in classical MD based on *ab initio* data is a daunting task. Due to these difficulties, current modeling approaches focus on single scale techniques, which can only poorly approximate the complexity of real materials.

The development of novel catalytic materials is an area where multiscale modeling can play a key role. As described earlier, catalysis is a major component of the DOE's energy mission, including for processes that create fuel stocks from

renewable energy sources. The physical processes that determine catalyst performance span a wide range of length and time scales (Figure 2.9), from chemical reactions at the atomic scale up to the design and performance of macroscopic chemical reactors. As such, for these applications, it is critical that the theoretical models of the catalytic materials capture the chemical complexity of the actual materials being studied experimentally. This rather "simple" statement has two corollaries for theoretical modeling that make this area of research extremely challenging.

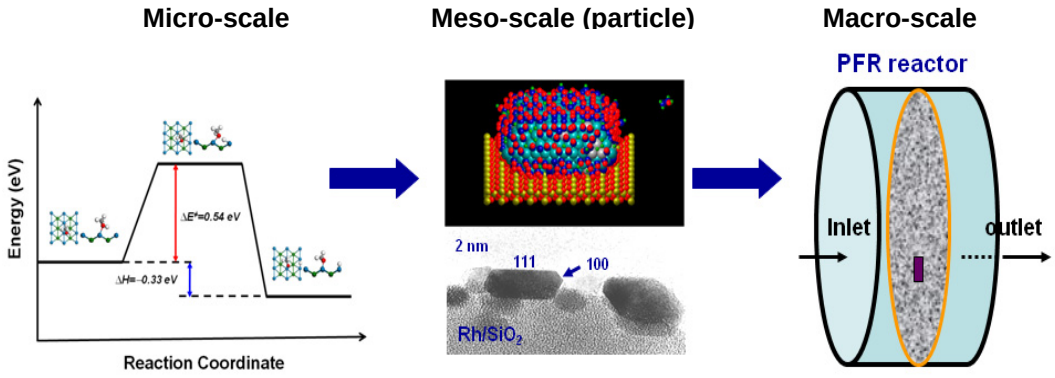


Figure 2.9. Multiscale hierarchy for catalytic reactor model informed from kinetic model and molecular reactivity obtained from experimental observations. (*Image courtesy of D. Mei, PNNL*).

Initially, it requires that atomistic models are on the order of 10^3~10^4 atoms, especially for heterogeneous processes that are modeled on surface slabs, often including defect sites in low concentrations. Secondly, for a given catalytic material, consideration must incorporate the potential of a large number of possible chemical reactions and their (free-) energy barriers in order to predict both the catalytic pathways and the side reactions, which can lead to unwanted byproducts or degrade the catalyst. Quantum chemical methodologies are necessary to explore the chemical space of the catalytic system because of the changes in electronic structure, which occur during the reactions. Currently, *ab initio* electronic structure approaches at the density functional level of theory are routinely used on small models of these catalytic systems for detailed investigations of the thermodynamics and kinetics of catalytic reaction mechanisms, which is accessible on today's large-scale computational platforms. However, the atomic length scale/time scale does not cover the entirety of the relevant physical processes necessary to understand catalysis. Currently, intermediate time and length scales (hundreds nanometer [nm] and up-to-second scales) are handled by kinetic modeling, such as kinetic Monte Carlo (KMC). In the KMC approach, the catalyst system (typically a small surface area or a nanoparticle) is modeled based on reaction energy barriers and kinetic parameters obtained from *ab initio* calculations carried out on smaller representations of the system. This approach is typically employed for heterogeneous phase catalysts. Extensions to include gas phase reactions (common in chemical reactors at elevated temperature and pressure conditions) or to encompass homogeneous phase catalysts are areas of current topical interest. A fundamental shortcoming of these models is the necessity to obtain kinetic parameters for all of the relevant reactions, including those which may not have been predicted *a priori*. Although some work has been conducted to combine KMC directly with electronic structure methods, the computational demand of the stochastic sampling of reactions for an unbiased determination of the relevant reactions and their rates at the DFT level is, at the moment, only possible for very simple systems.

Finally, large length and time scales (centimeters to meters, seconds to hours) are handled by reactor models involving the numerical solution of partial differential equations (PDEs). Coupling this approach with kinetic models and *ab initio* electronic structure methods offers the possibility to investigate how inhomogeneities in the heat and mass transport encountered in a chemical reactor can influence catalyst performance and operation in real-world applications. This multiscale approach, which offers the possibility to address the relevant physics occurring at all time and length scales on

real catalysts, is truly uncharted territory and will involve both efficient fluid dynamical methods (fast and parallel PDE solvers) and all the necessary advances in the KMC and atomistic models already outlined. It is envisioned that this approach, which will require multilevel parallelization hierarchies ranging from tightly coupled processor subgroups to loosely coupled Monte Carlo walkers or MD replicas, will require platforms of the exascale and beyond.

Multiscale modeling also is crucial in problems involving charge transport in nanoscale devices, such as sensors, which require careful treatment across multiple length scales (atomic to the nano length scales). At the atomic scale, the details of bonding between atoms of the device and an interaction with external stimuli (light, molecules, and mechanical deformation) and substrate atoms should be accurately modeled by electronic structure methods. The experimentally measurable current flow or charge transport, however, occurs over a length scale that is one to two orders of magnitude larger (typically tens to hundreds of nm). Accurate electronic structure methods at these length scales are impossible at this time as 1 to 2 nm^3 is about the size that can be considered today for semi-quantitative accuracy. In this case, the main computational methodology that requires development is embedded schemes, where information obtained at the atomic length scale using *ab initio* methods forms the basis for semi-empirical methods that are used to model system sizes with many hundreds to thousands of atoms. The embedding schemes should be robust to model realistic experimental conditions in typical applications, such as finite temperature and electric biases, and should include interactions between electrons and vibrational modes. The typical material systems of interest are broad and include nanotubes, graphene, nanowires, and floppy biological molecules such as protein and DNA.

Realistic, large-scale materials simulations

Realistic, large-scale simulations of material behavior are needed to understand the fundamental mechanisms associated with materials phenomena, such as fracture, sliding friction, and high strain-rate deformation in nuclear materials. Micron-sized systems need to be simulated by MD to represent grain structure, dislocations, voids, bubbles, and other microstructural features faithfully. At the same time, simulations need to be extended well beyond the picoseconds scale of current large-scale simulations. Unique data management and analysis challenges can arise out of such peta- and exa-scale simulations.

Exascale computations will challenge the data storage and bandwidth projections of the current technology paths. The simulations will consume and generate large amounts of data. From the combination of complexity and sheer quantities, there will be bandwidth bottlenecks in the knowledge management aspects of data-intensive computing. Certainly, this will require new solvers with more favorable scaling for analyses of these knowledge sets. The generated data sets are likely to become so large that transferring and storing them offsite may not be practical. Data has to be analyzed and visualized quickly and effectively on-the-fly using massively parallel algorithms. Quad precision floating point may be required for stochastic dynamics simulations. The data sets generated by massively parallel simulations using new architectures (e.g., trillion atom simulations) will tax data transfer and storage capabilities. Mining such large data sets for scientific insights also poses significant challenges. To accelerate knowledge extraction and categorization, steered simulations based upon knowledge consideration are needed. In designing the computing platform, a balance has to exist between speed and memory with the optimal balance varying with the problem of interest.

Rational materials design and screening

In the near future, theory-based design of materials with desired functional properties will revolutionize the materials discovery process. Indeed, there already have been some success stories of computational materials design, for example, ones based on surface reactivity calculations from first principles electronic structure theory and the replacements of the chlorofluorocarbons (CFCs). This is especially important since new materials underpin fundamental studies of current

science drivers, such as catalysis, high-temperature superconductivity, biomaterials, and photovoltaics. A portfolio of such "materials by design" might include materials that efficiently catalyze conversion of a hydrogen-rich molecule, such as methanol to hydrogen for fuel cells; new molecular or compound semiconductors that efficiently absorb sunlight across the full solar spectrum; or hierarchical organic-inorganic assemblies that provide the spin-coherent host as a basis for quantum computing. Given the high cost of experiments, computational materials design can improve the efficiency of the search process for new materials and have a major impact on science and engineering, although the inverse materials design problem, i.e., given a required set of properties, then design the material, is quite difficult and has not yet happened. Thus, HPC will play an essential role that will enable generation of physical and chemical property data on realistic models of the candidate materials needed for design from first principles and, just as importantly, be able to achieve the fast throughput required for the vast search space of potential new materials.

Rational design of materials poses several computational challenges. A typical example is the computational design of catalysts. A number of different spatial and temporal scales need be considered; heterogeneous structures have to be treated realistically; environment effects have to be considered; and complex metal-metal interactions need to be treated accurately as many catalysts involve *s*-, *d*-, or *f*-block metals. Additionally, catalysts need to be designed for use in real environments, such as chemical reactors, which means that effects, including diffusion, sintering, mobility, and catalyst protection, need to be considered. Finally, there are important issues in computational catalysis related to coupling multiple spatial and temporal scales.

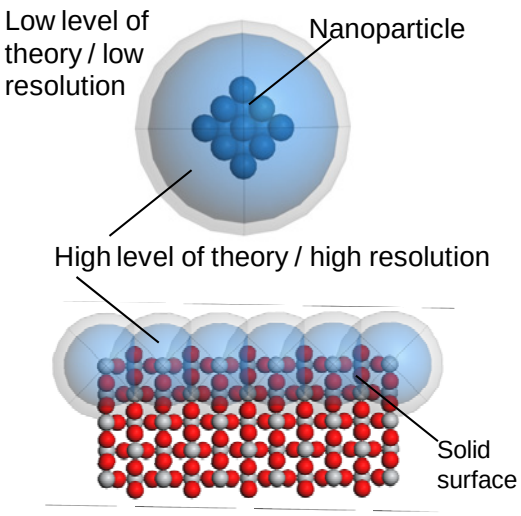


Figure 2.10. Multiscale chemical dynamics of complex systems can combine CCSD(T) (blue) with classical molecular dynamics in fluids (*Image courtesy of A. Heyden, University of South Carolina, and D. Truhlar, University of Minnesota*).

Addressing these issues will require access to the next and future generations of hardware, as well as software that readily runs on the hardware. Solving model problems qualitatively must shift to solving the correct real problem quantitatively (for example, Figure 2.10). Continued development of electronic structure methods must be implemented in usable, general purpose codes on advanced computer architectures. There also is a need to go beyond the atomistic scale to cover the entirety of physical processes required to understand the problem at hand. Truly multiscale approaches in length and time are uncharted territory because they require smooth coupling of different methods, starting from Coupled-Cluster with Single and Double and Perturbative Triple excitations [CCSD(T)] with large basis sets and proceeding up the scale to KMC methods and fluid dynamics or finite element methods. Multilevel parallelization hierarchies, ranging from tightly coupled processor subgroups to loosely coupled replicas, also will require exascale computing resources.

Electronically excited states in solids and solid surfaces

Many phenomena related to energy harvesting and utilization and radiation chemistry involve the creation of electronically excited states. For instance, the photovoltaic harvesting of energy can involve either the e-h pair excitation directly in semiconductors or the ultrafast interfacial charge transfer from photoexcited chemisorbed dyes or nanoparticles into the semiconductor substrate. The goal of photocatalysis is to convert photon energy into energy-rich fuels through photo-induced, e-h pair-driven redox processes on semiconductor surfaces.

Further progress in efficient harvesting of solar energy or modeling radiochemical processes requires the appropriate theoretical tools to describe electronically excited states in solids and solid surfaces. Highly successful methods for

describing the ground state properties of materials, such as DFT, are known to substantially underestimate the quasiparticle gap in semiconductors and electronically excited adsorbates on solid surfaces. Hybrid functional methods increase the band gap energy according to the extent of admixture of Hartree-Fock exchange, but the accuracy of the quasiparticle wave functions is unknown. Quantum chemical methods can be highly accurate for small systems, but when applied to condensed matter, the prediction of excitations requires embedding schemes, which are material-specific and difficult to implement. Many-body perturbation theory (MBPT) methods, such as the Green's function-based GW methodology together with the Bethe-Salpeter equation (BSE), have had considerable success in describing the band gaps of semiconductors such as silicon, a variety of metal oxides, and highest occupied molecular orbital-lowest unoccupied molecular orbital (HOMO-LUMO) gaps of chemisorbed molecules. The imaginary part of the quasiparticle self-energy obtained from such calculations also provides information on the electronic decay time scales through the screened Coulomb interaction. Such MBPT methods are costly and, so far, have only been applied to few systems.

High-level coupled-cluster methodologies are needed to account for electron correlations in combination with embedded cluster techniques for describing the solid. Achieving this level of sophistication in modeling electronically excited states in solids has only recently become possible with development of petascale computers and new, powerful, and scalable methods for treating electron correlation. For instance, the equation of motion coupled-cluster (EOMCC) formalism has evolved into a widely used and accurate method. This formalism provides a well-defined way of categorizing the basic correlation effects and possibility of constructing a hierarchy of increasingly more accurate approximations from the rudimentary EOMCC method with singles and doubles (EOMCCSD) to more sophisticated and accurate models, including triply (EOMCCSDT) or triply and quadruply excited configurations (EOMCCSDTQ). Efficient approaches are needed to account for the effect of triply (or higher) excited configurations and mediate cost-wise between iterative EOMCCSD and EOMCCSDT methods, at the same time providing a substantial improvement of the EOMCCSD results in calculating vertical excitation energies and excited-state PES and dynamics.

The additional computational resources available from the proposed system will make MBPT methods increasingly practical. Moreover, extensions to the higher level of perturbation theory, e.g., by describing the short-range interactions in highly correlated materials through the T-matrix approach, will provide increasingly more accurate descriptions of excited states. The accurate description of quasiparticle wave functions will lead to a better understanding of excited carrier-driven chemistry at solid surfaces.

Spectroscopic properties

Optical spectroscopy is a powerful characterization tool, and much knowledge of the electronic properties of chemicals and materials is obtained from it. These techniques, including electron spectroscopy, X-ray photoelectron spectroscopy, fluorimetry, NMR, EPR, vibrational, etc., are applied to all Science Themes across EMSL. Theory is critical to assist in correctly interpreting optical spectra by providing insights to understand the excitation transitions. Time-dependent DFT or wave functional theory for calculating excited states properties are computationally challenging. The demand for computing time-dependent properties of larger systems, such as nanomaterials or biological molecules, is pressing. Currently, these types of simulations are limited by the computing resources to small sizes or more approximate theory. In order to address size-dependent optical properties, it is highly desirable to be able to calculate real systems with more accurate methods, such as nanoparticles. Furthermore, including the surrounding environment provides added complexity.

Electron transfer

Charge transfer across semiconductor-molecule interface is fundamentally important across several fields, including photo-, electro-, and analytical chemistry; molecular electronics; and photography. Charge transfer processes are fundamentally important for the conversion of solar to electrical and chemical energy. Issues affecting interfacial charge transfer include the electronic coupling matrix elements, the role of adiabatic versus nonadiabatic processes, the correlation of proton and electron motion, electron relaxation and delocalization inside the semiconductor or localization at the interface, back-transfer processes, and e-h recombination that reduce the quantum yield of energy conversion. Describing semiconductor molecule interfaces presents difficulties related to the appropriate description of molecules with discrete, localized electronic states; unique vibrational spectra; and well-defined directional bonds, as well as solid-state materials with continuous bands of delocalized electronic and vibrational states. In the strong-coupling regime that is not captured by Marcus Theory, the charge transfer processes involve interfacial states created by, for example, the strong H_2O-TiO_2 interaction for photochemical splitting of water on titanium dioxide, and that do not exist separately in either material. Addressing these challenges will require advances in nonadiabatic MD implemented within time-dependent DFT and the next generation of computing resources.

Rare-event discovery and complex dynamical processes

Calculations of rate constants require identifying reaction pathways and saddle points. For activated reactions, which have potential energy barriers for transition from one region of a PES around a local minimum (a potential basin) to another potential basin, the transition rate is much longer than the dynamics of molecular motion in the potential basins. These events are characterized as infrequent or rare events and can require extremely long dynamical simulations for their discovery. When both the initial and final states of a reaction are known, the reaction pathway and saddle point can be found readily using existing methods. However, when only the initial state of the transition is known, finding the relevant saddle point(s) becomes a challenging problem. The time scale of the simulation is limited by the computational cost of evaluating the PES so that use of electronic structure methods directly in the simulations of rare events is prohibited. Rare-event discovery tools, such as accelerated dynamics methods, provide approaches to improving the search reaction pathways, although they have been applied only to relatively simple problems. Even with the advanced methods emerging for treating these systems, complex reactions in condensed phases and interfacial systems will require considerable advances in computational capabilities.

Condensed phase and interfacial molecular systems can exhibit complex behavior in which processes of interest are controlled by collective motions of the molecules in the system. Examples of these types of processes are phase transitions (e.g., freezing of liquids and crystallization), which are rare events where the reaction pathway is described by a complex order parameter involving collective MD. These types of processes provide extreme challenges for rare-event discovery tools, and, often, the only way to unveil the reaction pathways is to perform extremely long simulations. Increases in computational resources are essential for these types of studies.

Chemical accuracy for thermodynamics

The computational design of new materials and molecules for real applications requires the ability to predict, at the molecular level, the detailed behavior of large complex molecules, as well as solid-state materials, together with their reaction environments. Although intermediate-level computations, such as DFT, often can provide insight into chemical properties of molecules and materials, true computational design of practical relevance will require the ability to predict accurate thermodynamic (±1 kcal/mol or less) and kinetic (rate constants to at least an order of magnitude, initially)

results. The requirement for such accuracy means that we must be able to predict thermodynamic and kinetic quantities to high accuracy—currently a daunting computational task particularly for systems that involve metal atoms.

The core methods needed to attain the goal of designing catalysts for complex chemical process systems are electronic structure theory, including high-level *ab initio* molecular orbital theory, such as coupled-cluster methods with large correlation-consistent basis sets that can be extrapolated to the complete basis set limit; DFT with large basis sets; or complete numerical solutions for both static and dynamic (e.g., Car-Parrinello) approaches. In addition, for heavy transition metal and main group atoms, relativistic effects must be included, often by the use of effective core potentials. Furthermore, reliable techniques are necessary for the prediction of spin orbit effects in ground and excited states, and these are not readily available. Advanced kinetic modeling methods, e.g., variational transition state theory (VTST), reaction path sampling, and direct dynamics simulations, need to be employed. By using path integral methods, quantum effects can be included in the dynamics simulations, which will lead to improved predictions of reaction rates, especially when light atoms such as hydrogen are involved. At this time, CCSD(T) or multi-reference configuration interaction (MRCI) methods with basis sets that allow extrapolation to the complete basis set limit provide the highest accuracy. Such calculations are extremely difficult as they scale as N^7 or higher for N basis functions. To obtain chemical accuracy for chemical systems containing heavier elements, additional effects such as core-valence interactions, spin-orbit coupling, and reliable predictions of vibrational frequencies need to be incorporated into the computational models. In addition, the availability of reliable computational and experimental benchmarks are essential to provide insight into the quality of the more approximate DFT methods that have substantially improved computational performance scaling (often on the order of N^2 to N^3). In addition to the energies at optimized structures (gradients) and frequencies (second derivatives), calculations of rate constants beyond using simple transition state theory (TST) require more points on the PES than just a transition state and minima for reactants and products, raising the computational cost.

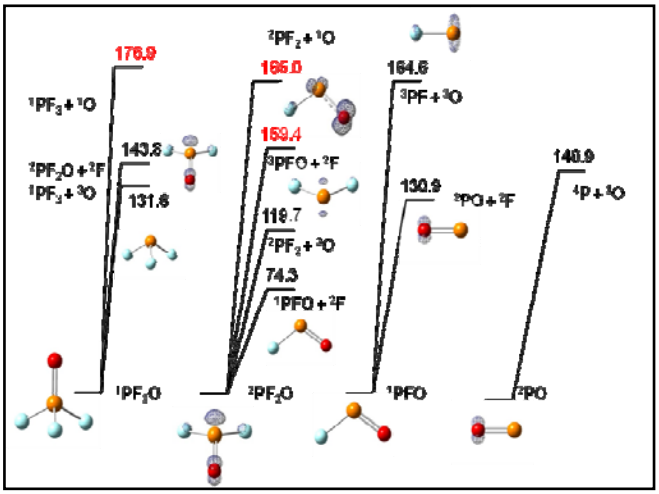


Figure 2.11. Bond dissociation energies of second row compounds at chemical accuracy using CCSD(T) and complete basis set calculations (*Image courtesy of Dixon Group, The University of Alabama*).

Another critical issue that must be addressed is the ability to predict the effects of entropy in complex systems in order to predict rate and equilibrium constants that require free energies. Today, microscopic solvation environments for single ions are predictable, and reasonable structural, thermodynamic (±2 kcal/mol), and ground state spectroscopic properties are obtainable. The issue is how to predict solvent effects reliably at different concentrations, temperatures, pressures, pH, ionic strengths, and media. Today, the rates of simple reactions in the gas phase are predictable using electronic structure theory to evaluate the PES and a kinetic theory, such as TST.

Currently, a number of issues affect the prediction of chemically accurate results for reactions in solution (see example Figure 2.11), including:

- how to predict the rate of the reaction in solution to the same accuracy
- how to predict absolute free energies of binding in solution
- how to predict kinetics of ion binding
- how to reliably predict excited electronic states and properties in solution.

Today, the energetics of excited electronic states for reasonable-size molecules can be predicted with different electronic structure methods, but it is difficult to obtain 0.1 eV accuracy. And, the common time-dependent DFT methods cannot treat two-electron excitations or long-range charge-transfer.

In order to address the preceding issues, a number of basic scientific advances must be made, including:

- new DFT functionals that can be used for the reliable prediction of weak interactions and relativistic effects, including multiplet splittings and excited states
- improved spin orbit treatments in electronic structure methods
- the ability to readily go beyond CCSD(T) for the "valence" electrons with large basis sets
- improved solvation models for thermodynamics beyond parameterized self-consistent reaction field approaches, e.g., to treat different temperatures, pressures, pH, and ionic strengths.

Addressing these issues will require access to the next and future generations of hardware, as well as software that readily operates on the hardware. Again, the paradigm must shift from solving model problems qualitatively to solving the correct real problem quantitatively.

Accurate molecular dynamics simulations

Most dynamics simulations are completed using classical mechanics not because it is accurate, but because it is the only practical method for the problem of interest. Classical MD simulations are acceptably accurate for a wide range of problems. However, there are situations and conditions where quantum effects are important, and the classical approximation is not valid. In fact, even in the modeling of liquid water, quantum effects can be important and can affect the nature of the hydrogen bonding, diffusion, orientational relaxation, and other properties. Fundamentally, classical mechanics can lead to the wrong heat capacity. In chemical reactions, classical mechanics can result in serious overestimations of rates. Perhaps, the most serious problem is neglecting proper behavior of intramolecular zero-point energy (or, more generally, vibrational energy). Quantum mechanically, each vibrational mode is expected to contain an amount of energy at least equal to the zero-point energy of that mode. However, in a classical mechanical simulation, the energy can flow among the modes without this restriction, yielding behavior that does not correspond to that of the real system. Clearly, this unphysical behavior is more important in polyatomics molecules, but even the quantum effects in phonon modes may be greater than appreciated. These effects may be partially minimized by the use of force fields that have been parameterized using classical simulations. The trend is toward greater usage of *ab initio* forces computed directly with quantum chemistry theory, where quantum effects may be more significant.

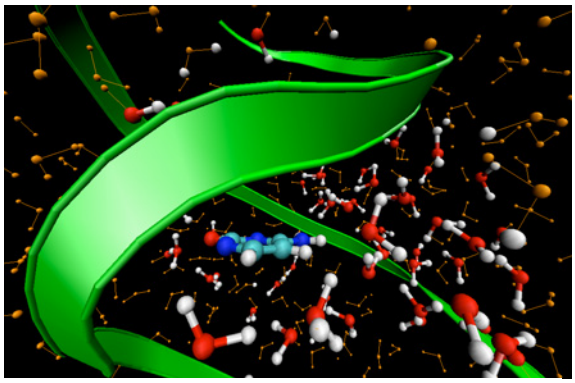


Figure 2.12. Cytosine base (blue) embedded in the relatively anhydrous DNA environment in water (*Image courtesy of E. Cauet and J. Weare, University of California, San Diego, and M. Valiev, PNNL*).

There are various approaches that provide some remedy, yet none is widely used. As demands for more and more accurate predictions of the full range of processes being simulated increase, the need to correct for the neglect for quantum effects becomes more pressing. Direct solution of the Schrödinger equation is only an option for small model systems. Any approach must be an approximate one. An obvious way to deal with quantum effects is to use a mixed quantum-classical approach (Figure 2.12). Unfortunately, this not suitable for many problems and has the inherent problem of how to properly couple the classical and QM modes. More general

approaches are semiclassical methods that treat all of the modes on the same basis. The semiclassical initial value representation has been incorporated in at least one widely used standard MD code. Soon, path integral-based approaches are likely to become widely applicable. Whatever the final practical solution(s) for dealing with quantum dynamics effects, it is critical in preparing to meet the demands of accurate predictions for almost all practically important processes that consideration be given to the need for bearing the costs of more expensive CPU simulations that account for interference effects.

Software considerations

The next-generation computer must be able to run computational chemistry software codes efficiently and in a scalable fashion. Typical codes include NWChem; plane-wave/periodic codes, such as VASP, DACAPCO, and the NEGF simulator; codes based on GW methods; BSE; CPMD; CP2K; MOLPRO; ADF; QMCMD; MC; KMC; and AMBER.

2.2.3 High-Performance Computing Requirements

Parallel computing platforms, fast data transfer and storage, scalable codes, algorithm development, and the availability of technical scientific consulting all are crucial components of using HPC for the scientific discovery process. Domain scientists are interested almost exclusively in using computational tools to solve a scientific problem without getting bogged down in hardware customization, software development, and code parallelization—areas in which they are not experts. New capabilities need to be developed to manage the possibilities and limitations of massively parallel computing resources. For example, fault-tolerance must be built into the code. Since large amounts (exabytes) of data will be generated by HPC, the traditional model of archiving data and analyzing it later may not be feasible. Parallel data analysis needs to be built into the code, which means the type of information being sought must be known *a priori*. Knowledge management is important when information is passed between different scales of a hierarchical problem. The following are important aspects of the new modeling paradigm:

- development of environment-dependent force fields on-the-fly
- information passing between different levels of theory embedded in a simulation
- algorithms that can steer massively parallel simulations based on the results
- algorithms that scale to more than 100,000 processors
- data localization and efficient local communication
- parallel data analysis, management, and visualization during the simulation.

Hardware requirements

For electronic structure codes, most of the current algorithms are built on a cache-blocked architecture. Therefore, cache latency, bandwidth, and size are important to performance and scalability. Also, the ability to interleave computation, communication, and input/output (I/O) operations (e.g., use of asynchronous I/O; use of non-blocking communication operations) is key to performance. Because many of the calculations involve large matrix operations, communication bandwidth and latency are especially important. In addition, both large local memories and access to copious amounts of fast storage to store intermediate results temporarily benefit the calculations. A key issue is the ability to deal with distributed data structures on a fully distributed memory system.

New approaches to electron and nuclear dynamics, especially ones that will scale to large processor counts, are also necessary. These include quantum Monte Carlo and path integral Monte Carlo for the simulation of multi-electron systems at finite temperature and quantum effects in nuclear motion. However, quantum Monte Carlo approaches currently are not readily available for general purpose electronic structure simulations as previously described, especially for extending beyond single-point calculations. The preceding description of the algorithms clearly shows that a balanced architecture is needed—balanced in terms of fast single-processor performance, low latency switches, fast I/O, and substantial amounts of memory and I/O capacity. The research also requires these resources be easy to use and computationally efficient.

Other computing requirements and specifics

Computing architecture also should be tailored to the problem at hand. For example, MDGRAPE-3 is a special-purpose computer system for MD simulations developed by the RIKEN Institute in Japan. This architecture has been custom designed to calculate non-bonded forces in an MD simulation and can attain a peak performance of 0.2 petaflops or more. Additionally, other requirements should include:

- network bandwidth and massive data storage to manage large data sets
- maintain memory—at least 4 Gbyte/core up to 6–8 Gbyte/core would be preferred; > 6 Mb cache
- fast interconnect speed and I/O
- speed of processor favored over number of processors; optimize delivered performance
- high-performance hardware for scientific visualization
- hybrid architectures.

2.3 Environmental Sciences

In the near future, the DOE must make landmark decisions regarding cleanup alternatives for many existing sites, future storage repositories for high-level nuclear waste, the development of alternative energy sources (e.g., methane hydrates, oil shales, and geothermal), and amelioration of global climate change (e.g., geologic sequestration of CO_2). More importantly, DOE must defend these choices before Congress, federal regulators, and the public and confirm any choices made are based on sound science. By incorporating increasingly accurate mechanistic descriptions of physicochemical processes, subsurface models are becoming more realistic, assimilating new data of enhanced resolution and quality. As researchers develop increasingly sophisticated and mechanistic simulation tools that leverage HPC to more accurately describe subsurface processes, it has improved the subsurface scientist's ability to develop realizations of large and complex data sets and simulate these out from tens to hundreds of years or more within reasonable turnaround times.

The environmental science panel identified major research themes to be CO_2 sequestration, alternative energy systems, and subsurface and groundwater remediation. Research in these areas continues to drive the need for future complex systems level descriptions of the subsurface across various length and time scales, as well as the requirements for HPC resources at EMSL.

2.3.1 Science Drivers

Carbon dioxide sequestration in geological systems

Capturing and storing CO_2 and other greenhouse gases in deep geologic formations represents one of the most promising options for mitigating the impacts of greenhouse gases on global warming due to the potentially large capacity of these formations and their broad regional availability. CO_2 sequestration technology is still unproven and conveys significant risk to stakeholders. The critical issue is to demonstrate that CO_2 will remain stored over the long term in the geological formation where it is injected. This requires a fundamental understanding and simulation of the physicochemical processes from the molecular to macroscopic scale.

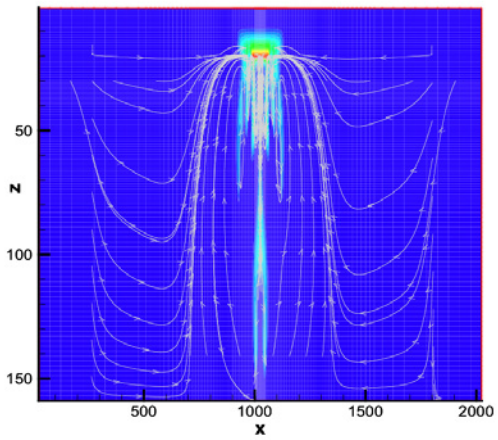


Figure 2.13. PFLOTRAN simulation applied to CO_2 sequestration in a deep aquifer (*Image courtesy of P. Lichtner, LANL*).

At the molecular scale, mineral-fluid interactions are of prime importance since such reactions can result in the long-term sequestration of CO_2 by trapping in mineral phases, such as carbonates, as well as influencing the subsurface migration of the disposed fluids via creation or plugging of pores or fractures in the host rock strata. Previous research on mineral-fluid interactions for subsurface CO_2 storage has focused almost entirely on the aqueous phase, i.e., reactivity with aqueous solutions or brines (Figure 2.13) containing dissolved CO_2. However, interactions with near-to-water-saturated non-aqueous fluids are of equal, if not more, importance over the long term for several reasons. First, the introduced supercritical CO_2 ($scCO_2$) is less dense than the aqueous phase (or oil if injected into an empty petroleum reservoir) over almost all conditions relevant to geologic sequestration. Consequently, a buoyant $scCO_2$ plume forms and ultimately dominates the contact area with the isolating caprock. Second, the injected $scCO_2$ is likely to contain water initially or soon after injection. The $scCO_2$ is highly diffusive, owing to its low viscosity, and the presence of water alters the mineral wettability of the $scCO_2$ phase. Consequently, it is possible that wet $scCO_2$ will permeate the overlying caprock pore network to a greater extent than previously considered. Third, mineral reactions in near-to-water-saturated non-aqueous fluids fundamentally differ from those in aqueous solutions. Specifically, direct interaction of the $scCO_2$-dominated fluid with mineral surfaces results in localized mineral replacement or transformation reactions unlike dissolution/re-precipitation reactions involving solution phase ion transport. Furthermore, reactive layers that form in $scCO_2$-dominated fluids cannot back dissolve into a solution phase. Such transformation reactions are more likely to result in pore or fracture plugging than the corresponding reactions in aqueous solution. Alternatively, low-water $scCO_2$-dominated fluids are expected to drive some mineral transformations via dehydration/hydration processes (such as interlayer water stripping from clays), which could lead to mineral volume decrease and enhanced permeability in caprocks. Collectively, mineral interactions with near-to-water-saturated $scCO_2$-dominated fluids are pivotal and could result in either the stable sequestration of CO_2 by trapping in mineral phases, such as metal carbonates, or degraded seals via creation of permeable zones in the caprock. To provide critical data for transport models and appropriate decisions, it is necessary to unravel molecular mechanisms governing the reactivity of mineral phases important in the geologic sequestration of CO_2 with variably wet $scCO_2$ as a function of temperature, pressure, mineral structure, and solution phase composition and to use this data in reactive transport models to make reliable predictions.

Macroscopic scale evaluation of CO_2 sequestration will require simulating the fate of injected $scCO_2$ in 3-D basin-scale (tens to hundreds of kilometers) domains to determine the displacement of brine, chemical interactions, and leakage to the

surface. In these simulations, multiphase, multicomponent, reactive fluids flow over large extents within geologic formations, which can experience geomechanical deformation. Specialized Equations of State (EOS) that span a wide range of temperatures and pressures are used to determine equilibrium conditions between gas phase CO_2 and surrounding waters and brines. Simulators must account for mass, heat, and mechanical stress balances within the domain of interest for all fluid, solute, energy, and structural equations, resulting in larger numbers of tightly coupled degrees of freedom in the systems of equations being solved. Grid resolution must be sufficiently fine, likely through adaptive mesh refinement (AMR), to capture injection wells and geologic features, simulate the response to invading CO_2, and preserve numerical accuracy. The success of geologic sequestration of CO_2 will depend on understanding: 1) changes in porosity that affect storage capacity, 2) multiphase flow in the context of additives and catalysts to improve CO_2 stability, 3) the impact of biological and chemical reactions on reservoir permeability (particularly caprock permeability and fracture sealing), 4) buoyancy effects and capillary trapping, and 5) rates of mixing between reacting and non-reacting fluids. The response of a reservoir to injected $scCO_2$ involves both spatial and temporal scaling challenges. To date, simulations have relied on a number of simplifying assumptions to make the modeling tractable. Improved understanding of the behavior of CO_2 reservoirs will require a comprehensive multiscale computational framework. In 3-D, simulation of CO_2 sequestration is impossible without the use of HPC to refine grid resolution and incorporate adequate coupling of processes to improve accuracy.

Minerals as records of atmospheric compositional variations on geologic timescales

The history of the composition of the Earth's atmosphere on geologic time scales forms a rich source of information with implications for the consequences of anthropogenic changes in the amount of CO_2. This knowledge can help answer questions such as how much the average temperature of the Earth will rise as the partial pressure of CO_2 increases and how the distribution of temperatures changes from the poles to the equator as a result. Atmospheric samples trapped as bubbles in ice cores have been a rich source of information, but they only reach back about half a million years in time, during which the atmospheric pressure of CO_2 appears to have been anomalously low. CO_2 levels can also be recorded in mineral phases growing in contact with atmospheric and marine reservoirs over time scales going back as far as 500 million years. The record here is not via direct "bubbles" of atmospheric composition, but it is imprinted in the mineralogy at a molecular level. According to best current estimates, CO_2 levels may have been 15 to 20 times higher in the past than have been characteristic of the Anthropocene.

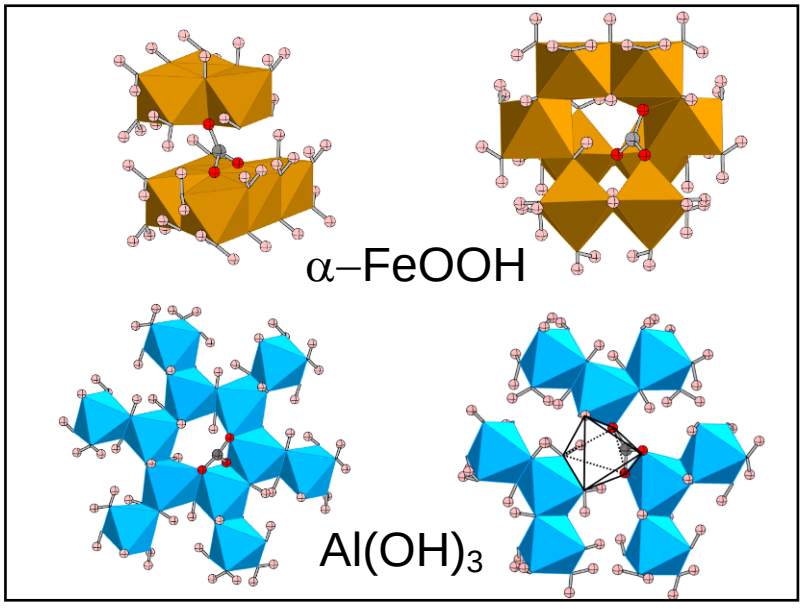


Figure 2.14. Structures of CO_2 defects in two oxy-hydroxide minerals. Carbon is grey, and oxygen is red (*Image courtesy of J. Rustad, University of California, Davis*).

To effectively read this "molecular stratigraphy" in the geologic record requires improved molecular-level knowledge of how atmospheric signatures are recorded in minerals. Currently, mineral-based estimates of the pressure of CO_2 are derived either indirectly through the influence of ocean pH on the boron-isotope composition of borate included in marine calcite or through the carbon isotope composition of the CO_2 component of pedogenic oxy-hydroxide minerals, such as goethite (α-FeOOH) and gibbsite ($Al(OH)_3$) (Figure 2.14). While this is a promising approach, the degree of uncertainty in CO_2 estimates based on these techniques needs improvement. Computational

methods already have made important contributions in this direction. For example, quantum chemistry calculations were instrumental in discovering errors in the equilibrium constant for boron isotope exchange between boric acid and borate ion in seawater. These calculations, together with sensitive new experimental measurements, have yielded a new equilibrium constant that changes predicted Miocene (15 million years) ocean pH levels by an order of magnitude. Because such changes are probably geologically unrealistic, much more work needs to be done to understand the molecular-level mechanism of boron incorporation into calcite and other marine minerals and how this might affect its boron-isotope signature. The boron-isotope pH proxy only reaches back 20 million years since it is limited by the residence time of boron in the oceans.

A second source of information on atmospheric CO_2 levels on the time scale of 10–100 million years is the isotopic composition of CO_2 incorporated into iron oxide minerals in paleo-soil profiles recorded in the geologic record. This technique depends crucially on distinguishing distinct isotopically heavy atmospheric CO_2 from isotopically light microbial CO_2. Until recently, it was assumed there was no isotopic fraction induced by incorporation of CO_2 in goethite. Recent *ab initio* MD calculations of CO_2 defect structures in oxyhydroxide minerals, coupled with quantum chemical calculations of the isotopic compositions associated with defects, suggest two distinct defect structures with different carbon isotope compositions—one of them light enough to be mistaken for organic carbon. Recent high-resolution temperature-programmed two-dimensional (2-D) correlation infrared (IR) spectroscopy of the goethite-CO_2 system, conducted by experimentalists at EMSL, provided vibrational frequencies in good agreement with calculated frequencies in both defect structures, suggesting the carbon isotope signature could be heterogeneous and dependent on growth conditions favoring one or the other structural types of defects.

Refinements of the paleo-soil oxyhydroxide CO_2 proxy would benefit from computational and experimental reactive transport capabilities at EMSL that could be used to explore diurnal/seasonal fluctuations in temperature and variations in barometric pumping. These types of issues could affect the uptake of CO_2 at different levels of pore-scale heterogeneity and oxyhydroxide mineralization. Thus, there is significant interest in being able to predict isotope fractionation factors not only for carbon and boron, but also for a wide range of metals including iron, magnesium, and calcium.

Alternative energy resource development in geologic systems

Over the past few years, geothermal energy has regained the spotlight as a potential source of energy. Subsurface science plays a critical role in evaluating the viability of such technology. Numerical simulations analyzing the viability of geothermal energy systems must account for more than solely groundwater flow and heat transport. Chemical interactions between injected/extracted waters, well bores, and mineral formations must be included along with consideration of alternative fluids such as CO_2. Simulations must incorporate the response of geologic formations to stresses and strains introduced by circulating fluids, and fractured porous media must be coupled to less permeable matrix continua, all of which are inherently 3-D in nature.

Additionally, DOE is exploring the feasibility of alternative energy resources, such as oil shale and submarine methane hydrates. Designing approaches for exploiting subsurface carbon-based energy resources depends heavily on process coupling across multiple scales. Examples include the installation of freeze walls and the sequence of heat-driven reactions involved in oil shale development, optimization and performance of geothermal reservoirs affected by fluid-rock interactions, and the impact of procedures and amendments used to modify permeability and fluid flow. High-fidelity simulations have been extremely computationally intensive, requiring several months to execute. Along with consideration of alternative fluids such as CO_2, chemical interactions between injected/extracted waters, well bores, and mineral formations must be included. Novel multiscale approaches will provide valuable insight on how to improve the efficiency of simulations.

Environmental remediation of heavy elements in geologic systems

DOE faces the imminent and expensive cleanup of existing contaminated sites within the DOE complex. For instance, a critical issue at DOE waste sites is determining whether natural processes are sufficient to contain or remediate contaminants, or if physical, chemical, and/or biological intervention is necessary to sequester or remove contaminants *in situ*. The predictive capability provided by subsurface simulation can serve to greatly reduce both cleanup costs and the DOE's liability to stakeholders. These simulations must be founded upon the best defensible science, accurately describe the contaminant behavior at molecular and macroscopic length and timescales, and include the complex physicochemical processes to capture reality.

The ability to reliably and readily predict the properties and behavior of heavy elements at the molecular scale and provide new insights into materials for storing nuclear waste and the degradation processes that can occur in waste storage tanks and the environment requires using advanced computational methods on petascale and higher performance computers. The chemical behavior of the heavy elements in the environment depends on their oxidation state and speciation, which relies on the available mineral surfaces, counter ions in aqueous solutions, and the pH. A critical need is the ability to reliably predict the electronic structure of chemicals and materials containing heavy elements to obtain reliable information about the thermodynamics of the systems and the kinetics of critical reactions and processes, especially those in solutions and at mineral interfaces. Materials composed of atoms and molecules with open 4*f* and 5*f* shells exhibit strongly correlated electron behavior, which, thus far, has prevented reliable predictions of how the physical properties of a material system changes in response to external conditions such as temperature, pressure, and impurities (Figure 2.15).

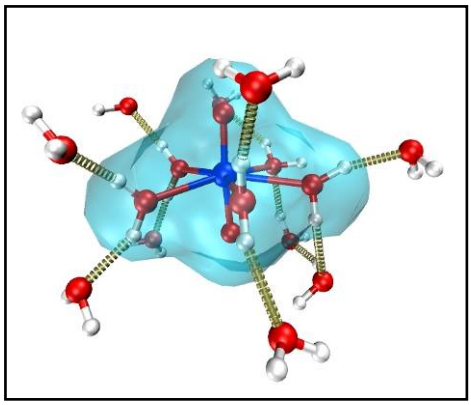


Figure 2.15. Equatorial part of the 2nd solvation shell of uranyl cation from a 23 ps Car-Parrinello simulation (*Image courtesy of W.A. de Jong and E. Bylaska, PNNL*).

Analytical methods that separate and identify species in complex mixtures are often tailored to the chemical composition of the sample to maximize the key intermolecular forces that dictate separations mechanisms. These mechanisms are difficult to elucidate from empirical observation and are ideally suited for study by computational approaches that combine multiple spatial and temporal scales. While much progress has been made toward "linking across scales," the accuracy of these multiscale approaches is dictated by the correct description of intermolecular forces, which may be weak and span long distances. Consider solvent extraction of groups of metals within the nuclear fuel cycle. Here, metal-specific chelating ligands in an optimal organic medium are added to a mixture of aqueous metals, generally under extreme pH, to create a biphasic system wherein the metal-complexes migrate to the organic phase. Understanding and controlling the intermolecular forces that dictate the chemistry at the organic-water interface are crucial to a successful separation. Currently, multiscale descriptions of separations processes are hindered by the challenge of accurately describing intermolecular interactions, either classically or using *ab initio* methods employing realistic systems that mimic experimental conditions.

A predictive capability (e.g., chemical accuracy for equilibrium constants and rate constants) for modeling solutions and interfacial phenomena for actinide-containing systems under extreme conditions of pressure, temperature, pH, and high radiation fields for aqueous media, as well as other solvents and media such as molten salts and ionic liquids, can lead to advances that will provide 1) a predictive capability for nuclear materials under "real" irradiated conditions, enabling the design of separation systems for current and future fuel cycles and materials for waste management, and 2) input data for large-scale simulations of nuclear plants and separation plants in terms of construction design, optimal operating

conditions, and catastrophic events. This work will have a substantial impact on the design of fuels, separations systems for current and future fuel cycles, and waste systems.

Simulation of contaminant fate and transport at the field scale depends on the contaminants involved and the complexity of site geology. Within the DOE complex, there are clearly perplexing contaminants than persist in domains composed of complex geology and surrounded by intricate boundary conditions (e.g., uranium(VI) [U(VI)] contamination at the Hanford 300 Area). At the Hanford 300 Area, PFLOTRAN has been employed using HPC to provide an increasingly mechanistic description of river/groundwater interaction and the physical and geochemical processes that control U(VI) persistence at the site. A critical component in the success of this work is the use of a fully 3-D flow and transport model with high temporal and spatial resolution. In this conceptual model, 15 chemical components are coupled through transport and geochemical reaction within an extremely dynamic groundwater flow field to describe the sorption of U(VI) to fine-grained sediments at the site. Simulations employing this 28-million degrees of freedom model require six to 12 hours on 4,096 processor cores to complete a year of simulation time with hourly time steps. Simulations of this scale or larger will become increasingly common at select contaminated DOE sites in the future.

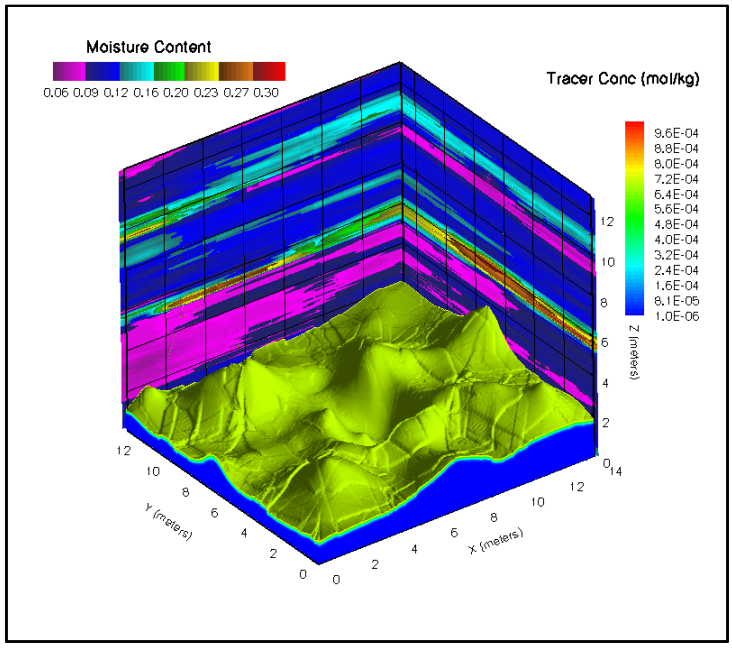


Figure 2.16. Migration and reaction of strontium-magnesium solution through the vadose zone. The green isosurface is the tracer front after migrating downward for 475 days (*Image courtesy of S. Yabusaki, PNNL*).

To reduce future cleanup costs, it is becoming increasingly clear that risk-based corrective action will be employed at contaminated sites, with diminished funding shifted to higher risk sites. In order to categorize sites based on risk, scientists must address parameter sensitivity and quantify uncertainty within risk-based models, both of which require numerous simulations, sometimes on the order of hundreds to thousands. These simulations are referred to as multi-realization simulations. To date, algorithms exist for such multi-realization simulations, but most employ a single processor core for each simulation and are limited to 2-D or simplified 3-D domains that reside within the memory available on a serial workstation (Figure 2.16). With PFLOTRAN, it has been demonstrated that tens of thousands of processor cores can be employed to simulate multiple parallel simulations of high-resolution 3-D, variably saturated flow, and multicomponent geochemical transport (i.e., 10 simulations run simultaneously under a single 40,960-core job, each repeatedly solving nonlinear systems of equations composed of 28-million degrees of freedom using 4,096 cores for 11 to 12 hours). This capability will reduce model uncertainty, solidifying the defensibility of simulations performed in support of risk-based corrective action.

An additional example of the complexity in such remediation models is the inclusion of the behavior of microbes in the subsurface. For instance, with respect to uranium bioremediation, molecular biological techniques (genome sequencing, gene expression, and transcriptomics) have provided the basis for important modeling assumptions, such as the ability for metal- and sulfate-reducing microorganisms to be present and actively participating in terminal electron accepting reactions. Current modeling of biologically mediated reactions is crude and empirical. This is due to the variable stoichiometry of the reactions (i.e., the energy for cell synthesis is a function of metabolic status, which is not addressed in the model) and the simplistic first order rate laws, which do not account for threshold behavior.

2.3.2 Computational Challenges

Realistic molecular scale models

Substantially increased computational resources are needed to address considerable issues dealing with solvation processes and predicting reaction equilibria and rates in solution. A critical issue that must be addressed is the ability to predict, manage, and control the effects of entropy in condensed media and at interfaces to predict rate and equilibrium constants that require free energies. Today, microscopic solvation environments for single ions can be predicted with reasonable structural, thermodynamic (±2 kcal/mol), and ground state spectroscopic properties. Still, how to predict solvent effects reliably at different concentrations, temperatures, pressures, pH, ionic strengths, and media is an open question. Today, the rates of simple reactions in the gas phase can be predicted using electronic structure theory to evaluate the PES and a kinetic theory, such as TST. It is necessary to predict the rate of the reaction and absolute free energies and kinetics of binding in solution to the same accuracy, and this is not routinely possible. Today, the energetics of excited electronic states for reasonable-size molecules can be predicted with different electronic structure methods, but it is difficult to obtain 0.1 eV accuracy. Even so, the common time-dependent DFT methods cannot treat two-electron excitations or long-range charge-transfer. Methods are needed to predict excited electronic states and properties in solution reliably, especially for the complex spin states. Although the acidity of most metal dications in water can be calculated to within about 1 pK_a unit, it does require MP2 calculations with a reasonably sized basis set (at least, aug-cc-pVTZ). However, the prediction of the pK_a of a +3 metal ion in water requires at least a second solvent shell at the MP2 level, and more positively charged ions may require even larger explicit solvent calculations. In addition, it is well-established that the treatment of anions requires substantially larger amounts of explicit solvent for reliable predictions. Such calculations push current capabilities in terms of the sampling issue compounded by the fact that it is not yet known how many explicit water molecules are adequate.

Micellar and colloidal systems are ubiquitous in subsurface and environmental systems, as well as playing a key role in many waste treatment and environmental remediation strategies. Simulations of these systems require large numbers of atoms to represent the molecules that are part of both the micelle or colloid and the solvent that compose these complex inhomogeneous systems. In addition to the large spatial extent needed to capture structures properly, the simulations are further complicated by the fact that many of the phenomena of interest in these systems occur on a very slow time scale, at least from the point of view of traditional MD simulations. Explicit simulation of atomic motion usually requires time steps on the order of femtoseconds, so events occurring on even a microsecond time scale require on the order of billions of time steps to simulate directly. Dynamical phenomena of interest to characterize include transport of species through membranes, colloid formation, micelle formation, and evaluation of the critical micelle concentrations. Simulation at the microsecond time scale is being targeted by several groups, including the developers of NAMD (Illinois), Blue Matter (IBM), and Desmond (D.E. Shaw), primarily in the context of protein folding. In addition to requiring very highly scalable code, performing these calculations on long time scales will require large numbers of processors. Alternative simulation approaches can be used to address the structural and dynamical properties of these systems that, nonetheless, remain computationally intensive. For example, if a low dimensional coordinate surface can be identified, then transition state approaches can be combined with free energy evaluation techniques to characterize rates. Free energy calculations can be structured to realize large amounts of parallelization, so they can be distributed over large numbers of processors. Recent work on nucleation in the gas phase has identified free energy-based methods that could be extended to condensed phase problems, such as colloid formation and micelle formation. Variations in the solution, such as pH, ionic strength, and biphasic makeup, along with variability in the composition of the amphiphilic molecules, will require large amounts of computational resources to sample the spatial and temporal extents of these systems properly.

In current approaches, *ab initio* methods are combined with classical statistical simulations to probe the fundamental intermolecular interactions and mechanisms that govern separations processes. Within this arena, accurate representations of intermolecular interactions (particularly weak, long-range forces) are elemental to obtaining realistic multiscale descriptions. In the case of condensed phase separations, DFT is used to perform large cluster calculations to elucidate the electronic and structural effects of intermolecular forces. Subsequently, specialized density functionals or correlated wave function methods are implemented to map the PES for solvent and ligand exchange reactions under different external perturbations (e.g., acid anions in the solvent shells). These PESs are used to fit highly accurate force fields designed to capture the essential physics of these interactions, e.g., to probe the kinetic and thermodynamic effects of weak intermolecular interactions of bulk phenomena. By systematically incorporating new physics within the intermolecular potential, their effects on the thermal averages of observables within a statistical simulation will lead to a better understanding of implications of accurate intermolecular descriptions upon large-scale simulations. This computational approach can be applied to understanding the intermolecular forces at the organic-water interfaces to characterize the fundamental reaction mechanisms for biphasic solvent extraction relevant to the nuclear fuel cycle. The transport process of actinide ions and their ligand complexes across water-alkane interfaces is determined as a function of 1) interfacial order and packing and 2) hydrophobic and hydrophilic solvation characteristics.

Owing to the sheer scale of the condensed phase systems, electronic structure-based methods will be somewhat limited in scope but are essential because they provide the necessary benchmarks for the more prevalent simulations that will dominate the field—those based on classical statistical mechanical methods. Continued efforts within force field development have been dedicated toward creating force fields that accurately represent intermolecular interactions within separations-based systems. However, such potentials may be complex, and new algorithms are needed to turn on and off these interactions dynamically to maximize the efficiency of the simulation. Ongoing developments to make on-the-fly interaction modification possible include using graphical algorithms that follow how all particles of the system interact, e.g., the H-bonding in solvent shells around solutes or association at the interfacial layer where two immiscible liquids meet.

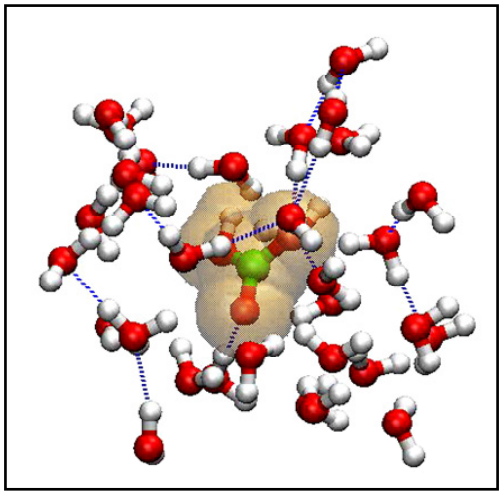


Figure 2.17. Representative 32-water Cluster for HCO_3^-(aq) (*Image courtesy of Dixon Group, The University of Alabama*).

The calculation of geochemical isotope fractionation in aqueous interfacial systems requires two components: 1) generation of candidate structures using MD from both *ab initio* and classical simulations and 2) accurate calculation of equilibrium fractionation factors, which requires calculating the most likely vibrational partition function at the *ab initio* electronic structure level. In complex interfacial environments, the former requires a combination of parameterized MD with reactive force fields and *ab initio* MD to find key interfacial and bulk defect structures that incorporates the recording molecule (borate or CO_2) or atom, e.g., the metals. For the latter, reliable calculations of aqueous boron and carbon isotope fractionation at the required 0.001 percent (parts per million) level require high-level quantum chemistry calculations. Thus, there is a clear need to be able to carry out vibrational frequency calculations with high-level methods (at least MP2/aug-cc-pVTZ) in interfacial and solution environments requiring at least 50 to 100 atoms. Large cluster models of minerals are needed, for example, so the vibrational frequencies can be calculated at the same level for all species in any environment. For instance, the prediction of the isotope fractionation factors of the critically important $CO_2/HCO_3^-/CO_3^{2-}$ system requires the ability to predict reliable frequencies in aqueous solution (Figure 2.17). Currently, it is impossible to predict the harmonic frequencies for such a system at the MP2/aug-cc-pVTZ level, much less predict the anharmonic frequencies that may be needed. In homogeneous aqueous and bulk phases, boundary conditions can often be simple. For example, if the

system size is large enough, continuum solvation models may be useful. However, the representation of interfacial environments will require much larger systems and will benefit from substantial increases in computational power, as well as using methods such as QM/MM, which have been recently implemented in NWChem.

The study of the carbon sequestration processes will require a combination of electronic structure calculations with MD simulations. Quantum chemical calculations of critical regions of the PES for the reactions of CO_2 and H_2O in solution and supercritical media, as well as with these media and surfaces/interfaces, will be needed. The calculations cannot be performed with most current DFT exchange-correlation functionals as they cannot be used to describe the important non-bonding features, especially in CO_2-rich environments. Thus, these calculations need to be completed at the correlated molecular orbital theory level (the minimum level is MP2, and CCSD(T) may be needed for accurate energetic predictions) with large basis sets (aug-cc-pVTZ and higher). Because of the importance of interpreting experimental vibrational spectroscopy measurements, it will be necessary to go beyond the harmonic approximation in the vibrational frequency calculations. An important issue is the appropriate sampling of phase space for the different molecular configurations in weakly interacting systems for which there will be many low energy conformations. It will be necessary to predict chemical reactions on representative mineral surfaces, which are predominantly some form of a metal (including silicon) oxide, and the effect of water layers on the reactivity of the mineral surfaces.

Many of the reactions important in CO_2 sequestration will involve proton and electron transfer, and it will be necessary to incorporate new techniques to deal with the quantum dynamics of these processes. For the light atom dynamics, e.g., the motion of H atoms or protons, path integral approaches can be used to treat the quantum dynamics. This involves the calculation of many trajectories simultaneously, substantially increasing the cost of the simulation, but also being highly parallel.

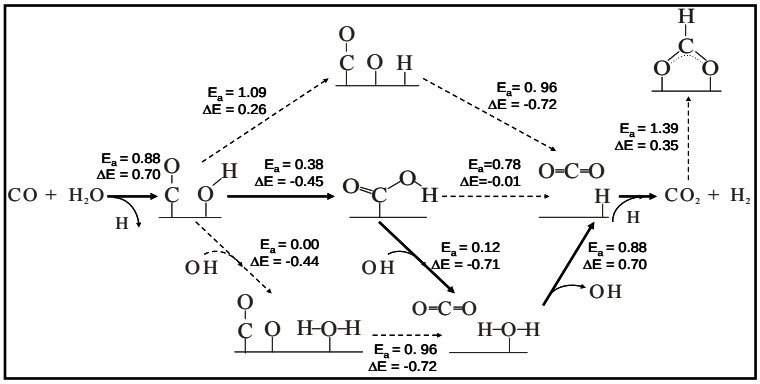


Figure 2.18. Mechanism of the water gas shift reaction on Pt with calculated DFT energies (*Diagram courtesy of M. Mavrikakis, University of Wisconsin-Madison*).

The first-principles prediction capabilities are an essential starting point for incorporation into averaged (thermodynamic) models, which are required for practical long-term prediction. In the geosciences and subsurface modeling already described, EOS are used to summarize the averaged behavior at the atomic level by providing thermodynamic potentials and related kinetic properties as a function of pressure, temperature, and composition, which are required as input to higher-level predictive tools (e.g., reactive flow simulators). While the development of EOS has been a topic of research for many years (Figure 2.18), adequate EOS for many of the key mixtures for CO_2 sequestration are not available. These EOS, for example, must incorporate the yet-to-be-discovered behavior of the complex systems relevant to CO_2 sequestration. Unfortunately, the current methods for the development of EOS are poorly based in theory, do not provide a large range of extrapolation or interpolation, and are highly dependent on the availability of data. Thus, a key synergistic role of theory is to inform the development of EOS. Appropriate equations need to be developed that can more efficiently describe the chemical and physical properties of gaseous, liquid, and solid mixtures. For the foreseeable future, reliable EOS must be heavily based on experimental data and will require inverse modeling techniques and associated uncertainty assessments.

Compounds containing heavy elements, such as actinides, lanthanides, and third-row transition metal atoms, require a proper treatment of relativity, which includes both scalar relativistic and spin-orbit components. Notably, the inclusion of the latter is necessary to predict even qualitative trends. For example, the quantitative modeling of redox reactions, which is critical for the interpretation of speciation in the environment, requires substantially improved energetic accuracy and

the ability to treat different numbers of *f* electrons at the same level of accuracy. Here, problems associated with multiplet interactions arise that cannot be reliably captured with conventional molecular orbital or DFT approaches based on a single determinant. There is a clear need to develop new approaches that build upon existing correlation methods for treating molecules and the solid state to address strong correlations, spin-orbit corrections, relativistic effects, and the issue of multiplet complexity. New approaches that show promise include: improved DFT exchange-correlation functionals and dynamical mean field theories, quantum Monte Carlo methods, and new highly correlated molecular orbital theory approaches. The theoretical development needs to incorporate results from innovations in measurement techniques, which provide data for benchmarking the calculations. Additional experimental measurements are needed to obtain high-accuracy thermodynamic, kinetic, and spectroscopic data for benchmark purposes, especially for radioactive elements. This will help not only to benchmark and validate the methods on simple systems, but in the design of appropriate computational models and the development of new methods to estimate the error in the simulation, the error in the model, and how they propagate across scales. This could be accomplished in the new EMSL Radiological Capability and focus on thermodynamic and kinetic type measurements as an integrated capability between theory and experiment.

Computational approaches also must be able to capture the unconventional phase behavior/stoichiometry and strong correlation effects in these systems. The development of a predictive understanding of the behavior of aggregated nanophase species containing heavy elements that can form under geological conditions will have an impact on understanding the transport and modeling of environmental contaminants. These likely will be colloidal-charged nanoparticles with a significant number of heavy element atoms. Intractable, small aggregates in nuclear-waste streams can impair cleanup, forcing a low-level waste stream to be treated as high-level waste, thereby increasing treatment costs. Furthermore, the associated counter ions can play an important role. The development of a fundamental understanding of the chemistry underlying nanophase formation, structure, stability, and reactivity will have an impact on understanding transport and modeling of environmental contaminants.

To deal with the phase space sampling for these complex reactions on surfaces, there still is a need to develop new MD techniques. Understanding the role of the surface and the surrounding solvent environment to get the reaction dynamics and kinetics correct is a difficult problem because of the amount of sampling required. This is akin to the impact parameter sampling of early reaction dynamics simulations. Some examples of systems sizes requiring consideration are showcased as examples of why larger-scale computational facilities are required. For example, a 100 nm by 100 nm by 100 nm box of water molecules would contain 3.3×10^7 H_2O molecules. Neutral pH requires 10^7 H_2O molecules per H^+/OH^- pair, and the minimum number of atoms in a MD trajectory study will be 10^5 to 10^6 atoms for microseconds (10^{-6} s) with femtosecond (10^{-15} s) time steps. In addition, the formation of micelles and colloids will require the development and use of appropriate statistical mechanics sampling techniques. Additionally, new classical MD methods will need to be developed, especially with reacting, polarizable force fields. New reactive subsurface modeling systems also will need to be developed to deal with reacting multiphase flow.

The Computational Grand Challenge for environmental sciences is to achieve a predictive, systems-level understanding of complex subsurface contaminant fate and transport systems. Such systems-level understanding requires multiscale models that span a broad range of length scales (from nanometers to kilometers) and time scales (from femtoseconds to millennia). Macroscopic models rely on chemical information at the molecular scale, as well as the biological behavior of species interactions with soil and groundwater. Accurate data are needed, and great care must be taken to minimize errors in the simulation data used as input. Sophisticated environmental and chemical process model algorithms must be designed so errors do not accumulate, propagate, and ultimately invalidate the macroscopic-scale model. The continued development of efficient and accurate methods that couple phenomena from one size and time domain and pass relevant parameters to an adjacent size and time domain is paramount to successful multiscale modeling. This is germane not only to environmental sciences, but also biological and chemical sciences.

Modeling paradigms for subsurface flow and transport models

Typically, phenomena in a particular size and time domain are described with systems of PDEs. Various solution methods are employed to solve the systems of PDEs governing variably saturated groundwater flow coupled to chemical reactions, heat transport, and geomechanical deformation. Modeling approaches in reactive transport simulations generally revolve around fully implicit schemes versus operator splitting. These methods both have advantages and disadvantages. A fully implicit method allows larger time steps with the potential time truncation errors. Fully implicit schemes also require larger memory; more intensive computing needs; and, consequently, more processors compared to operator splitting. Generally, operator splitting is restricted to much smaller time steps limited by the Courant condition to avoid operator splitting errors. Using total variation diminishing (TVD) methods with operator splitting is more robust, but it requires adding dispersion explicitly to the simulation for heterogeneous velocity fields (at present, an area of research because of the generation of negative oscillations that can occur in the numerical solution). Compared to implicit methods, operator splitting algorithms may be better suited for taking advantage of heterogeneous architectures involving vector GPU processors.

Although still experimental in nature, AMR is being implemented in a number of reactive transport codes using different approaches. Issues involved include acceptable mass conserving discretization schemes across changes in grid resolution and the overhead introduced compared to a fixed grid. Potential benefits of AMR are a significant reduction in the number of degrees of freedom needed to model a large 3-D problem and the gain in resolution where it is needed.

Within a multiscale framework, it will be necessary to propagate uncertainty through process dynamics across a range of scales to accurately quantify uncertainty in predictions of system behavior and risk associated with non-compliance. The mathematical frameworks for uncertainty propagation (e.g., Monte Carlo techniques, approximate covariance propagation equations, or polynomial chaos theories) lead to computational problems that are vastly different from their deterministic counterparts. The challenge of quantifying uncertainty in nonlinearly interacting multiscale subsurface systems will require novel computational concepts and, in turn, enhance the computational science arena. Also, there is a need to develop novel flexible frameworks that implement a range of upscaling models (from classical homogenization to quasi-empirical constitutive relations to direct pore-scale simulation) that adaptively balance accuracy and computational burden.

Accurate, reliable, and scientifically defensible predictions of coupled subsurface processes are fundamental to critical decisions facing DOE. There is, however, considerable uncertainty in the predictions made using current numerical models of coupled subsurface phenomena. Unlike engineered systems, natural geologic formations are highly complex. Material properties display high variability and complex spatial correlation structures that span a rich hierarchy of length scales. In these subsurface systems, coupled subsurface processes involving multiphase–multicomponent interactions exhibit a wide variety of behaviors across a large range of spatial and temporal scales.

HPC provides the processing and large memory to simultaneously address long simulation periods, comprehensive treatment of coupled processes, the resolution of spatial and process-level details in the context of multiscale variability in material properties, and uncertainties in conceptual process models and parameters. While advances in computational science have enabled progress, two significant issues remain: 1) characterization of material properties (initial and boundary conditions at model resolvable scales) and 2) characterization of sub-grid scale process representations. Accurate and detailed characterization of spatial property distributions in the subsurface remains a first-order challenge facing reactive transport modeling efforts. Subsurface properties, such as permeability and reactivity, are known to vary over orders of magnitude across a wide range of spatial scales. This heterogeneity gives rise to preferential flow paths or barriers, incomplete mixing of reactants, limited effectiveness of remediation, and slow release of contaminants through diffusion-controlled mass transfer processes. Conversely, it often is a struggle to identify which small-scale (i.e., sub-grid

resolution) features, processes, and/or conditions impact/control the field-scale behaviors of interest. It is neither possible nor desirable to include the processes and process details for all length and time scales in mathematical models of coupled flow and reactive transport. In part, this is due to the severe computational burden of resolving a wide range of length and time scales to obtain accurate predictions of the subsurface processes. Consequently, the current practice is to develop models based on the continuum scale where Darcy's law applies, and the geometrical complexity of the pore space is averaged and replaced by empirically determined effective parameters or constitutive relations. For complex, coupled, nonlinear processes (e.g., multiphase flow, biogeochemical reactions, and geomechanics), the lack of mechanistic upscaling undermines the reliability of predictive simulations used to inform decisions affecting the environment and human health. Simulations that include as much of the correct physics and chemistry as possible on the finest appropriate scales are critical to developing the benchmarks needed for upscaling models and testing approximations. In terms of scale and components, these accurate simulations require substantial computational resources.

Another clear issue is the heterogeneity of the subsurface. Natural subsurface systems are physically, chemically, and biologically heterogeneous across a spectrum of scales. Unlike many engineered systems, the heterogeneity is large, difficult to quantify, and can evolve with time as a result of engineered intervention. Disordered pore structure and variability in physicochemical surface properties at the micron to millimeter (secondary porosity and grain surface) scale impacts the definition of Darcy-scale porous medium properties, such as permeability, dispersivity, and reaction rates. Heterogeneity at the centimeter-meter (core) scale influences equivalent permeability and dispersion tensors and reaction rates at the 10- to 100-meter scale. Geologic heterogeneity is manifest at the 10-meter to kilometer scale and ultimately impacts site-scale behavior. Currently, it is not possible, or perhaps even practical, to deterministically represent property fields (permeability, porosity, stratigraphy, composition, reaction rates, etc.) at all scales at a typical field site or formation, especially since there is a lack of experimental data for the models. It is necessary to develop valid methods for correlating the averaged properties of increasingly large volumes to the smaller-scale heterogeneities within the volumes and to understand how heterogeneous distributions of properties (e.g., porosity, reactive surfaces, and biomass) can change in response to various natural or engineered stimuli. Progress in such scaling approaches will require using advanced computing applied to the continuum of scales for different types of systems. An example of the different scales is shown in Figure 2.19.

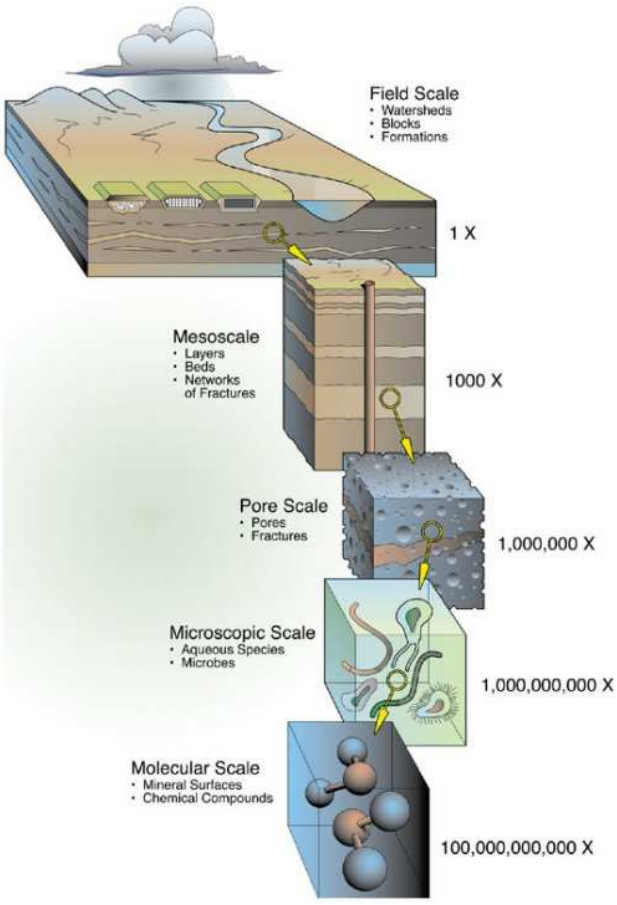


Figure 2.19. Continuum of scale for subsurface processes.

A second critical area involves process coupling. Coupling between physical, chemical, and biological processes involving multiphase–multicomponent interactions and alteration of media properties produces system behavior with enormous complexity across a large range of spatial and temporal scales. Unless there is explicit effort to obtain information at multiple scales and under ranges of conditions that can elicit the integrated response of process coupling, such variability can go unnoticed. For example, geochemical precipitation/dissolution processes and biological growth can alter pore space geometry resulting in

permeability changes, which, in turn, can affect the flux of reactive fluids and yield nonlinear feedbacks. Understanding process coupling and scaling becomes even more critical in systems that are characterized by strong, changing, and/or episodic gradients. Although they represent relatively small volumes or even micro-environments, biogeochemical reactions and physical characteristics of geologic systems along thermal and chemical gradients often can be where the most critical changes take place. Engineered and natural gradients also can change rapidly in degree and location.

In addition, the characteristic time scales for interacting processes often are highly disparate. The governing equations and/or constitutive relationships for the same process potentially can take on different forms at different scales. In a given spatial region, temporal evolution, as a result of coupled processes, can dramatically alter the local dynamics and controlling processes (i.e., associated governing equations). Understanding macroscopic phenomena (e.g., large-scale transport of fluids, heat, and reactive components in complex, geomechanical systems) in terms of fundamental processes at the scales of molecules, cells, solid-solution interfaces, and pores is increasingly being explored to address long-standing issues with the characterization of appropriate field-scale models and parameterizations. Coupled subsurface processes typically involve phenomena occurring across a wide range of scales, such that a single Representative Elementary Volume (REV) assumed to exist for many of the discrete-multiscale models may be different for different processes or may not even exist. For example, biologically mediated processes involving micro-environments surrounding biofilms or microbial colonies typically are not explicitly represented in contemporary models formulated at the Darcy- or REV-scale—nor is the evolution of these micro-environments and colony structures explicitly related to changes in REV-scale nutrient concentrations resulting from field-scale injection schemes. Significant challenges likely to be encountered in developing such a framework to address these issues include potentially different sets of governing equations at different scales involving different sets of property fields, different types of reaction kinetics at different scales, and vastly disparate time scales associated with coupled interacting phenomena.

2.3.3 High-Performance Computing Requirements

These environmental application areas share common attributes, including naturally complex, multiscale heterogeneous porous media (physical; chemical; biological properties) and coupled interacting scale-dependent processes (e.g., hydrologic, thermal, multiphase, biogeochemical, and geomechanical). As the applications of subsurface science continually increase in sophistication and complexity, they all will require petascale computing resources and beyond. The impacts these scientific modeling advances will have on DOE energy, security, and environmental cleanup missions cannot be underestimated. Costly mistakes can be avoided in the remediation of radioactive and other legacy waste that presently contaminates vast quantities of water at the Nevada Test Site and the Hanford Site. EMSL's MSC capability is providing the crucial computational power (i.e., hardware optimized for subsurface science and supporting software) to support such research.

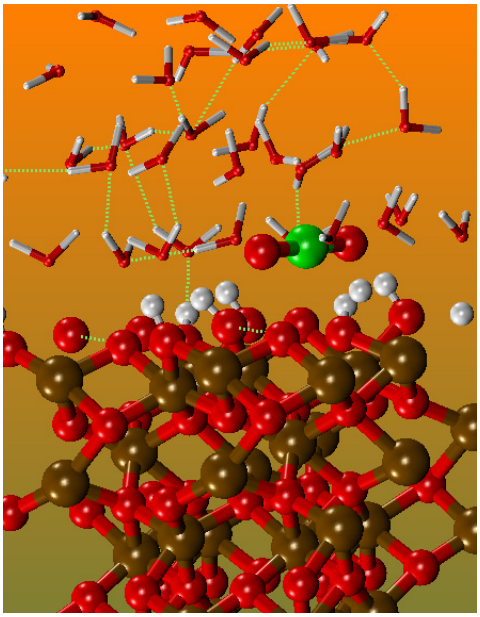


Figure 2.20. Solvated actinide species on the hydroxylated Fe_2O_3 surface (*Image courtesy of J. Rustad, University of California, Davis, and E. Bylaska, PNNL*).

Molecular scale simulations

To continue to make substantial progress in gaining fundamental insight into the molecular scale processes critical to carbon sequestration and the environmental remediation of heavy elements (Figure 2.20), computational resources in the petaflop regime, combined with the appropriate scalable software, will be required. Most of the molecular scale simulations make direct connections

with EMSL experiments in the areas of molecular structure determination, such as: crystallography and extended X-ray absorption fine structure (EXAFS) of solution structures; vibrational IR and Raman; NMR; EPR; and ultraviolet-visible (UV-Vis) spectroscopes, equilibrium measurements, and macroscopic thermodynamic measurements. These simulations provide data for input to reactive transport models, which are directly connected to field-scale experiments, as well as to column and batch experiments.

The computational resources needed for these calculations is beyond what is available at most academic institutions, and it is only through using the focused resources at EMSL that this research is able to proceed, aiding the DOE mission to safely store and dispose of radioactive waste and protecting the environment and the public. EMSL has the opportunity to provide the same focused resources to make significant advances in the area of carbon sequestration. Addressing these issues will require access to the next and future generations of hardware, as well as software that readily runs on the hardware. The paradigm must shift from solving model problems qualitatively to solving the actual problems quantitatively, which will require the continued development of electronic structure methods implemented in usable, general purpose codes on advanced computer architectures. Electronic structure methods are at the core and include: coupled-cluster methods, such as CCSD(T) and MRCI, for high-accuracy correlated results; MP2 for medium-accuracy correlated results; DFT and *ab initio* MD (Car-Parrinello) for medium-accuracy results; solid state approaches at the DFT and molecular orbital theory levels; and relativistic approaches. For electronic structure codes, most of the current algorithms are built on a cache-blocked architecture. Therefore, cache latency, bandwidth, and size are important to performance and scalability. The ability to interleave computation, communication, and I/O operations (e.g., use of asynchronous I/O; use of non-blocking communication operations) also is critical to performance. Because many of the calculations involve large matrix operations, communication bandwidth and latency are especially important. In addition, large local memories and access to large amounts of fast storage for temporary storage of intermediate results benefit the calculations.

A key issue is the ability to deal with distributed data structures on a fully distributed memory system. New approaches to electron and nuclear dynamics, especially ones that will scale to large processor counts, will be required. The same techniques described under the chemistry section will be necessary to advance the environmental sciences. These include quantum Monte Carlo and path integral Monte Carlo for the simulation of multi-electron systems. And again, the algorithms clearly show that a balanced computer architecture is needed. Balanced in terms of fast single-processor performance, low latency switches, fast I/O, and substantial amounts of memory and I/O capacity. The research also requires these resources be computationally efficient and easy to use.

In order to address the molecular-scale problems associated with the scientific drivers outlined in the environmental sciences section, a number of basic scientific advances must be made, including strategic advances in the fundamentals of electronic structure theory, kinetics, and statistical mechanics, such as:

- new DFT functionals that can be used for the reliable prediction of weak interactions, multiplet splitting, and excited states
- improved relativistic spin orbit treatments for multiplet splitting and excited states
- improved implementations of advanced high-accuracy electronic structure methods
- improved solvation models for thermodynamics to treat different temperatures, pressure, pH, and ionic strength
- improved sampling methods of the appropriate phase space for chemical reactions in and at interfaces and approaches for the quantitative prediction of reaction rates in solution at temperatures, pressures, pH, etc.
- computational techniques for long-time dynamic events, which are needed for diffusion, self assembly, self healing/repair, and kinetics

- methods to predict the properties of alternate media, e.g., molten salts, ionic liquids, or supercritical fluids.

Subsurface flow and transport simulations

In recent years, HPC has been pushing the state-of-the-art with respect to subsurface simulation using Scientific Discovery through Advanced Computing (SciDAC)-funded PFLOTRAN code. To date, simulations composed of up to 2 billion degrees of freedom have been employed at the petascale on ORNL's Jaguar XT5 supercomputer using up to 131,072 processor cores as a demonstration of capability. On the more practical side, PFLOTRAN has been executed on problems composed of tens of millions of degrees of freedom for highly transient flow and transport scenarios (e.g., the Hanford 300 Area) and hundreds of millions of degrees of freedom for steady-state runs. The key barrier to increasing problem size is a limitation in iterative linear system solver scalability and I/O.

Likely, future modeling approaches for parallel computation will remain the same, employing master-slave (for multi-realization simulations) and domain decomposition paradigms founded upon message passing to distribute the workload across processor groups and processor cores within groups. The success of future efforts to employ increasing HPC resources to the field of subsurface reactive transport will require advances in algorithmic developments (e.g., more scalable solvers, improved parallel I/O, and restructuring for GPUs), as well as advances in hardware. Although subsurface simulators using domain decomposition generally rely on nearest neighbor communication, one of the key impediments to solver scalability is the time required to calculate global reductions over larger numbers of processor cores. Iterative solvers perform various numbers of vector inner products and norms within each solver iteration—all of which require global reductions. As the number of processor cores increases, these global reductions become more expensive due to the increased communication time (e.g., higher latency due to more switches being traversed in the interconnect). To exacerbate the situation, it is well known that conventional iterative solvers fail to scale to extreme processor core counts due to breakdown in preconditioning algorithms, which results in increased iteration counts and growing numbers of expensive global reductions. Therefore, the efficacy of iterative solvers employed in extreme-scale subsurface simulations will hinge greatly upon the speed of the interconnect (i.e., lower latencies) used in future supercomputers.

Another critical feature is hardware fault tolerance for simulations using large numbers of processor cores. The PFLOTRAN code demonstrated the ability to use more than 40,000 processor cores for days for multi-realization simulation scenarios. However, at such a scale, the high probability of a single-node failure may limit HPC utility to the end user since conventional approaches to fault tolerance, such as simulation checkpointing, become unwieldy with this size of simulations and, more importantly, the increasing number of simultaneously run simulations.

Memory bandwidth and cache size play critical roles in the efficiency of algorithms employed in subsurface simulators since the expense of caching drastically reduces the effective use of floating point operations per second (FLOPs) by subsurface codes. Due to the large number of conditional statements embedded within flow, transport, and reaction algorithms, it is unlikely that increased processor speeds will improve performance. In fact, the use of slower processors will decrease power consumption and improve effective FLOPs counts. Thus, slower processors may be more practical.

Visualization for environmental science

Computationally intensive visualizations will require support from EMSL's MSC capability. Visualizations of dynamics simulations will be required because these calculations typically contain a much larger number of atoms, and many individual images in a trajectory need to be processed. In order to provide insight into the calculations themselves, the large number of atoms in dynamical simulations also means that more sophisticated visualizations will be required. This

includes viewing molecular interactions as dynamic graphs that allow the tracking of structural motifs (e.g., cliques and cycles in graph theory language), such as topologies of membrane proteins and regions of associated membrane lipids (rafts). In addition, the lighting schemes in many standard molecular displays can make it difficult to identify the important features, such as pockets and grooves in large molecular systems, which are critical to understanding their function. More sophisticated visualizations, such as ray tracing, can convey a much clearer sense of the molecular topography, but they require more computational resources to process, as well as capable rendering software. Performing these visualizations interactively will require parallel rendering capabilities, both at the hardware and software levels.

Visualization capabilities are also necessary for large continuum simulations of reactive transport in the subsurface and chemical and physical processes in the atmosphere. These simulation techniques are being pushed to include larger numbers of reactive species and use higher resolution grids on larger simulation domains. The data volumes emerging from large-scale simulations of this type are too immense to fit on a single processor and need to be distributed across multiple CPUs, which again require parallel rendering capabilities. These data sets also will require sophisticated visualizations that combine multiple surfaces using different coloring and display styles to render useful visualizations for publications and presentations. Animations add the dynamic time dimension and, therefore, require even more computation as suggested in Figure 2.21. Even more intensive renderings, such as ray tracing, will be useful in creating visualizations that reveal features in topologically complex systems containing multiple surfaces. Often, visualization is an important component in developing and debugging application software and will need to be available to developers on the computer platform itself.

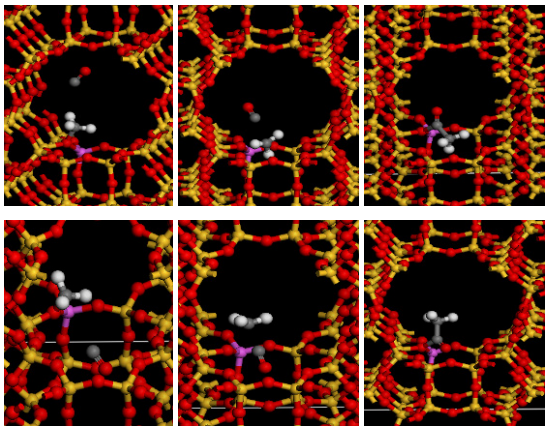


Figure 2.21. DFT-calculated carbonylation of methanol in acidic zeolites. Animation of such reactions could reveal subtle differences (*Image courtesy of M. Neurock, University of Virginia, and E. Iglesia, University of California, Berkeley*).

Currently, several parallel visualization packages are freely available, including VisIt and ParaView software. Ray tracing is available through the Persistence of Vision Raytracer (POV-Ray) software package, but it requires extensive development to use and does not provide an interactive visualization capability. To allow ready access to processors for interactive visualization, provisions will need to be made in allocating MSC capability resources.

3.0 Recommendations

Recommendations in this section summarize the HPC needs as described by the three sub-panels in biological, chemical, and environmental sciences. It reflects the clear need for the next-generation computing capability and infrastructure to enable the EMSL user community to continue to deliver world-class science.

As outlined in this *2010 Greenbook*, the computational resources necessary to make major advances toward the EMSL science drivers are beyond anything available at most academic institutions. Significant scientific discovery supporting the DOE's missions only can be accomplished through the availability of scientifically focused and optimized computational resources, such as those within EMSL's MSC capability. Next and future generations of parallel computing capabilities, fast data transfer and storage, scalable codes, algorithm development, and the availability of technical scientific consulting are all crucial components of the scientific discovery process and are especially important for paradigm shifts.

3.1 High-Performance Computing Needs

The science drivers indicate that the next-generation computing capability should be able to address the needs from two types of computing: 1) model-driven and 2) data-driven. Model-driven applications are the types of computing traditionally supported by MSC computing resources, i.e., computational chemistry, flow- and transport-models, and model-driven computing described in the *2005 Greenbook* (which drove the HPCS-3 procurement). Model-driven science requires a balanced architecture with at least an order of magnitude of performance increase in terms of fast processor performance for integer and floating point operations, low latency-high bandwidth networks, fast I/O, and substantial amounts of memory with low latency access and I/O capacity. Data-driven applications, i.e., proteomics, sequence analysis, discovery of biological networks, and image analysis, were emerging areas of computing in the MSC *2005 Greenbook* and have become an important new aspect of computing at EMSL. The different aspects of the architecture with respect to the two computing models are discussed briefly as follows.

Processor and memory

Model-driven applications require high-performance floating-point capabilities. However, the peak performance and efficiency that actually can be achieved by the algorithms underlying these applications strongly depends on the balance of floating-point operations with memory bandwidth, latency, and cache size and hierarchy. Most data-driven applications in the biological sciences area revolve around data and pattern matching and require the capability to perform high-performance integer operations.

Memory size is an important aspect for both model- and data-driven applications. The availability of four Gbyte of memory per compute core on Chinook is an essential and unique aspect of the MSC architecture that should be maintained or increased. Especially, data-driven applications have a need for large, preferably shared, amounts of memory to store the massive and complex data sets that need to be analyzed to extract scientific information. Optimally, all of a data set would fit in available local memory (RAM) and be available equally to all processor cores.

Although it is recognized these architectures have only limited demonstrated scalability, many scientific applications can benefit from shared-memory architectures. Having both memory models in a single computational environment is a desirable capability.

In the last couple of years, computer architectures have undergone a major paradigm shift that now delivers multi- and many-core systems with multiple threads per core and tens to hundreds of processing elements per node. Effectively capitalizing on these new hardware architectures is a significant issue facing scientists in planning future research at EMSL's MSC capability. Generally, the speed of a processor core is favored over number of cores in a compute node, but the trend is that hardware manufacturers are providing more cores in lieu of faster cores. This increases the need for well-designed applications to take full advantage of the hardware.

High-performance network communication

Model-driven applications require high-performance communications to share (possibly large) data blocks among computing cores. Careful attention must be paid to the network that provides interprocess communications between compute nodes. Communication patterns can be very regular or irregular and nearest neighbor or distributed across the machine. To achieve efficient and scalable communication among compute cores requires a network with high bandwidth and low latency. Because many of the calculations involve large matrix operations, communication bandwidth and latency are especially important. Network latency is the dominant factor for parallel performance for MD and domain decomposed software, such as the subsurface flow and transport simulators. Although these rely on nearest neighbor communication, a key impediment to performance and scalability is the time needed for global reductions or synchronizations over larger numbers of processor cores. The ability to interleave computation, communication, and I/O operations through asynchronous and non-blocking communication operations is critical to performance.

For the vast majority of data-driven applications, efficient interprocess communications are not an important capability requirement for next-generation MSC architecture. However, it is vital that every processor core in a job have fast access to all parts of a large shared memory space. If all of a data set cannot fit in memory local to all the processor cores, the next best alternative is to pass data between cores across a high-performance network. The high-performance network will be critical to scaling data-driven computation up to massive data sets. Bandwidth and latency likely will be important to data-driven computation.

Disk storage and access

Most model-driven applications need disk storage to accomplish their simulations. Disk storage local to the compute core facilitates fast access to intermediate data that is too large to be stored in memory. Fast disk I/O has been one of the unique hallmarks for each generation of computing capability at the MSC and has enabled scientific study of problems in sizes that are difficult to compute on other architectures. The need for local disk storage is tightly coupled with the available memory on a compute core. Of the data-driven applications, proteomics has demonstrated the need for high-performance local disk access.

Once the simulation is completed, very large, globally accessible, petabyte size disk storage capabilities are required to store the data generated by large model-driven time evolution simulations, such as MD or subsurface flow and transport, and enable further analysis. Data-driven models also require large global disk storage to enable access to volumes of experimental data from proteomics/genomic experiments by the simulation software. This storage should be available via network to all MSC compute resources. Experience with Chinook has shown that sharing one network for both interprocess communications and shared disk storage can have negative performance impact on both types of traffic. The storage network should be separate from the interprocess communications network.

The ability to access huge data sets also places requirements on the data network in the MSC capability. Users need to be able to move data to and from the global disk storage. The production of scientific data can reach many petabytes.

Therefore, the storage capability of MSC's Aurora archiving system will need to be increased appropriately with relatively fast data transfer rates.

Other considerations

For a next-generation computing capability based on high-end multithreaded commodity processors, the workhorse for most scientific computing applications, a 10x increase in level of performance is realistic and will significantly advance scientific discovery at EMSL. The compute capabilities will enable simulations to become more interactive because computational tasks that previously took hours now take minutes, and simulation work that previously took days now can occur overnight.

The proliferation of GPGPU hardware and GPGPU-enabled applications and algorithms that deliver one to two orders of magnitude speedup in performance over conventional processors have the potential to affect scientific research disruptively and fundamentally by removing time-to-discovery barriers. Computational tasks that previously would have required a year to complete can be finished in days. Better scientific insight becomes possible because researchers can work with more data and have the ability to use more accurate, albeit computationally expensive, approximations and numerical methods. For the experimentalist in particular, the results of new high-throughput instruments (or collections of many instruments) can be used to create higher-resolution, more informative pictures of what is occurring in nature, potentially in real time.

To benefit from such disruptive technologies requires a commitment to rewrite many of the computationally intensive portions of current applications to engage a large number of simultaneous threads effectively. As massively parallel hardware becomes more inexpensive, capable, and ubiquitous in the worldwide scientific community, rewriting certain computational applications appears necessary to keep computation-dependent research at the MSC capability competitive.

Another critical feature is hardware fault tolerance for simulations using large numbers of processor cores. A single-node failure currently leads to a failed simulation, which needs to be rerun and often requires the user to wait (again) for scientific results to be calculated. Despite the hardware failure, fault-tolerant scientific applications allow a simulation to complete and enable users to get the fastest time-to-solution for their large scientific problems.

3.2 Infrastructure

MSC's HPC capabilities, leading-edge scalable software, scientific and technical expertise, and data analysis and visualization all are crucial components necessary to allow users to deliver scientific discoveries that impact DOE and the nation.

Staff

To enable scientific discovery, MSC's computational resources need to be easy to use and computationally efficient. The MSC staff consists of experts with extensive knowledge and experience in HPC and operations and domain experts with detailed scientific knowledge to support EMSL's users.

- MSC's HPC Operations staff deliver a stable, robust, and highly available hardware and system software environment to users. These staff members are experts who have delivered solutions and software capabilities to improve performance that are now in use at other computing facilities across the United States. Staff responsible for the

Aurora archive are instrumental in determining that EMSL's scientific data is safe and easily and quickly available to EMSL's users.

- Capitalizing on next-generation hardware technologies is a significant issue facing scientists in planning future research using EMSL's MSC capability. HPC experts are essential to support users with improving the scalability and performance of their software and development of advanced new algorithms that make efficient use of the computing architecture.
- Scientific domain experts with knowledge of scientific applications and HPC are critical in helping and guiding expert, non-expert, or experimental EMSL computing users in designing simulations, achieving fastest time-to-solution, and integrating theory and experiment across EMSL.

Data analysis and visualization

The next generation of simulations will use and produce large volumes of data that need to be analyzed to extract scientific knowledge. Analyzing these large data sets will require sophisticated, computationally intensive data analytics software tools and visualization capabilities, as well as support from EMSL's MSC. Because these calculations typically contain a much larger number of elements and many individual images in a trajectory that need to be processed, interactive visualization of time-evolved dynamics simulations also will be required. Additionally, these capabilities are necessary for large continuum simulations of reactive transport in the subsurface and of chemical and physical processes in the atmosphere. Many of the advanced visualization techniques, such as ray tracing, can reveal features in topologically complex scientific systems not accessible by conventional approaches, but they require more computational resources to process, as well as capable rendering software. Performing these visualizations interactively will require ready access to computing resources with interactive, high-performance parallel rendering capabilities, both at the hardware and software levels.

The amounts of data generated from a simulation may become so large that the traditional model of archiving data and analyzing it later may not be feasible. Parallel data analysis may need to be built into the scientific software, which means the type of information being sought must be known *a priori*.

Additionally, data analysis and visualization is an important component in developing and debugging scientific application software and will need to be available to developers on the HPC capability itself.

3.3 Summary of Recommendations

Table 3.1 reflects the essence of the recommendations outlined in this section. The reader is strongly encouraged to review the entire recommendations section of the *2010 Greenbook*, as well as the HPC needs outlined in the science drivers sections developed by the MSC Science Panels.

Table 3.1. Summary of Recommendations

High-Performance Computing Needs				
	Data-driven	**Model-driven**		
	Biological	**Biological**	**Chemical**	**Environmental**
Floating point operations	X	X	X	X
Integer operations	X			
Memory size	X	X	X	X
Memory bandwidth	X	X	X	X
Memory latency	X	X	X	X
Interconnect bandwidth		X	X	X
Interconnect latency		X	X	X
Disk storage size Local to compute Global/WAN			X X X	X X X
Disk access bandwidth Local to compute Global/WAN	X[1]	 X	X	 X
Disk access latency Local to compute Global/WAN	X[1]	 X	X	 X
Massive threading and GPGPU technologies[2]	X	X	X	X
Infrastructure needs				
Expert staff: HPC operations and systems Domain experts in scientific disciplines supported by MSC capability Parallel and new technology programming expertise				
Software: Leading-edge scalable application software Fault tolerant applications				
Sophisticated and dedicated data analysis and visualization capabilities				
Sufficient and fast accessible archive storage				

[1]Within data-driven applications, this need was identified for proteomics only.

[2]Disruptive technology with the potential to affect scientific research fundamentally by removing the time-to-discovery barrier.

4.0 Perspective: Directions in High-Performance Computing

As the second decade of this millennium dawns, HPC finds itself at a crossroads. Processors, which do the real work of computation, are becoming available in unprecedented numbers in both old and new flavors. The potential exists that up to two orders of magnitude more raw computational power is available within the same size box as compared to just two years ago. However, before this potential can be realized, crucial changes must be made to HPC software to take full advantage of the millions (and potentially billions) of threads that are available. There are two reasons for this: 1) the prevalence of multi-core processors and 2) new availability of graphical processors for computation.

A few essential components of any high-performance computer are:

- General purpose processing unit (or processor), which performs simple operations such as arithmetic or comparisons on "words" of data (words typically represent one or more letters or numbers)
- Memory, which temporarily stores data and is the fastest storage relying on efficient caches
- Network interface, which moves data from memory in one computer to memory in another computer
- Storage, which holds data for longer periods but is much slower to access than memory
- Power consumption and excess heat dissipation (which are flat at best) continue to increase.

Traditionally, HPC has focused on getting as much work as possible out of the available processors, but this situation is suddenly turning itself inside out. In the past two years, processors have become significantly less expensive and more plentiful. Now, it is becoming more difficult and increasingly important to determine that processors can use memory, storage, and network interfaces efficiently enough to remain busy. This requires new thinking about how HPC is conducted and additional investments in code to reflect that new thinking.

A few key points must be understood about the relationships between these components:

- *All* of the words a processor operates on must be fetched from memory and written back to memory.
- Memory is *much* slower than the processor. This memory bottleneck has existed for more than decade. Elaborate hardware and software mechanisms exist to attempt to hide it as much as possible, but they only mitigate the problem; they do not solve it.
- Understanding this and other bottlenecks and making appropriate compromises in system and code design is crucial to attaining high performance.
- Processors have a *clock* that regulates how fast they operate.
- Processors are meant to be general purpose devices. As such, they are designed to do many things reasonably well rather than being tuned to excel at a particular type of operation.

In the 1990s, relatively inexpensive desktop processors from Intel and AMD made massive strides in performance by continual increases in clock speed. Clock speeds increased from 20 or 30 megahertz to a few gigahertz. In doing so, they displaced more expensive processors and systems by other manufacturers. For a time, processors were expected to get faster with each new product release, and computation would speed up "for free" on new computers, i.e., with no major changes to the code. *This is no longer true!*

With its Pentium 4 product line, Intel discovered that increasing clock speeds was no longer a viable way to improve performance. The processors would waste power and generate excessive heat as they approached the 4 gigahertz clock speed. Intel and AMD stopped pursuing ever-increasing clock speeds and, instead of faster processors, offered *more* processors ("cores") on a single chip. It should be noted that while this change helped reduce the rate at which power demands grew, it did not stop growth. Typically, each new generation HPC system will draw and dissipate about twice the power of its predecessor as heat. Thus far, each new system has generated 10 or more times the performance, so this tradeoff is worthwhile. As before, the power consumption trend is expected to continue, as well as a disruptive increase in performance in the next-generation HPC system.

In the past two to five years, most HPC cluster nodes have had one or two processor sockets in them. Where there once was one processor in a socket, now there are typically four. It is evident that there will be even more (easily 12 or more) imminently. This presents two major challenges. First, most software is not written to use multiple cores. Such software could easily waste the "extra" cores on the chip, sacrificing three-quarters or more of the processors' potential performance. Second, not only is memory already much slower than processors, but now more cores are feeding from the same memory with not many additional paths to memory. At best, an incremental increase in memory speed is being made as the number of cores increases. This is analogous to a couple suddenly having to feed a family of four with only a 25 percent or 50 percent increase in budget. Significant thought and some careful software engineering must be applied to keep those extra cores from "starving" and being wasted.

The potential for more radical improvements in performance comes from the use of GPUs to do computation. Born in the late 1990s, GPU technology was developed to provide 3-D graphics for computer games. The technology revolutionized gaming, and some kind of 3-D graphics processor exists in every desktop and laptop computer sold today. The demand for ever more realistic 3-D graphics has driven the development of progressively more powerful GPUs. Now, GPUs have more transistors in them than general-purpose processors, and they are optimized to do tremendous amounts of math in a short time. This is terrific news for HPC because these devices are plentiful, inexpensive, and manufacturers (NVIDIA and AMD) now are making HPC-oriented products out of them. By using a single GPU for computation, early research results have shown some computational chemistry algorithms can speed up by a factor of 100 to 200. However, there are challenges to overcome. First, because the GPUs are specialized devices, new specialized code will need to be written to use them at all. Second, while GPUs can do much more math than a general-purpose processor, they still have to fetch all of their data from memory, and they have *even slower* access to that memory. In order to not waste its potential, careful decisions must be made concerning how data is moved into and out of the GPU.

The potential of these technology trends is the same money can buy 100 to 200 times the compute power that it could just a few years ago. The challenge is the software *must* be redesigned in order to tap this potential. This is not the first time this has happened in HPC. The last decade experienced the transition to parallel computation on clusters. The previous two decades before that saw vector computing dominate. In both cases, code had to be redesigned and rewritten to take advantage of the new hardware. To take advantage of multi-core processors and GPUs, comparable code reworking will need to be done. This will not be a nominal investment. It will take millions of dollars to update major software packages. However, the reward should be on the order of 100 or more times the computation out of a new system, which, on a multimillion-dollar HPC setup, would rapidly pay for itself both in terms of efficiency and time-to-solution.

Appendix A: List of Molecular Science Computing Science Panel Members

Biological Sciences Panel

Jeffry Madura*
Biochemistry Center for Computational Sciences
Duquesne University
jdmadura@duq.edu

T.P. Straatsma*
Computational Sciences and Mathematics Division
Pacific Northwest National Laboratory
tps@pnl.gov

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

Bill Cannon
Computational Biology & Bioinformatics
Pacific Northwest National Laboratory
william.cannon@pnl.gov

Mihaly Mezei
Department of Structural & Chemical Biology
Mt. Sinai School of Medicine
mihaly.mezei@mssm.edu

Chris Oehmen
Computational Biology & Bioinformatics
Pacific Northwest National Laboratory
christopher.oehmen@pnl.gov

Anuj Shah
Scientific Data Management
Pacific Northwest National Laboratory
anuj.shah@pnl.gov

Chemical Sciences Panel

Larry Curtiss*
Center for Nanoscale Materials
Argonne National Laboratory
curtiss@anl.gov

Ram Devanathan*
Material Science
Pacific Northwest National Laboratory
ram.devanathan@pnl.gov

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

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

Donald Thompson
Department of Chemistry
University of Missouri-Columbia
thompsondon@missouri.edu

Donghai Mei
Catalysis Science
Pacific Northwest National Laboratory
donghai.mei@pnl.gov

Kerwin Dobbs
Central Research & Development
DuPont
kerwin.d.dobbs@usa.dupont.com

* Science Panel Lead

Haluk Resat
Computational Biology & Bioinformatics
Pacific Northwest National Laboratory
haluk.resat@pnl.gov

Anant Anantram
Department of Electrical Engineering
University of Washington
anant@ee.washington.edu

Environmental Sciences Panel

David A. Dixon*
Department of Chemistry
The University of Alabama
dadixon@as.ua.edu

Glenn E. Hammond*
Hydrology Department
Energy and Environment Directorate
Pacific Northwest National Laboratory
glenn.hammond@pnl.gov

L. René Corrales
Department of Chemistry
University of Arizona
lrcorral@email.arizona.edu

Eric J. Bylaska
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
eric.bylaska@pnl.gov

Jim Rustad
Geology Department
University of California, Davis
james.rustad@gmail.com

Steve Yabusaki
Hydrology Department
Energy and Environment Directorate
Pacific Northwest National Laboratory
yabusaki@pnl.gov

* Science Panel Lead

† Chair MSC Science Panel

Speakers and Participants

Bert de Jong†
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
bert.dejong@pnl.gov

Thom Dunning, Jr.†
NCSA
University of Illinois at Urbana-Champaign
tdunning@ncsa.uiuc.edu

Dave Cowley
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
david.cowley@pnl.gov

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

Allison Campbell
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
allison.campbell@pnl.gov

Don Baer
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
don.baer@pnl.gov

Rob Farber
Global Security Technology & Policy
Pacific Northwest National Laboratory
rob.farber@pnl.gov

Niri Govind
Environmental Molecular Sciences Laboratory
Pacific Northwest National Laboratory
niri.govind@pnl.gov

Appendix B: List of White Paper Contributors

M.P. Anantram (Anant)
Department of Electrical Engineering
University of Washington
anant@ee.washington.edu

David Baker
Department of Biochemistry
University of Washington
dabaker@u.washington.edu

Eric J. Bylaska
EMSL Molecular Science Computing
Pacific Northwest National Laboratory
eric.bylaska@pnl.gov

Donald Camaioni
Chemical & Materials Sciences Division
Pacific Northwest National Laboratory
donald.camaioni@pnl.gov

Aurora Clark
Department of Chemistry
Washington State University
auclark@wsu.edu

L. René Corrales
Department of Materials Science and Engineering
The University of Arizona
lrcorral@email.arizona.edu

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

David A. Dixon
Department of Chemistry
The University of Alabama
dadixon@bama.ua.edu

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

Michel Dupuis
Chemical & Materials Sciences Division
Pacific Northwest National Laboratory
michel.dupuis@pnl.gov

Roland Faller
Department of Chemical Engineering & Materials Science
University of California, Davis
rfaller@ucdavis.edu

James Franz
Chemical & Materials Sciences Division
Pacific Northwest National Laboratory
james.franz@pnl.gov

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

Bojana Ginovska
Chemical & Materials Sciences Division
Pacific Northwest National Laboratory
bojana.ginovska@pnl.gov

Niri Govind
EMSL Molecular Science Computing
Pacific Northwest National Laboratory
niri.govind@pnl.gov

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

Glenn E. Hammond
Hydrology Department
Energy and Environment Directorate
Pacific Northwest National Laboratory
glenn.hammond@pnl.gov

Artem Oganov
Center for Computational Sciences
State University of New York
artem.oganov@sunysb.edu

Timothy Johnson
Chemical & Materials Sciences Division
Pacific Northwest National Laboratory
Timothy.Johnson@pnl.gov

David Baker
Department of Biochemistry
University of Washington
dabaker@u.washington.edu

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

Jeffry Madura
Biochemistry Center for Computational Sciences
Duquesne University
jdmadura@duq.edu

Donghai Mei
Chemical & Materials Sciences Division
Pacific Northwest National Laboratory
donghai.mei@pnl.gov

Mihaly Mezei
Department of Structural and Chemical Biology
Mount Sinai School of Medicine, New York
Mihaly.Mezei@mssm.edu

Barbara Mooney
Department of Chemistry and Biochemistry
The University of Arizona
barbaram@email.arizona.edu

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

Erich R. Vorpagel
EMSL Molecular Science Computing
Pacific Northwest National Laboratory
erich.vorpagel@pnl.gov

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

Andrew Rappe
Department of Chemistry
University of Pennsylvania
rappe@sas.upenn.edu

Simone Raugei
Chemical & Materials Sciences Division
Pacific Northwest National Laboratory
simone.raugei@pnl.gov

Haluk Resat
Computational Biology & Bioinformatics
Computational Science & Mathematics Division
Pacific Northwest National Laboratory
haluk.resat@pnl.gov

Ruben Santamaria
Department of Theoretical Physics
Instituto de Física, UNAM
rso@fisica.unam.mx

George Schatz
Northwestern University
Department of Chemistry
schatz@chem.northwestern.edu

Stephen Williams
Department of Chemistry
Appalachian State University
willsd@appstate.edu

Constantinos Zeinalipour-Yazdi
Department of Chemistry
University of Cyprus
zeinalip@ucy.ac.cy

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

Xue-Bin Wang
Chemical & Materials Sciences Division
Pacific Northwest National Laboratory
Xuebin.Wang@pnl.gov

John H. Weare
Department of Chemistry & Biochemistry
University of California, San Diego
jweare@ucsd.edu

Appendix C: List of Supporting Documents

In addition to the *2010 Greenbook* document, the following documents are contained on the CDROM in separate folders:

1. White papers
2. Presentations from the opening session of the November 2009 Science Panel held at EMSL.