DISCLAIMER

This report was prepared as an account of a workshop sponsored by the U.S. Department of Energy. Neither the United States Government nor any agency thereof, nor any of their employees or officers, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of document authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Copyrights to portions of this report (including graphics) are reserved by original copyright holders or their assignees, and are used by the Government's license and by permission. Requests to use any images must be made to the provider identified in the image credits.

Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.

Ernest Orlando Lawrence Berkeley National Laboratory

University of California

Berkeley, California 94720 U.S.A.

NERSC is funded by the United States Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program. David Goodwin is the NERSC Program Manager and James Davenport, Mark Pederson, Nicholas Woodward and Van Nguyen serve as BES allocation managers for NERSC.

NERSC is located at the Lawrence Berkeley National Laboratory, which is operated by the University of California for the US Department of Energy under contract DE-AC02-05CH11231.

This work was supported by the Directors of the Office of Science, Office of Basic Energy Sciences, and the Office of Advanced Scientific Computing Research, Facilities Division.

High Performance Computing and Storage Requirements for Basic Energy Sciences: Target 2017

Report of the HPC Requirements Review

Conducted October 8-9, 2013

Gaithersburg, MD

DOE Office of Science

Office of Basic Energy Sciences (BES)
Office of Advanced Scientific Computing Research (ASCR)

National Energy Research Scientific Computing Center (NERSC)

Editors

Richard A. Gerber, NERSC

Harvey J. Wasserman, NERSC

Table of Contents

1 Executive Summary 5

2 DOE Basic Energy Sciences Mission 6

3 About NERSC 8

4 Meeting Background and Structure 10

5 Workshop Demographics 11
- 5.1 Participants 11
- 5.2 NERSC Projects Represented by Case Studies 13

6 Findings 16
- 6.1 Requirements Summary 16
- 6.2 Computing and Storage Requirements 17
- 6.3 Additional Observations 20

7 BES and NERSC Trends 22

8 Materials Sciences Case Studies 24
- 8.1 Computational Resources for the Nanomaterials Theory Institute at the Center for Nanophase Materials Sciences 24
- 8.2 The Materials Project 30
- 8.3 Transport Phenomena in Novel Energy Materials 38
- 8.4 Computational Design of Novel Energy Materials 46

9 Chemical Sciences Case Studies 52
- 9.1 Combustion of Alternative Fuels for Transportation Systems – Fundamental Investigation using Direct Numerical Simulations 52
- 9.2 Rational Catalyst Design for Energy Production 64
- 9.3 Condensed Phase Studies with CP2K 71
- 9.4 Accurate Scalable Calculations for the Ground and Excited States of Complex Molecular Assemblies 75
- 9.5 Molecular Dynamics of PNIPAM Agglomerates and Composite Architectures 81
- 9.6 Sampling Diffusive Dynamics on Long Timescales, and Simulating the Coupled Dynamics of Electrons and Nuclei 90

10 Geoscience Case Studies 93
- 10.1 Large Scale Geophysical Inversion & Imaging 93
- 10.2 Computational Studies in Molecular Geochemistry 100
- 10.3 Direct Numerical Simulation of Poisson-Nernst-Planck Equation in Charged Clays 108
- 10.4 Imaging and Calibration of Mantle Structure at Global and Regional Scales Using Full-Waveform Seismic Tomography 114

11 Scientific User Facility Case Studies 121
- 11.1 Introduction 121
- 11.2 Advanced Light Source 122
- 11.3 Advanced Modeling for Next-Generation BES Accelerators 131

Appendix A. Attendee Biographies 139

Appendix B. Meeting Agenda 143

Appendix C. Abbreviations and Acronyms 145

Appendix D. About the Cover 147

1 Executive Summary

The National Energy Research Scientific Computing Center (NERSC) is the primary computing center for the DOE Office of Science, serving approximately 5,000 users working on some 700 projects that involve nearly 600 codes in a wide variety of scientific disciplines. In addition to large-scale computing and storage resources NERSC provides support and expertise that help scientists make efficient use of its systems.

In October 2013 NERSC, DOE's Office of Advanced Scientific Computing Research (ASCR) and DOE's Office of Basic Energy Sciences (BES) held a review to characterize High Performance Computing (HPC) and storage requirements for BES research through 2017. This review is the tenth in a series that began in 2009 and it is the second for BES. The report from the previous BES review is available at http://www.nersc.gov/science/hpc-requirements-reviews/target-2014/.

The latest review revealed several key requirements, in addition to achieving its goal of characterizing BES computing and storage needs. High-level findings are:

1. Scientists will need access to significantly more computational and storage resources to achieve their goals and reach BES research objectives.
2. Users will need assistance from NERSC to prepare for Cori (NERSC-8) and follow-on manycore systems.
3. Research teams need to run complex jobs of many different types and scales.
4. BES is a leader in innovative use of HPC and requires a diverse set of resources and services from NERSC.
5. BES facilities need computational analysis and data storage resources beyond what they can provide.

This report expands upon these key points and adds others. The results are based upon representative samples, called "case studies," of the needs of science teams within BES. The case study topics were selected by the NERSC meeting coordinators and BES program managers to represent the BES production computing workload. Prepared by BES workshop participants, the case studies contain a summary of science goals, methods of solution, current and future computing requirements, and special software and support needs. Also included are strategies for computing in the highly parallel "many-core" environment that is expected to dominate HPC architectures over the next few years.

2 DOE Basic Energy Sciences Mission

Basic Energy Sciences (BES) supports fundamental research to understand, predict, and ultimately control matter and energy at the electronic, atomic, and molecular levels in order to provide the foundations for new energy technologies and to support DOE missions in energy, environment, and national security. The BES program also plans, constructs, and operates major scientific user facilities to serve researchers from universities, national laboratories, and private institutions. The BES program funds work at more than 160 research institutions through the following three Divisions:

- Materials Sciences and Engineering Division
- Chemical Sciences, Geosciences, and Biosciences Division
- Scientific User Facilities Division

The Materials Sciences and Engineering (MSE) Division supports fundamental experimental and theoretical research to provide the knowledge base for the discovery and design of new materials with novel structures, functions, and properties. This knowledge serves as a basis for the development of new materials for the generation, storage, and use of energy and for mitigation of the environmental impacts of energy use.

The Chemical Sciences, Geosciences, and Biosciences (CSGB) Division supports experimental, theoretical, and computational research to provide fundamental understanding of chemical transformations and energy flow in systems relevant to DOE missions. This knowledge serves as a basis for the development of new processes for the generation, storage, and use of energy and for mitigation of the environmental impacts of energy use.

The Scientific User Facilities (SUF) Division supports the R&D, planning, construction, and operation of scientific user facilities for the development of novel nano-materials and for materials characterization through x-ray, neutron, and electron beam scattering; the former is accomplished through five Nanoscale Science Research Centers and the latter is accomplished through the world's largest suite of synchrotron radiation light source facilities, neutron scattering facilities, and electron-beam microcharacterization centers. These facilities provide unique capabilities to the scientific community and are a critical component of maintaining U.S. leadership in the physical sciences. Annually, more than 15,000 scientists and engineers in many fields of science and technology visit the BES user facilities.

The energy systems of the future - whether they tap sunlight, store electricity, or make fuel from splitting water or reducing carbon dioxide - will revolve around materials and chemical changes that convert energy from one form to another. Such materials will need to be more functional than today's energy materials. To control chemical reactions or to convert a solar photon to an electron requires coordination of multiple steps, each carried out by customized materials with designed nanoscale structures. Such advanced materials are not found in nature; they must be designed and fabricated to exacting standards using principles revealed by basic science.

This report highlights the growing need for state-of-the-art computational resources through such activities as the interagency Materials Genome Initiative, advanced simulation capabilities in Geoscience, and enhanced analysis and real-time simulation of data from the BES suite of user facilities.

† U.S. Department of Energy Strategic Plan, May 2011

(http://energy.gov/sites/prod/files/2011_DOE_Strategic_Plan_.pdf)

3 About NERSC

The National Energy Research Scientific Computing (NERSC) Center, which is supported by the U.S. Department of Energy's Office of Advanced Scientific Computing Research (ASCR), serves more than 5,000 scientists working on over 700 projects of national importance. Operated by Lawrence Berkeley National Laboratory (LBNL), NERSC is the primary high-performance computing facility for scientists in all of the research programs supported by the Department of Energy's Office of Science. These scientists, working remotely from DOE national laboratories; universities; other federal agencies; and industry, use NERSC resources and services to further the research mission of the Office of Science (SC). While focused on DOE's missions and scientific goals, research conducted at NERSC spans a range of scientific disciplines, including physics, materials science, energy research, climate change, and the life sciences. This large and diverse user community runs hundreds of different application codes. Results obtained using NERSC facilities are citied in about 1,500 peer reviewed scientific papers per year. NERSC activities and scientific results are also described in the center's annual reports, newsletter articles, technical reports, and extensive online documentation. In addition to providing computational support for projects funded by the Office of Science program offices (ASCR, BER, BES, FES, HEP and NP), NERSC directly supports the Scientific Discovery through Advanced Computing (SciDAC[1]) and ASCR Leadership Computing Challenge[2] Programs, as well as several international collaborations in which DOE is engaged. In short, NERSC supports the computational needs of the entire spectrum of DOE open science research.

The DOE Office of Science supports three major High Performance Computing Centers: NERSC and the Leadership Computing Facilities at Oak Ridge and Argonne National Laboratories. NERSC has the unique role of being solely responsible for providing HPC resources to all open scientific research areas sponsored by the Office of Science.

This report illustrates NERSC alignment with, and responsiveness to, DOE program office needs; in this case, the needs of the Office of Basic Energy Sciences. The large number of projects supported by NERSC, the diversity of application codes, and its role as an incubator for scalable application codes present unique challenges to the center. However, as demonstrated its users' scientific productivity, the combination of effectively managed resources, and excellent user support services the NERSC Center continues its 40-year history as a world leader in advancing computational science across a wide range of disciplines.

NERSC provides an important computational resource for BES scientists. During the 2013 allocation year, there were 290 BES projects at NERSC, which is the largest number of projects of the six Office of Science program offices. These BES projects, which consumed about 40% of the total 2013 DOE-allocated time at NERSC, supported principal investigators and approximately 1,000 graduate and postdoctoral students addressing fundamental issues in predictive materials and chemical sciences, actinide chemistry,

---

[1] http://www.scidac.gov

[2] http://science.energy.gov/~/media/ascr/pdf/incite/docs/Allocation_process.pdf
[2] http://science.energy.gov/~/media/ascr/pdf/incite/docs/Allocation_process.pdf

[3] http://www.nwchem-

energy storage, carbon capture, catalysis, combustion, geosciences, magnetism, polymer science, solar energy, and superconductivity. In addition to core research programs, NERSC resources support BES scientific user facilities, the BES accelerator and detector research program, the BES SciDAC programs, the Energy Frontier Research Centers and the Fuels from Sunlight Energy Innovation Hub.

For more information about NERSC visit the web site at http://www.nersc.gov.

4 Meeting Background and Structure

In support of its mission to provide world-class HPC systems and services for DOE Office of Science research NERSC regularly gathers user requirements. In addition to the requirements reviews, NERSC collects information through the Energy Research Computing Allocations Process (ERCAP); workload analyses; an annual user survey, and discussions with DOE program managers and scientists who use the facility.

In October 2013, ASCR (which manages NERSC), BES, and NERSC held a review to gather HPC requirements for current and future science programs funded by BES. This report is the result.

This document presents a number of findings, based upon a representative sample of projects conducting research supported by BES. The case studies were chosen by the DOE Program Office Managers and NERSC staff to provide broad coverage in both established and incipient BES research areas. Most of the domain scientists at the review were associated with an existing NERSC project, or "repository" (abbreviated later in this document as "repo").

Each case study contains a description of scientific goals for today and for the future, a brief description of computational methods used, and a description of current and expected future computing needs. Since supercomputer architectures are trending toward systems with chip multiprocessors containing hundreds or thousands of cores per socket and perhaps millions of cores per system, participants were asked to describe their strategy for computing in such a highly parallel, "manycore" environment.

Requirements presented in this document will serve as input to the NERSC planning process for systems and services, and will help ensure that NERSC continues to provide world-class resources for scientific discovery to scientists and their collaborators in support of the DOE Office of Science, Office of Basic Energy Sciences.

NERSC and ASCR have been conducting requirements workshops for each of the six DOE Office of Sciences offices that allocate time at NERSC (ASCR, BER, BES, FES, HEP, and NP). A first round of meetings was conducted between May 2009 and May 2011 for requirements with a target of 2014. This second round of reviews target needs for 2017.

Findings from the review follow.

5 Workshop Demographics

5.1 Participants

5.1.1 DOE / NERSC Participants and Organizers

Name	Institution	Area of Interest/Title
James Davenport	DOE / BES	Program Manager, Materials Sciences and Engineering Division
Sudip Dosanjh	NERSC	NERSC Director
Jack Deslippe	NERSC	NERSC User Services Group; materials science application support
Richard Gerber	NERSC	Meeting Organizer, NERSC Senior Science Advisor
Dave Goodwin	DOE / ASCR	NERSC Program Manager
Eliane Lessner	DOE / BES	Program Manager, Accelerator and Detector R&D
George Maracas	DOE / BES	Program Manager, Nanocenters
Van T. Nguyen	DOE / BES	Program Manager, Scientific User Facilities (SUF) Division
Mark R. Pederson	DOE / BES	Program Manager, Computational and Theoretical Chemistry
David Skinner	NERSC	NERSC Strategic Partnership Lead; high-throughput materials science applications
Harvey Wasserman	NERSC	Meeting Organizer
Nicholas Woodward	DOE / BES	Program Manager, Geosciences

5.1.2 Domain Scientists

Name	Institution	Area of Interest	NERSC Repo(s)
Michael Banda	Lawrence Berkeley National Laboratory	Advanced Light Source	als
Jacqueline Chen	Sandia National Laboratories, California	Combustion	mp241
Sanket Deshmukh	Argonne National Laboratory	Molecular dynamics in chemistry	m1524, m1528
Andrew R. Felmy	Pacific Northwest National Laboratory	Geochemistry	mp119
Scott French	University of California, Berkeley	Geophysics	m554
Andreas Heyden	University of South Carolina	Rational catalyst design	m1065
Paul Kent	Oak Ridge National Laboratory	Materials science	m526, m641
Yun Liu	Massachusetts Institute of Technology	Materials science	m655, m1797
Thomas Miller	California Institute of Technology	Chemical science	m822
Jeffrey Neaton	Lawrence Berkeley National Laboratory	Materials science	mp149, m716, m1793, mp173, m387
Gregory Newman	Lawrence Berkeley National Laboratory	Geophysics	m372
David Skinner	NERSC	High throughput materials science	matgen, matcomp, m1290
Carl Steefel	Lawrence Berkeley National Laboratory	Porous media transport	
Sotiris Xantheas	Pacific Northwest National Laboratory	Chemical physics	m1513, mp329, m452

5.2 NERSC Projects Represented by Case Studies

NERSC projects represented by case studies are listed in the table below, along with the number of NERSC hours they used in 2013. The BES allocation at NERSC is the largest of the six offices that allocate time at NERSC, with approximately 300 projects (about 40% of the NERSC total) and over 800M hours (also about 40%). BES and ASCR program managers, along with NERSC staff, chose participants to best represent the BES workload at NERSC. The projects listed below include one repository ("matcomp") that NERSC considers a "sponsored" project. One of two such projects at NERSC (the other is sponsored by HEP), matcomp compute hours at NERSC are allocated by BES program managers but are not part of the annual NERSC ASCR allocation target.

NERSC Project ID (Repo)	NERSC Project Title	Principal Investigator	Workshop Speaker	Hours Used at NERSC in 2013 (M)	Archival Data at NERSC 2013 (TB)	Shared Data on Disk (TB)
Geoscience						
m372	*Large Scale 3D Geophysical Inversion & Imaging*	Gregory Newman	Gregory Newman	29	0.172	0.02
mp119	*Computational Studies in Molecular Geochemistry*	Andrew Felmy	Andrew Felmy	2.2	0.449	0
m554	*Global-scale full-waveform seismic imaging of Earth's mantle*	Barbara Romanowicz	Scott French	3.0	0.081	0.1
m1792*	*Chombo-Crunch: Advanced Simulation of Subsurface Flow and Reactive Transport Processes Associated with Carbon Sequestration**	David Trebotich	Carl Steefel	63	274	2
m1516*	*Advanced Simulation of Pore Scale Reactive Transport Processes Associated with Carbon Sequestration**	David Trebotich	Carl Steefel	33	249	0
Materials Science						
m526	*Computational Resources for the Nanomaterials Theory Institute at the Center for Nanophase Materials Sciences*	Paul Kent	Paul Kent	19.4	15.4	4.1
matgen, matcomp	*The Materials Project*	Kristin Persson	David Skinner	18.8	46.1	51
m1793	*Excited-State and Charge Transport Phenomena in Novel Energy Materials*	Jeffrey Neaton	Jeffrey Neaton	18.7	45.0	0
m1797 m655	*Computational Design of Novel Energy Materials*	Jeffrey Grossman	Yun Liu	15.4	7	7
Scientific User Facilities						
ALS	*Advanced Light Source*	Michael Banda	Michael Banda	4.3	536	75
m669	*Advanced Modeling for Next-Generation BES Accelerator*	Robert Ryne	Robert Ryne	5.2	38.7	0.25
Chemical Sciences						
mp241	*Direct Numerical Simulations of Clean and Efficient Combustion with Alternative Fuels*	Jacqueline Chen	Jacqueline Chen	73.4	828	8.5
m1065	*Rational Catalyst Design for Energy Production*	Andreas Heyden	Andreas Heyden	4.5	0	0
m452	*Condensed Phase Studies with CP2K*	Christopher J. Mundy	Sotiris Xantheas	6.4	27.4	2.7
m1513	*Accurate Scalable Calculations for the Ground and Excited States of Complex Molecular Assemblies*	Sotiris Xantheas	Sotiris Xantheas	5	0.622	0
m1524 m1528	*Molecular dynamics simulation of PNIPAM-coated gold nanoparticles;*	Derrick Mancini Sanket Deshmukh	Sanket Deshmukh	12.9	0	0
m822	*Sampling diffusive dynamics on long timescales, and simulating the coupled dynamics of electrons and nuclei*	Thomas Miller	Thomas Miller	21.4	33.0	0.002
Total Represented by All Case Studies**				**343 M**	**2,320 TB**	**150 TB**
All BES usage at NERSC in 2013 (280 projects)**				**820 M**	**2,362 TB**	**188 TB**
Percent of BES 2013 Allocation Represented by Case Studies**				**37.5 %**	**98%**	**80%**

* These projects were allocated under ASCR, but are included here for purposes of evaluating future needs. The science research falls under BES and will need to be accommodated in BES in the future.
** Includes m1516 and m1792

6 Findings

6.1 Requirements Summary

The following is a summary of requirements derived from the case studies. Note that many requirements are stated individually but are in fact closely related to and dependent upon others.

6.1.1 Scientists will need access to significantly more computational and storage resources to achieve their goals and reach BES research objectives

- Researchers attending the review anticipate that BES scientists will need 15.8 billion Hopper-equivalent hours of computing time in 2017. This is 17 times what BES used in 2013 and 31 times what BES used in 2012.
- BES scientists will need more than 3 PB of shared real-time-access file storage space and 36 PB of archival data storage at NERSC. Both of these values are approximately 16 times what BES used in 2013.
- System stability, reliability, availability, and usability are important features scientists need to make use of their allocations.
- BES facilities (e.g., light sources) are expected to produce a large amount of data that will require storage and computational resources for analysis beyond the requirements given above.

6.1.2 Users will need assistance from NERSC to prepare for Cori (NERSC-8) and follow-on manycore systems

- Scientists need guidance, advice, and training from NERSC to transition codes for efficient computation on manycore systems like Cori.
- BES researchers depend substantially on third-party ISV and community software (full applications and several key libraries and partial differential equation (PDE) solvers) and there is an expectation that this software will be available and run well on future systems.
- Connections with computer science experts are needed to develop new algorithms.

6.1.3 Research teams need to run complex jobs of many different types and scales.

- Researchers need to run codes at both high and low parallel concurrencies, with run times from seconds to weeks. Some codes require large-memory nodes up to 100 GB or more.

- Workflows are becoming more complicated and tools are needed to accommodate this need. Some teams' workflows involve multiple resources across different sites.
- Time to solution is the most important metric for success. This requires schedulers and policies that can support High Throughput Computing, with the ability to accommodate episodic computing needs.

6.1.4 BES is a leader in innovative use of HPC and requires a diverse set of resources and services from NERSC.

- Science teams need to continue delivering data and results of calculations performed at NERSC to their own communities of users.
- The Materials Project informatics approach to science needs supercomputing resources, high-throughput computing capabilities, shared data storage, complex workflows, and web portals.
- Other projects, like those at the Advanced Light Source, have similar requirements.
- Some projects need on-demand computing to serve their users.

6.1.5 BES facilities need computational analysis and data storage resources beyond what they can provide.

- Scientific insight and discovery are now limited as much by computational capacity as by detector or accelerator technology.
- Since users of other facilities are investing considerable time in porting data and software to NERSC, multi-year NERSC account commitments are needed.
- Facilities need to overcome the challenge of data management, including ownership, stewardship, and provenance.
- Advanced data analytics techniques and software are needed to make scientific discoveries.

6.2 Computing and Storage Requirements

The following two tables list, respectively, the 2017 computational hours and storage needed at NERSC for research represented by the case studies in this report. "Total Scaled Requirement" at the end of each table represents the amount (hours or TB) needed by all 2013 BES NERSC projects if 2013 BES usage is increased by the same factor as that needed by the projects represented by the case studies. The "Factor Increase" listed for the project for which Steefel is PI was obtained by using the sum of the three ASCR projects listed above as the reference.

6.2.1 Computing Requirements

Case Study Title	Principal Investigator	Repo(s)	Compute Resources Needed in 2017	
			Million Hours	Factor Increase vs. 2013
Large Scale Geophysical Simulation and Imaging	Newman	m372	900	31
Computational Studies in Molecular Geochemistry	Felmy	mp119	22	10
Direct Numerical Simulation of Poisson-Nernst-Planck Equation in Charged Clays	Steefel	m1516 m1792	1,000	10
Global-Scale Full-Waveform Seismic Imaging of Earth's Mantle	Romanowicz	m554	25	8
Computational Resources for the Nanomaterials Theory Institute at the Center for Nanophase Materials Sciences	Kent	m526	500	26
The Materials Project	Persson	matgen, matcomp	500	26
Excited-State and Charge Transport Phenomena in Novel Energy Material	Neaton	m1793	250	13
Computational Design of Novel Energy Materials	Grossman	m1797	416	27
Advanced Light Source	Banda	als	45	10
Advanced Modeling for Next-Generation BES Accelerators	Ryne	m669	100	19
Combustion of alternative fuels for transportation systems – fundamental investigation using direct numerical simulations	Chen	mp241	500	6.8
Rational Catalyst Design for Energy Production	Heyden	m1065	20	4.8
Condensed Phase Studies with CP2K	Mundy	m452	18	2.8
Accurate Scalable Calculations for the Ground and Excited States of Complex Molecular Assemblies	Xantheas	m1513	500	38
Molecular Dynamics of PNIPAM Agglomerates and Composite Architectures	Deshmukh	m1528, m1524	500	39
Sampling Diffusive Dynamics on Long Timescales, and Simulating the Coupled Dynamics of Electrons and Nuclei	Miller	m822	150	7.0
Total from by Case Studies			**5,946**	
Percent of NERSC 2013 BES Allocations Represented by Case Studies			**37.5 %**	
All BES at NERSC Total Scaled Requirement for 2017			**15,856**	**17.3**

6.2.2 Storage Requirements

Case Study Title	PI (Repo)	Archival Data Storage Needed in 2017		Shared Online Data Storage Needed in 2017	
		TB	Factor Increase	TB	Factor Increase
Large Scale Geophysical Simulation and Imaging	Newman (m372)	10	58	1.0	62.5
Computational Studies in Molecular Geochemistry	Felmy (mp119)	10,000	-	1,000	-
Direct Numerical Simulation of Poisson-Nernst-Planck Equation in Charged Clays	Steefel (m1516) (m1792)	10,000	13	100	50
Global-Scale Full-Waveform Seismic Imaging of Earth's Mantle	Romanowicz (m554)	0.5	6.2	0.1	2
Resources for the Nanomaterials Theory Institute at the Center for Nanophase Materials Sciences	Kent (m526)	600	39	10	2.4
The Materials Project	Persson (matgen) (matcomp)	1,000	22	1,000	20
Excited-State and Charge Transport Phenomena in Novel Energy Material	Neaton (m1793)	500	11	20	-
Computational Design of Novel Energy Materials	Grossman (m1797) (m655)	70	10	70	10
Advanced Light Source	Banda (als)	5,000	9.3	200	2.7
Advanced Modeling for Next-Generation BES Accelerators	Ryne (m669)	300	7.8	4	16
Combustion of alternative fuels for transportation systems – fundamental investigation using direct numerical simulations	Chen (mp241)	8,300	10	100	12
Rational Catalyst Design for Energy Production	Heyden (m1065)	0	-	0	-
Condensed Phase Studies with CP2K	Mundy (m452)	70	2.6	5	2
Accurate Scalable Calculations for the Ground and Excited States of Complex Molecular Assemblies	Xantheas (m1513)	200	300	1	-
Molecular Dynamics of PNIPAM Agglomerates and Composite Architectures	Deshmukh/ Mancini (m1528) (m1524)	20	-	20	-
Sampling Diffusive Dynamics on Long Timescales, and Simulating the Coupled Dynamics of Electrons and nuclei	Miller (m822)	75	2.3	4	-

Total Represented by Case Studies	**36,145**	**15.6**	**2,535**	**16.8**
Percent of NERSC 2013 BES Allocations Represented by Case Studies	**98.2%**		**80.3%**	
All BES at NERSC Total Scaled Requirement for 2017	**36,800**		**3,159**	

6.3 Additional Observations

Participants at the meeting noted a number of observations that are not listed in the high-level findings, the most significant of which are listed here.

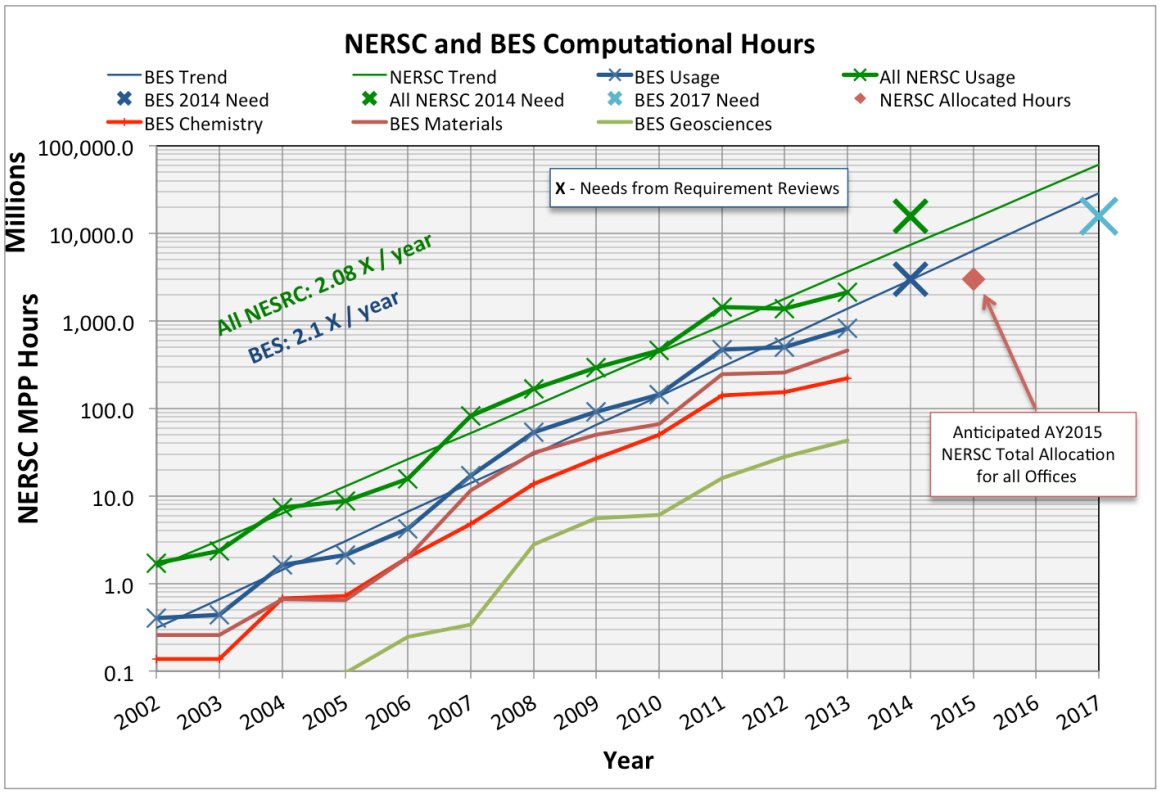
- Many projects could make good use of long queue run limits – at the very least 24 hours – using moderate MPI parallelism.
- Many MD, DFT, and electronic structure codes do not have built-in checkpoint ability.
- Many projects could use large-memory nodes for analysis, especially for Density Functional Theory (DFT) for excited state calculations (some users do not use all the cores on a node in order to get access to more memory per core) and correlated wave functions.
- There are ongoing porting efforts for manycore systems now within the BES community.
- There is a portion of the BES workload that has invested in creating software that uses GPUs and can use GPUs as part of its end-to-end workflow today. Some codes also have OpenMP but many don't yet have fine-grained parallelism.
- Projects like the Materials Project and ALS "touch all of NERSC."
- Users are looking for way to create fault-tolerant applications and workflows; some mechanism for probing the health of a system and its components would be helpful.
- The table below summarizes some characteristics of projects that took part in this review. In this table, "HTC" is High-Throughput Computing and the "Software" column omits more "routine" software, such as MPI, OpenMP, LAPACK, HDF5, even though projects may need it. Visualization products are also not listed here.

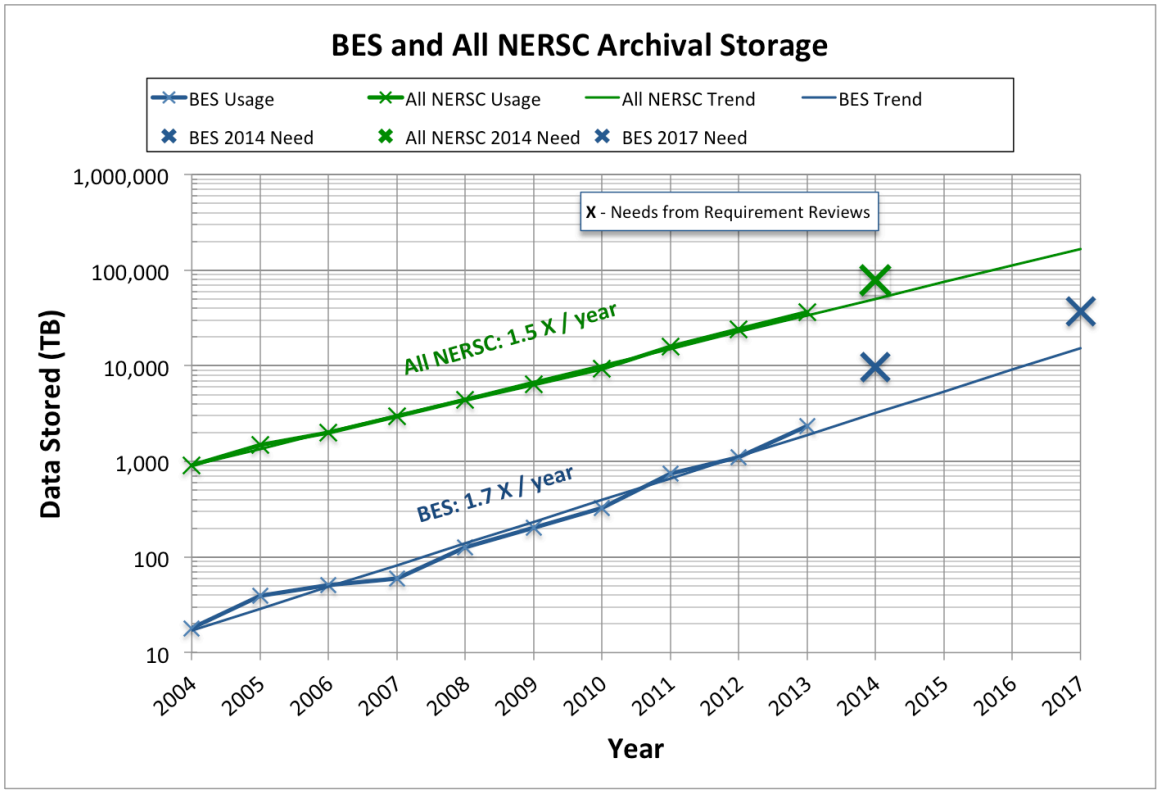
Project Title	Principal Investigator	HTC Important?	Manycore Ready Now?	Software NERSC Needs to Provide	Strong or Weak Scaling Needed
Large Scale Geophysical Simulation and Imaging	Newman	No	Some	MUMPS, PETSc, SuperLU, Trilinos	Strong
Computational Studies in Molecular Geochemistry	Felmy	No	Some	FFT, Global Arrays, ScaLAPACK	Strong
Direct Numerical Simulation of Poisson-Nernst-Planck Equation in Charged Clays	Steefel	No	No	PETSc	Weak
Global-Scale Full-Waveform Seismic Imaging of Earth's	Romanowicz	No	No	FFT, ScaLAPACK	Strong

Mantle					
Computational Resources for the Nanomaterials Theory Institute at the Center for Nanophase Materials Sciences	Kent	Occasional	Some	LAMMPS, VASP, ABINIT, QE, FFT, ScaLAPACK	Both
The Materials Project	Persson	Yes!	No	VASP, BerkeleyGW, ABINIT, Zeo++, and Boltztrap, MongoDB	Both
Excited-State and Charge Transport Phenomena in Novel Energy Material	Neaton	Not now, maybe in the future	No	Siesta, BerkeleyGW, QE, VASP, PARATEC, FFT, ScaLAPACK	Both
Computational Design of Novel Energy Materials	Grossman	Some	No	BerkeleyGW, LAMMPS, VASP, FFT, ScaLAPACK	Strong
Advanced Light Source	Banda	No	Some	FFT	Both, but mostly weak
Advanced Modeling for Next-Generation BES Accelerators	Ryne	No	No	FFT, ExaHDF5	Both
Combustion of alternative fuels for transportation systems – fundamental investigation using direct numerical simulations	Chen	No	Yes	Adios	Weak
Rational Catalyst Design for Energy Production	Heyden	Yes	No	VASP, FFT, ScaLAPACK	Strong
Condensed Phase Studies with CP2K	Mundy	No	Yes	FFT, ScaLAPACK	Both
Accurate Scalable Calculations for the Ground and Excited States of Complex Molecular Assemblies	Xantheas	No	Some	Global Arrays, ScaLAPACK	Both, but mostly strong
Molecular Dynamics of PNIPAM Agglomerates and Composite Architectures	Deshmukh	No	Yes	NAMD, CHARMM, FFT	Both
Sampling Diffusive Dynamics on Long Timescales, and Simulating the Coupled Dynamics of Electrons and Nuclei	Miller	No	Some	NAMD, GROMACS, DL_POLY, AMBER, MOLPRO	Strong

7 BES and NERSC Trends

The following plots show the historical usage of computational hours (adjusted for performance relative to Hopper hours) and archival storage for BES and all of NERSC. The Xs are the anticipated needs from the Requirements Reviews. The solid lines are trend lines fit to the historical usage. The materials science, chemistry, and geoscience portions of the BES usage are also shown on the computational hours plot.




BES and All NERSC Archival Storage
BES Usage
All NERSC Usage
All NERSC Trend
BES Trend
BES 2014 Need
All NERSC 2014 Need
BES 2017 Need
X - Needs from Requirement Reviews
All NERSC: 1.5 X / year
BES: 1.7 X / year
Data Stored (TB)
1,000,000
100,000
10,000
1,000
100
10
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
Year

8 Materials Sciences Case Studies

8.1 Computational Resources for the Nanomaterials Theory Institute at the Center for Nanophase Materials Sciences

Principal Investigator: Paul Kent (Oak Ridge Ngational Laboratory)
NERSC Repository: m526

8.1.1 Project Description

This project is from the Center for Nanophase Materials Sciences (CNMS) at Oak Ridge National Laboratory, which is supported by the BES Scientific User Facilities Division. It is included here with Materials Sciences case studies because of similarity in subject matter and because NERSC allocations for this project are handled by Materials Sciences.

8.1.1.1 Overview and Context

Our project supports a varied set of calculations in support of the externally reviewed user proposals and scientific thrusts of the Nanomaterials Theory Institute (NTI) of the CNMS. We perform predictive calculations into the behavior and properties of nanoscale systems, ranging from new energy efficient nanoscale catalysts to simulations of DNA used for molecular electronics. A general trend is the study of increasingly realistic systems, in most cases incorporating large length scales to properly simulate experimental conditions. Our simulations are primarily first-principles (quantum mechanics) or classical molecular dynamics-based atomistic simulations to determine the structure and properties of each system. The calculations are therefore challenging, requiring extensive use of high performance computing facilities at NERSC. The majority of calculations are too large or complex to perform on midrange computer clusters.

Our largest runs are either performed or supervised by NTI staff using appropriate applications, processor counts, and run configurations, as determined by careful benchmarking. This process enables the most appropriate tools to be selected and optimized for the task at hand. E.g., we have considerable experience in selecting between and optimizing the different numerical implementations of density functional theory that are available. We also make extensive use of classical molecular dynamics where suitable potentials are available and multiscale simulation approaches for polymeric materials. The former tends to utilize popular packages such as LAMMPS, while the latter tend to utilize homegrown codes.

In addition to using conventional modeling based on molecular dynamics and quantum mechanics, we also use methodology for global optimization of structures to help locate ground state geometries (e.g., of organic monolayers on metal substrates for molecular electronics applications) and to identify reaction intermediates and reaction pathways of catalytic processes. Techniques include basin hopping, cluster expansions, and traditional numerical global optimization approaches such as parallel tempering and genetic algorithms. These methods are commonly used in materials-discovery/genome type approaches and are computationally expensive, often consisting of thousands of single total energy calculations. Our main goal is to improve our ability to effectively collaborate with experimentalists, since the same methods can be used where experimental resolution and

physical intuition are initially lacking, e.g., for complex surface reconstructions, and can increase overall confidence in predictions.

Overall storage is not a significant consideration compared to some other fields, with datasets typically in the Gigabyte range. Although these datasets will grow in proportion to simulated system size and timescale, this volume of data remains relatively facile to store or transfer to home institutions.

8.1.1.2 Scientific Objectives for 2017

Our goals include (1) a more thorough exploration and understanding of already identified materials and chemical systems, and (2) an increasingly common discovery and exploration of the properties of not-yet synthesized materials, using simulations of sufficient sophistication that we can have confidence in the predicted results. Despite a number of famous unresolved cases, such as high temperature superconductivity, there are currently broad classes of chemicals and materials for which predictive methods exist that are quite accurate. Instead, the key difficulty is often that these predictive methods are only effective when the exact atomic structure or molecular conformation is known. A key challenge is therefore to identify unknown structures, unknown reactions, and unknown reaction paths. Currently, simulations that can identify, for example, unknown reactions via forms of accelerated molecular dynamics (basin hopping, metadynamics, Monte Carlo etc.) remain too expensive to apply except for the simplest of processes using relatively simple methods for energies (classical force fields instead of quantum-mechanics based). A third objective is to improve the confidence in our predictions, primarily through the use of improved methodologies as they become computationally affordable (e.g., using quantum-based methods to validate classical force field results).

8.1.2 Computational Strategies (now and in 2017)

8.1.2.1 Approach

For simplicity, we concentrate on simulations of atomistic systems, which currently consume over 75% of our allocation.

In our atomistic simulations, the main task is to compute the energy and forces acting on a set of atoms. This is performed using either classical-mechanics based potentials or (more costly) using approximate solutions of the Schrodinger equation to incorporate the effects of quantum mechanics. The trajectory of the atoms can then be integrated using simple Newtonian mechanics. For classical simulations we make use of the very large number of developed force fields implemented in codes such as LAMMPS. For simulations based on quantum mechanics we primarily use density functional theory (DFT), as numerically implemented within several variants of the plane wave pseudopotential approximation in codes such a Quantum Espresso ("pwscf") and VASP. For more accurate quantum mechanical methods we utilize several forms of quantum chemistry and also quantum Monte Carlo. However, the majority of our time is consumed by DFT. Property computation is only occasionally a major consumer of time, for example, accurate band gaps of materials require use of the GW method which scales with a much higher power of system size than DFT. While previously highly specialized, these methods are becoming both commoditized and expected by reviewers in the scientific community. In part this drives our requirements for increased resources: minimum standards are improving.

8.1.2.2 Codes and Algorithms

LAMMPS: a domain decomposed classical molecular dynamics code, with many integration algorithms and numerical force fields.

Quantum Espresso, VASP, ABINIT: plane wave pseudopotential density functional theory. MPI distributed FFTs and dense linear algebra, non-linear optimization.

All these codes are well studied, but should not be considered as monolithic applications implementing a single algorithm. E.g. Both Quantum Espresso and VASP implement hybrid density functionals and density functional perturbation theory, both of which have very different costs and performance characteristics to conventional/traditional ground-state DFT calculations.

8.1.3 HPC Resources Used Today

8.1.3.1 Computational Hours

We used 19.4 million hours in 2013, considerably more than our allocated 8.25 million, thanks to early user time on Edison.

8.1.3.2 Parallelism

Typical production runs are in the hundreds of cores.

The scalability varies with simulated system size and convergence settings (basis sets, k-points, spin). Although we have run DFT calculations on 40K processors with large enough simulated systems, in practice scalability is limited to a few cores per atom (per k-point, per spin). For todays production runs the largest number of cores is typically in the low thousands. We typically use fewer than the maximum afforded by the code/system/algorithm because of improved efficiency and sometimes for better throughput.

Sometimes we have many independent tasks to run, but these can each use a few nodes, so they are submitted as multiple jobs.

Both strong scaling and weak scaling are important for our project since both determine overall time to scientific solution. Sometimes we have a single system to investigate, hence strong scaling (and queue time) are very important.

8.1.3.3 Scratch Data

We typically consume 1 – 10s of Gigabytes.

8.1.3.4 Shared Data

We have a project directory 'm526' that currently has about 4 TB stored in it. Project directories are a data sharing convenience, and very useful when, e.g., several students are working on the same investigation.

8.1.3.5 Archival Data Storage

We had 15 TB of data stored in NERSC's HPSS system in 2013.

8.1.4 HPC Requirements in 2017

8.1.4.1 Computational Hours Needed

Any one of our three goals can readily utilize an order of magnitude more computational hours. For example, to increase accuracy we would prefer to transition to the routine use of hybrid DFTs instead of local DFTs. These are normally more accurate the local DFTs, and it is becoming increasingly expected (by reviewers, amongst others) that these calculations are included in our studies. These methods scale notionally as N^4 instead of N^3 (N= number of atoms), resulting in a cost at least one order of magnitude higher even in small (tens-hundred atom) systems.

Although there are some exceptions, in general our calculations are not routinely large enough to justify an INCITE allocation. We have access to no other large sources, e.g., through NSF.

8.1.4.2 Parallelism

If current methods are used, only thousands of cores will be used in 2017. If more accurate methods are utilized and time is available, hybrid DFT can readily utilize tens of thousands of cores and such methods may scale up to hundreds of thousands.

We occasionally compute in a high throughput mode (many calculations to be performed simultaneously), or utilize a configuration sampling approach. Potentially this can involve dozens-to-hundreds of calculations.

8.1.4.3 I/O

We will probably write tens of Gigabytes per run and we would hope that I/O time is no larger that about 2.5% of the run time.

8.1.4.4 Scratch Data

We will need several terabytes of scratch space to store the computed quantum mechanics wave functions, which increase with system size.

8.1.4.5 Shared Data

The project directory requirement will probably increase to about 10TB, driven primarily by an increase in the number of configurations and longer trajectories.

8.1.4.6 Archival Data Storage

We estimate needing about 10 Terabytes per user in 2017 for NERSC HPSS, the driver again being an increase in the number of configurations and longer trajectories. With about 60 users (2013 value), that translates to 600 TB of archival storage.

8.1.4.7 Memory Required

We need 64 GB per node, particularly if one-sided communications are well supported.

8.1.4.8 Emerging Technologies and Programming Models

Some of the codes we use are ready, e.g., CUDAized or very highly threaded. We know how to do the translation, rewrite etc., but due to a shortage of human resources and the

requirements to publish, we focus only on what we absolutely must change to obtain good performance on installed and upcoming computer systems, meaning that we typically transform only the most critical paths.

We plan to standardize on open community codes as much as possible, to avoid the difficulties of contributing to proprietary codes and to benefit from international contributions. We hope NERSC focuses effort on open codes and actively steers users towards them.

8.1.4.9 Software Applications and Tools

We'll need the same applications and tools as 2013, only sufficiently scaled and updated for any new architectures:

LAMMPS: a domain decomposed classical molecular dynamics code, with many integration algorithms and numerical force fields.

Quantum Espresso, VASP, ABINIT: plane wave pseudopotential density functional theory. MPI distributed FFTs and dense linear algebra, non-linear optimization.

8.1.4.10 HPC Services

We'll need the same as we do in 2013, only updated.

8.1.4.11 Time to Solution and Throughput

For 2017, time to solution, throughput, turnaround, and job scheduling remain a concern.

8.1.4.12 Data Intensive Needs

We don't have any special requirements in this area.

8.1.4.13 Additional Comments

The most important feature of an HPC system is reliability. We look to NERSC to provide reliable FLOP/s combined with good consulting, training, and support.

8.1.4.14 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	19.2 M	500 M
Typical number of cores* used for production runs	200	2,000
Maximum number of cores* that can be used for production runs	10,000	200,000
Data read and written per run	0.05 TB	0.1 TB
Percent of runtime for I/O	2.5	2.5
Scratch File System space	TB	TB
Shared filesystem space	4 TB	10 TB
Archival data	15 TB	600 TB
Memory per node	64 GB	64 GB

* "Conventional" cores

8.2 The Materials Project

Principal Investigator: Kristin Persson (Lawrence Berkeley National Laboratory)
Additional Worksheet Author: Anubhav Jain (Lawrence Berkeley National Laboratory)
NERSC Repositories: matgen, matcomp

8.2.1 Project Description

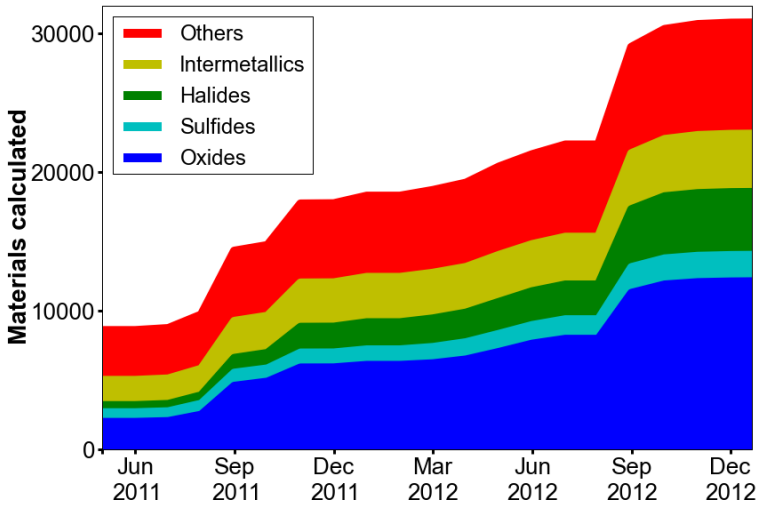
8.2.1.1 Overview and Context

Major technological advancement is largely driven by the discovery of new materials. The performance of materials like solar cells and batteries greatly influence important societal issues like the nature of our future energy supply. However, materials discovery today still involves significant trial-and-error: decades of research are needed to identify a suitable material for a technological application and decades more to prepare it for commercialization.

The goal of the Materials Project (online at https://www.materialsproject.org) is to accelerate materials discovery and education through advanced scientific computing and innovative design methods, scale those computations to cover all known inorganic compounds, and disseminate that information and design tools to the larger materials community. Stated differently, our goal is to *automatically compute* the properties of new materials so that experiments and detailed studies are focused on only the most promising options. This effort requires substantial cross-disciplinary efforts in materials theory, high-throughput computing, experimental and application specific materials knowledge, computer science, data mining, and database science.

Currently, over 4,500 users are registered for the Materials Project. Data are available for 30,000 compounds, and seven different interactive "apps" allow the user to explore the data and perform high-level analysis. External users that used the Materials Project dataset and apps to perform scientific studies have published at least four peer-reviewed papers.

The Materials Project is an example of a large-scale collaboration that leverages many of the resources offered by NERSC. In addition to raw CPU time for performing calculations, we depend on codes compiled by NERSC such as VASP. NERSC hosts and maintains the development and production MongoDB databases (used to drive the web site as well as our workflow) as well as the community-facing web site as a NERSC science gateway. We use the tape storage options for backup, and use the global project directories to store the raw output data. NERSC staff help coordinate web site deployment and releases. Finally, NERSC sets up, operates, and maintains a project-specific cluster ("Mendel") that is used for many of our computations.



Number of compounds available on the Materials Project web site since the initial release in October 2011, broken down by type of compound. Each compound was calculated using the electronic structure code VASP.

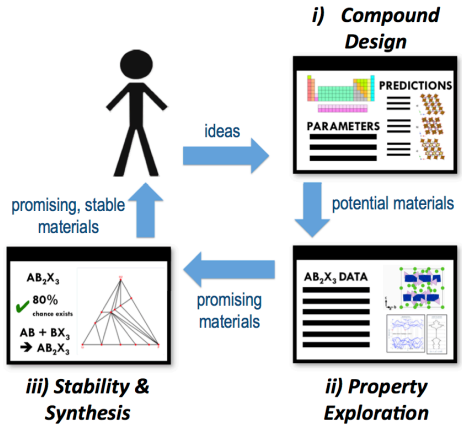
8.2.1.2 Scientific Objectives for 2017

The Materials Project's goal for 2017 is to become an indispensable tool for materials design that combines many achievements. First, we expect to provide researchers with calculated data on *hundreds of thousands* and perhaps *millions* of compounds, making it the largest single resource for materials data. Our computations will not only reveal properties of compounds known to exist, but will also *predict* new compounds and calculate their likelihood for existence using a combination of data mining techniques and ab initio calculations. Each compound will be comprehensively characterized using a suite of computational techniques (some too expensive to apply on a large scale today). These techniques will reveal not only the electronic structure of each material but also mechanical properties, thermal properties, optical properties, and defect character.

Rather than being a static data source, the Materials Project aims to be a full-featured platform for materials design. Researchers will be able to request calculations via the web or mobile platform, and submit them for computation on NERSC resources (a prototype of this functionality already exists). Thus, the space of materials that are computed will start to be "crowd sourced". We will build state-of-the-art APIs to the data in order to enable large-scale data analysis and data mining. Finally, we will allow users to build third-party "apps" that can be integrated with the main web site, and increase outreach and collaboration with experimental teams.

Achieving these objectives requires not only drastic increases in computational power (e.g., 100 times current levels), but also significant improvements in other areas. We expect that 1 petabyte of easily accessible (i.e., not tape) disk space will be needed to store the raw outputs of these calculations. The databases serving the web site and the automation software must similarly be scaled across multiple machines (*sharded*, in MongoDB language). Finally – and perhaps most importantly – queue policies and software must be built that allow ***millions of small, independent, high-walltime jobs to execute on NERSC supercomputers within a 1 year timeframe***. It is important to note that with high-throughput computing at HPC centers, the total number of CPU hours is rarely the limiting factor, even at relatively moderate allocations (e.g., 10 million CPU hours). Rather, hundreds

of jobs must be running continuously in order to make any significant dent in computing budget, which is not possible given (current?) queue submission limits, job turnaround time, and running job limits.



Rapid virtual materials prototyping and discovery envisioned by the Materials Project

8.2.2 Computational Strategies (now and in 2017)

8.2.2.1 Approach

Our current scientific workflow is a series of electronic structure calculations performed over tens of thousands of materials. Each electronic structure calculation typically consumes 100 – 1000 CPU-hours, and uses third-party software (VASP) that does not parallelize well past ~100 cores (we typically use 32-48 cores). The workflow for each material currently encompasses about 4 such electronic structure calculations. However, we expect that in the future each material will encompass up to 20 calculations or more as we expand the set of properties that are computed for each material.

We developed our own general workflow software ("FireWorks", http://pythonhosted.org/FireWorks) targeted largely at running high-throughput scientific workflows at HPC centers. The FireWorks software stores our workflows, launches jobs on the clusters with the proper dependencies, and handles failures/restarts/duplicate checking. The software submits jobs either to the Mendel cluster (owned by Materials Project, but maintained by NERSC) or to the "thruput" queue on Hopper. Currently, each job contains a single electronic structure calculation. However, we have recently built a "bundling" feature into FireWorks that can submit hundreds of concurrent electronic structure calculations within a single queue script (it is not yet used in production).

8.2.2.2 Codes and Algorithms

Our current main "workhorse" code is the Vienna Ab Initio Simulation Program (VASP). Almost all our computing budget for FY2013 used this code.

We expect that in the future, the number of codes used will diversify, perhaps including the Berkeley GW, ABINIT, Zeo++, and Boltztrap software packages.

8.2.3 HPC Resources Used Today

8.2.3.1 Computational Hours

We used 18.8 million CPU hours of computing on DOE/ASCR-allocated resources at NERSC in FY2013. This was actually a significant reduction in utilization from FY2012 due largely to almost 6 months of downtime spent building a new workflow system. We expect this number to go up drastically in FY2014. No significant computing was performed outside of NERSC.

8.2.3.2 Parallelism

We typically use 32 cores (Mendel cluster) or 48 cores (Hopper) for running VASP. In our internal tests, scaling is acceptable to 2 nodes (i.e., 1.7X performance using 2X processors) but continuously degrades as more nodes are added. We use 2 nodes because it provides good performance and sufficient memory to perform our calculations, but we have resisted further parallelism since scaling reduces VASP efficiency. If the VASP code demonstrated better scaling, we would use the additional processors in order to reduce walltime.

We run high-throughput computing, but we are submitting tens of thousands of individual jobs rather than packing calculations into a single job. We may in the future pack together ~100 calculations within a single job. This would utilize 3,200-4,800 cores total.

The two types of scaling important to our project are "high throughput bundling" and "strong scaling." The high-throughput bundling (a form of weak scaling) would allow us to run many small jobs within the constraints of an HPC environment. Strong scaling of the VASP code would allow us to reduce our walltimes and increase overall throughput, and should be combined with job bundling.

8.2.3.3 Scratch Data

About 1TB of scratch space is sufficient for our purposes. We regularly (and in fact, automatically) move the results of our calculations from scratch to permanent storage in projectdirs.

8.2.3.4 Shared Data

We have a projectdir called "matgen". It serves as the repository for all our raw outputs, currently hosting about 50 TB of calculations. One issue we've had is the need for about 75TB or more of projectdir space; we've maxed out our projectdir quota several times and had to delete raw outputs of low importance as well as shut down our entire workflow. Note that our calculations need to be easily accessible (i.e., not on tape) because new calculations depend on output files from older calculations, and adding new features to the database often requires reparsing old runs.

8.2.3.5 Archival Data Storage

We have used about 46 TB of space in HPSS at NERSC. We have backed up some of our old calculated data here, and store some of our very old DB snapshots here as well.

8.2.4 HPC Requirements in 2017

8.2.4.1 Computational Hours Needed

A total factor of about 100X over 2012 usage will be required in FY2017. That translates to about 1 billion current NERSC MPP hours. The number of computing cores needed scales linearly with the number of materials we expect to compute. We expect to compute about 10X the number of materials per year in FY2017 than we do today. Separately, we have built a system whereby individual research groups can contribute workflows for expanding the number of properties computed per material. These new properties tend to be more computationally expensive than basic electronic structure calculations, ranging from doubling the total computation time to adding perhaps a factor of 100X as much computing needed for a single material. We do not expect to run the most expensive calculations over every material in the database, but we do expect another factor of 10X in computing to come from performing higher order methods across the database. Thus, we get a total factor of about 100X our 2012 usage.

8.2.4.2 Parallelism

Parallelism in the future depends on the degree to which the third party codes we employ improve their concurrency. While we expect that these codes (e.g. VASP, ABINIT, etc.) will make significant strides towards increased parallelism (perhaps to 10,000 cores), but this is difficult to predict in advance.

We will strive to use the maximum parallelism afforded by the codes that we use. If the queuing systems in 2017 stay similar to those of today, we will almost certainly be bundling hundreds of jobs within a single queue submission in order to simultaneously use many cores.

We expect that our computational needs will diversify as we include more properties per material. For example, when adding optical properties of materials to our workflow, some of the calculations involve serial codes (in particular, BoltzTrap). Other codes (for obtaining a different set of optical properties) for calculating the GW or Bethe-Salpeter approximation may already scale to 1,000 processors. Thus, we expect to be running a spectrum of codes that range in parallelism from 1 processor to perhaps 10,000 processors.

8.2.4.3 I/O

Each run writes (on average) about 200GB of data (however, we have millions of runs to perform). In general, I/O time and bandwidth is not a large concern for our project.

8.2.4.4 Scratch Data

At this time, we expect that moderate scratch space (e.g., 5TB) will be sufficient in 2017. As mentioned previously, we automatically move completed runs to our shared global project directory immediately after completion through our workflow software. Our scratch footprint is therefore expected to remain light.

8.2.4.5 Shared Data

We expect to need at least 1 Petabyte of shared data in our global project directory in 2017 to store the raw output files. One way to reduce this need would be to devise a system

whereby information could be flexibly, quickly, and automatically retrieved from tape as needed. Such projects have previously been initiated at LBL's Computation Research Division, but to our knowledge have not been integrated at NERSC.

8.2.4.6 Archival Data Storage

We will need to be able to back up our expected 1 Petabyte of raw output files, plus perhaps a few hundred additional terabytes for DB backup.

8.2.4.7 Memory Required

At this time, we expect about 64 GB of memory per node (about 128 GB/node would be useful). The codes we currently use do not expect more memory than this. It is difficult to predict if we will be using theoretical methods that require more memory in 2017.

8.2.4.8 Emerging Technologies and Programming Models

Since we do not develop our own electronic structure codes and use third-party software, our ability to use these technologies depends on third party developers. Currently, the electronic structures codes we use cannot exploit emerging technologies like GPUs.

8.2.4.9 Software Applications and Tools

We will need compiled electronic structure codes (VASP, ABINIT, etc.), Python, MongoDB, and Git support.

8.2.4.10 HPC Services

We expect to continue needing shared project space, web gateway support and hosting, and database hosting.

8.2.4.11 Time to Solution and Throughput

We are generally fairly forgiving in "time to solution" (see exception in 5.13), but as mentioned previously we do have problems with throughput when trying to submit many small jobs and still utilize large allocations.

8.2.4.12 Data Intensive Needs

Our application like many others generate large amounts of data that needs to be accessed regularly. We have previously encountered some resistance (though never rejection) in increasing our project directory quota past 40TB. Either project directories should be allowed to significantly expand storage (a factor of 10 or more) or tape storage should be made as almost as easy to use as project directories (from a programming standpoint). For example, we would like to be able to write a Python program that seamlessly grabs specific data from tape storage as needed as easily as opening a file path.

8.2.4.13 What Else?

Almost every science application – ours included – involve times where "time to solution" is at baseline, and other times when "time to solution" is critical to scientific productivity or an upcoming review or conference. In our case, when initially sketching out and debugging new calculation workflows for materials, "time to solution" is especially important. Waiting 3 days in a queue just to find out that a workflow (which may only be 500 total CPU hours of computation) has a bug is a huge impediment to scientific output. Given that accounting for

nuances in new workflow design often requires many cycles of iteration and debugging, months of productivity can be "lost to the queue". This is not a problem for NERSC's utilization of its CPU cycles, but it serves as a series of very impactful speed bumps on the road to scientific breakthroughs.

Currently (and to our knowledge), NERSC provides two official solutions for quick turnaround: the "debug" queue and the "premium" queue. Neither of these is satisfactory. The "debug" queue is fundamentally flawed because it arbitrarily constrains the 'shape' of the job to 30 minutes wall time and ~500 nodes. One cannot use this to debug codes requiring larger wall times or more nodes. In our specific case, we cannot use this queue for debug purposes because our electronic structure codes will require more than 30 minutes to report failure or success. Similarly, the "premium" queue also restricts the shape of the job to an arbitrary wall time and core count. This queue may also have problems regarding users wanting to quickly use their unused allocations at the expense of the waiting time of other users.

NERSC should implement better strategies for variable "time to solution" needed by users. Clearly, *everyone* cannot be at the top of the queue *all* the time. But, perhaps everyone should be able to be at the top of the queue at *some* time of their choosing (without contacting NERSC), and unrestricted by the shape of the debug or premium queues for their high-priority jobs. For example, Materials Project, in addition to a 10M total allocation, might also be awarded 100,000 CPU hours of "high priority" time. These 100,000 CPU hours could be applied to *any* NERSC queue, but with high priority. It could be used before scientific conferences, or to increase turnaround during iterative workflow development (we would use it for the latter). This may not be a perfect solution, but a better strategy for variability in time to solution for all shapes of workflows should be targeted.

One other concern requires jobs requiring large walltimes. Not all codes can be strongly scaled to reduce walltimes. For example, the VASP code we run requires several days of walltime, and there are many such users at NERSC. NERSC has very little support for long walltime jobs, and almost no fallback strategy when long walltime is needed. An "automatic checkpoint/restart" strategy was once pitched to our team and would be fantastic for long wall time users, but it does not seem to have been released.

8.2.4.14 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	18.8 M	1,000 M
Typical number of cores* used for production runs	32-48	unknown
Maximum number of cores* that can be used for production runs	~100	unknown
Data read and written per run	0.001 TB	0.1 TB
Maximum I/O bandwidth	negligible GB/sec	negligible GB/sec
Percent of runtime for I/O	negligible	negligible
Scratch File System space	1 TB	5 TB
Shared filesystem space	51 TB	1,000 TB
Archival data	46 TB	1,000 TB
Memory per node	64 GB	64-128 GB
Aggregate memory	0.128 TB	unknown TB

* "Conventional" cores

8.2.5 Additional Storage and I/O

The I/O is generally not a big concern for running VASP. A run generally results in a few "large" files of about 200MB and some smaller files of about 20GB or less.

8.3 Transport Phenomena in Novel Energy Materials

Principal Investigator: Jeffrey Neaton (Lawrence Berkeley National Laboratory)
Additional Worksheet Authors: Jack Deslippe, Zhenfei Liu, Sahar Sharifzadeh (Lawrence Berkeley National Laboratory)
NERSC Repository: m1793

8.3.1 Project Description

8.3.1.1 Overview and Context

Our research concerns understanding the physics of charge transport and excited-state phenomena in condensed matter systems of relevance to energy, with the aim of predicting and designing new materials for energy conversion, storage, and carbon capture. Broad materials classes such as oxides, organics, metal-organic frameworks, and interfaces feature prominently. Although structurally distinct, these materials classes share astonishing structural and chemical diversity; highly-localized, sometimes strongly-correlated electronic states; and, in instances, appreciable non-covalent interactions. As such, they simultaneously present significant opportunities for discovery and drive the development of contemporary electronic structure theory.

A major theme of our work is to devise analytical and computational methods that exploit connections between these disparate materials classes to create general approximations and methods, design new materials, and understand novel phenomena. An ultimate aim is the development of new intuition – or "design rules" – connecting emergent properties and function to chemical composition and structure. As such, for many projects, we draw upon and develop contemporary "first-principles" density functional theory (DFT)-based approaches, theoretical methods at the nexus of condensed matter physics, quantum chemistry, and computational materials. Below, we describe three projects that make particular use of HPC.

Organic molecules and assemblies are of considerable interest for next-generation photovoltaics and other energy conversion applications. Their performance and utility hinges on the understanding and control of their spectroscopic properties, such as ionization potentials and electron affinities in gas-phase and solid-state environments, and orbital energy level alignment at interfaces. However, orbital energies and energy differences within common approximations to density functional theory (DFT) (such as the local density approximation, generalized gradient approximations, and hybrid functionals) are known to dramatically underestimate these quantities, and a GW-Bethe-Salpeter Equation (GW-BSE) approach is essential. We rely heavily on DFT codes and the BerkeleyGW program for computation. We used almost 20M hours in 2013 and currently require ~100 of TB of storage per year for computation and archival purposes.

A second research area is focused on studies to understand energy transport in biomimetic and natural photosynthetic systems. The efficient transport of excitation energy in natural photosynthesis is a functionality predicated on a complex and dynamic molecular architecture, the underlying principles of which remain to be elucidated. We study energy transfer between both dimers and periodic arrays of the light absorbing organic molecules. The principal objective of the work is to generate detailed models of the absorption and

transport of incident light energy by chromophore arrays with varying geometries, orientations, and environments.

A third area of research involves detailed computational studies of charge transport phenomena and spectroscopic properties of interfaces of complex dye molecules, such as porphyrins, and conducting surfaces, such as gold, graphite, and graphene. In close collaboration with experiment, we compute conductance, thermopower, and IV characteristics of molecular junctions – individual molecules wired up to metallic electrodes – to understand the relationship between chemical composition, atomic-scale structure, and level alignment and transport properties. Density Functional Theory with local or semi-local functionals are known to yield inaccurate level alignment between Fermi level of the junction and frontier orbital energies of the contacted molecule, often resulting in a significantly overestimated conductance. We use a GW-based method to correct the energy level alignment in the junction that leads to quantitative agreement with and understanding of experimental measurements. Practically, a large number of metal layers are needed for these calculations, and for sufficiently large molecules, in-plane lattice parameters of large lateral dimension are required. This leads to supercells containing 100s of atoms for the most complex junctions, and HPC is essential.

8.3.1.2 Scientific Objectives for 2017

One of our science goals is to achieve improved understanding and control of molecular-scale charge transport phenomena, which are central to the realization of next-generation energy conversion materials. The low efficiency of organic and other nanostructure-based solar cells can be connected to ineffective charge separation at donor-acceptor (and p-n) junctions, and charge collection across interfaces with metallic contacts. For many organic solids and interfaces of interest, e.g., organic semiconductors and donor-acceptor organic interfaces, only ordered model structures have been considered so far. More realistic structural models exhibit significant complexity and can involve hundreds or thousands of atoms.

Donor-acceptor interfaces are a crucial component of working organic photovoltaic devices, though the nature of charge transfer and dissociation at these interfaces is not well understood. By 2017, we would like to compute the spectroscopic properties of more realistic and experimentally realized organic donor-acceptor interfaces, with a particular focus on quantitatively understanding the relationship between interface structure and excited-state properties. The understanding gained in such calculations will allow the design of more efficient photovoltaic materials.

In recent years, “materials genome” type approaches have successfully demonstrated the utility of data mining in material design, particularly in the field of battery development. In order to apply a materials genome type approach to photovoltaic materials, one must be able to quantitatively predict both the ground and excited-state properties of the materials involved. For example, one would want to screen the relative energy level alignment in various donor/acceptor pairs. The GW-BSE methodology has proven extremely accurate at predicting such excited state properties. However, the approach has, until now, typically been applied to one system (typically bulk or periodic systems not containing transition metals) at a time with significant operator involvement. By 2017, we would like to be completing GW-BSE computations across wide datasets of materials and interfaces for the purposes of evaluating materials for energy applications. Further, we plan to achieve deep understanding excited states of complex, light-harvesting molecules adsorbed on a metal or

semiconductor surface, as it is crucial to the conductance, spectroscopy, or catalytic activity. Currently, approximate GW methods developed within our group work well for weak adsorption and negligible charge transfer between the surface and the molecule. We plan to extend the method to intermediate and strong coupling regimes through detailed, more rigorous GW-BSE calculations on these complex hybrid systems.

8.3.2 Computational Strategies (now and in 2017)

8.3.2.1 Approach

Generally speaking the goal is to compute a description of the many-body electronic states of a variety of materials from first-principles. Typically we use the DFT formalism to describe the ground state electronic-structure of a material and then use the GW-Bethe-Salpeter Equation approach (starting from DFT) to describe the excited state properties of a material. These approaches involve the construction and solution of dense eigenvalue problems. However, it is often the construction or application of the operators that consume the most resources. See the next section for more details.

For studies of electron transport, an important goal is to calculate transmission as a function of energy for a number of molecular junctions, often consisting of supercells of 100s of atoms. This is done via a non-equilibrium Green's function or scattering-state framework, and involves construction of the Landauer formula, which consists of multiplication of coupling matrices and Green's function matrices of the junction. This is a non-self-consistent step following the self-consistent convergence of density matrix of the junction. The coupling matrices are computed at DFT level, which involves generating surface Green's functions of the leads. In the approximate GW method we have recently developed, the Green's functions of the extended molecule are computed at DFT level first, and then the poles of the molecular block of the matrix are shifted using non-empirical approximate self-energy, resulting in accurate level alignment between Fermi energy and frontier orbital energies of the molecule. In the non-self-consistent calculation of Landauer formula, a large number of k-point is desired, because the transmission converges slower than density matrices over k-point. Also, a large number of energy points are needed, to generate a smooth transmission curve as a function of energy. Both of these require HPC resources.

8.3.2.2 Codes and Algorithms

DFT Codes (Quantum ESPRESSO / PARATEC / VASP / SIESTA etc.):

These are computer codes for electronic structure calculations and materials modeling at the nanoscale based on density-functional theory, plane waves, and pseudopotentials (both norm-conserving and ultrasoft).

Typically these codes construct and solve the Kohn-Sham equations self-consistently, where each self-consistent iteration involves the solution (or partial solution) of a Hermitian eigenvalue problem via iterative methods like conjugate gradients or Davidson. These approaches utilize the fact that the operation of our operator, H, on an arbitrary vector, scales as O(N) instead of the typical O(N^2).

Typically the bottlenecks include the application of the Hamiltonian matrix to a vector (done via parallel FFTs), and the construction and exact diagonalization of the Hamiltonian in a subspace (involving parallel matrix-multiplication and diagonalization via ScaLAPACK).

GW Codes (BerkeleyGW):

The BerkeleyGW Package is a set of computer codes that calculates the quasiparticle properties and the optical responses of a large variety of materials from bulk periodic crystals to nanostructures such as slabs, wires and molecules. The package takes as input the mean-field results from various electronic structure codes such as the Kohn-Sham DFT eigenvalues and eigenvectors computed with PARATEC, Quantum ESPRESSO, SIESTA, Octopus, or TBPW (aka EPM).

Typically the problem involves the setup and solution of the Dyson's equation – similar to the Kohn-Sham equations in DFT but consisting of an energy-dependent, non-hermitian self-energy operator.

The code is heavily dependent on FFTs (using libraries like FFTW and MKL) and dense linear algebra matrix-multiplication, diagonalization and inversion. The code typically uses threaded libraries and custom MPI/OpenMP parallelization around these libraries.

Transport Codes (Scarlet and TranSiesta)

The TranSiesta utility is part of the Siesta package (although Siesta requires special flags in compilation), and the additional operations on top of Siesta are two-fold: (1) after a regular DFT convergence of density matrix of extended molecule (typically consists of seven layers of metal atoms and the molecule and binding sites in between) using periodic boundary conditions, the code calculates an open-boundary density matrix, by integrating the imaginary part of the lesser Green's function of the extended molecule over energy up to Fermi level. The Fermi level is determined from the regular Siesta DFT calculation and is fixed in this step, and the integral is computed on a contour rather than directly along a real axis. In this way, the coupling of the extended molecule to two semi-infinite leads is taken into account in the calculation of the surface Green's function of the leads, and then these surface Green's functions are used in the calculation of self-energy in Green's function of the extended molecule. (2) After the convergence of the open-boundary density matrix, the Hamiltonian matrices of the two leads and of the extended molecule are stored in files on disk, and the transmission is calculated using Landauer formula in a non-self-consistent step, as described in Section 3.1. The second step is done with the "tbtrans" utility in the package. The code reads in the Hamiltonian matrices, and constructs the Green's function of the extended molecule and the coupling matrices of the extended molecule to the two semi-infinite leads via a surface Green's function of the leads. Finally, the Green's function and the coupling matrices are used to compute transmission matrix using Landauer formula. The transmission at a particular energy is the trace of the transmission matrix at that energy.

It is the second step that requires large number of cores, as a large number of k-points is needed to converge the transmission function. The tbtrans utility parallelizes over k-points, and the number of cores that can be used is up to the same number of k-points. In this way, each core does a single k-point calculation. After an initial loading of the Hamiltonian matrices of the two leads and the extended molecule, the code scales linearly as the number of energy points, as each energy point needs to be calculated separately. Overall the code

scales as O(N^3), where N is the number of basis functions in the extended molecule, due to the operations of matrix multiplication (Landauer step) and inversion (Green's function step).

Alternatively there is another version of the tbtrans utility in the trunk version of TranSiesta. This new utility parallelizes over energy points, and after the initial loading, the code scales linearly as the number of k-points, and each k-point needs to be calculated separately. Overall the code scales O(N^3), where N is the number of basis functions in the extended molecule, just as explained in the previous paragraph. Depending on whether one has more energy points or more k-points, one could choose from the two flavors of tbtrans to achieve greater computational efficiency.

8.3.3 HPC Resources Used Today

8.3.3.1 Computational Hours

In 2013 we used about 19 Million hours at NERSC. We have additional time via NSF and ALCC at ALCF (~10 Million hours)

8.3.3.2 Parallelism

At NERSC today we use anywhere from 100 to 30,000 cores in a run, depending on the program being used (see above). The maximum number of cores we could problem use is about 10,000 for DFT computations and about 100,000 for GW. We generally use fewer than that for faster turn around when running multiple computations/steps at once. We do many medium size systems, where the sweet spot (in terms of efficiency) is less than full scale. We do not currently compute using High Throughput Computing mode, although this is an area we would like to move into with GW computations on large sets of materials in the future.

Both strong scaling and weak scaling are important. We would like to be able compute large sets of medium size materials quickly (and potentially in parallel) as well as study larger, more complex systems such as defects and interfaces etc.

8.3.3.3 Scratch Data

We typically consume about ~10-20TB of temporary disk space.

8.3.3.4 Shared Data

We currently have two project directories, m1694 and mftheory, that are used mostly for sharing files among members.

8.3.3.5 Archival Data Storage

We currently have about 100 TB stored on HPSS.

8.3.4 HPC Requirements in 2017

8.3.4.1 Computational Hours Needed

We estimate that we'll need about 250 million hours from NERSC in 2017. We expect to have a significant allocation from NSF also.

The primary factor driving the need for more hours is that we plan to study more complex and realistic systems and apply GW approaches to large material datasets for genomics-like approaches.

8.3.4.2 Parallelism

We expect to use 50,000-100,000 for studying large systems and about 5,000-25,000 for high-throughput studies of materials databases. The maximum that could be used is greater than 100,000, assuming a hybrid MPI/OpenMP programming model. For our genomics-like approaches we may have as many as ten or so jobs running concurrently and about 10 multiple tasks per job.

8.3.4.3 I/O

We estimate having to write about 1-50TB for intermediate files (roughly equivalent of checkpointing).

An I/O bandwidth of 100s of GB/s in practice would be ideal and less than 20% of run time devoted to I/O would be ideal.

8.3.4.4 Scratch Data

We estimate that ~100-200 TB would be ideal. It would allow us to run multiple large concurrent calculations. The primary cause of this growth is increase in size and complexity of the systems we simulate. Data needs scale as N^2 (where N is number of atoms).

We estimate needing about 20 TB for use in sharing large intermediate files between users.

8.3.4.5 Archival Data Storage

We estimate a need to store about 500 TB on HPSS. This is due to the need to store larger amounts of input and resultant data for large sets of material systems and for storing some intermediate data.

8.3.4.6 Memory Required

BerkeleyGW can utilize a hybrid OpenMP/MPI programming model that alleviates the memory requirements somewhat on nodes with many compute cores. However, the DFT + GW workload is fairly memory intensive ~100 GB per node would be ideal to split among MPI tasks.

8.3.4.7 Emerging Technologies and Programming Models

Quantum ESPRESSO has basic OpenMP support. BerkeleyGW has fairly efficient OpenMP support. Both codes have very experimental GPU support.

8.3.4.8 Software Applications and Tools

Efficient BLAS/LAPACK/ScaLAPACK (or alternatives like ELPA, Elemental)
Efficient FFT
Fortran/C compilers
Efficient Parallel IO via HDF5

8.3.4.9 HPC Services

We would probably need all of the suggested services, including consulting or account support, data analytics and visualization, training, support servers, collaboration tools, web interfaces, federated authentication services, and gateways.

8.3.4.10 Time to Solution and Throughput

We don't have exceptional needs in this area.

8.3.4.11 Data Intensive Needs

We are "mostly" satisfied with HPSS. Our data management plan is that we currently rely on individual researchers using archival storage (HPSS) for their data. Are thinking about data management and which data needs to be archived for long periods as opposed to that which could be recreated fairly easily.

8.3.4.12 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	18.7 M	250 M
Typical number of cores* used for production runs	100-30,000	10,000-200,000
Maximum number of cores* that can be used for production runs	100,000	100,000+
Data read and written per run	0.1-2TB	1-20TB
Maximum I/O bandwidth	1-3GB/sec (in practice)	100GB/sec
Percent of runtime for I/O	20	20
Scratch File System space	10TB	>100TB
Shared filesystem space	0 TB	20TB
Archival data	45 TB	500TB
Memory per node	40GB	~100GB (assuming more cores per node).
Aggregate memory	~10TB	>100TB per 100,000 cores (1GB/core)

* "Conventional" cores

8.4 Computational Design of Novel Energy Materials

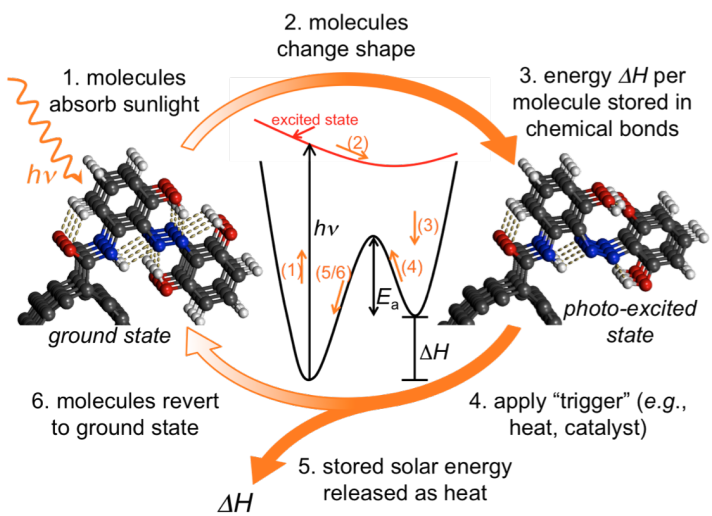
Principal Investigator: Jeffrey C. Grossman (Massachusetts Institute of Technology)
Worksheet Author: Dr. Yun Liu (Massachusetts Institute of Technology)
NERSC Repositories: m1797, Design of high-efficiency solar thermal fuels via first-principles computations; m655, Quantum simulation of nanoscale energy conversion

8.4.1 Project Description

8.4.1.1 Overview and Context

Current technologies that capture, convert, or store energy rely critically on materials innovation. For example, solar thermal fuels (STF) can store the energy of sunlight and release it later in the form of heat, offering an emission-free and renewable solution for both solar energy conversion and storage. However, this approach is currently limited by the lack of low-cost materials with high energy density and high stability in the charged state. During the past few years, computational simulations have demonstrated promising potentials in understanding, discovering, and designing novel materials to tackle such kind of problems. It has been shown in our group that by using quantum mechanical simulations it is possible to design new class of functional materials that have the potential to meet the criteria of STF applications. In this project, we specifically looked into the following research topics:

1. Solar Cells Based on Monolayer Materials
2. Energy alignment in Metal/Organic Interfaces and its effect on solar cell efficiency
3. Optical properties and excited state dynamics of STF
4. High-throughput search of novel STF materials
5. Thermal and electronic transport in pattered graphene for thermal electric applications



Basic operation of a solar thermal fuel, with an azobenzene-derivitized carbon nanotube, one of the proposed HybriSol fuels. H, C, N, and O are white, gray, blue, and red, respectively.

To achieve our goals, we have applied a combination of various computational techniques, including empirical force field molecular dynamics (MD), density functional theory (DFT), time-dependent density functional theory (TDDFT), GW and Bethe-Salpeter calculations. Because of the complexity of the systems we studied and the high accuracy method we employed, our simulations are challenging and the use of HPC resource is thus critical.

8.4.1.2 Scientific Objectives for 2017

As for 2017, we are expecting to extend our current research fields in all aspects. We are planning to look into patterning and functionalization in 2D materials for solar cell and thermal electric applications, photophysics of STF with large templates and functionalized groups. With the increase in system size and method accuracy, those calculations are expected to be computationally very intensive.

8.4.2 Computational Strategies (now and in 2017)

8.4.2.1 Approach

The methods we primarily used are empirical MD, DFT, TDDFT, GW and Bethe-Salpeter calculations. Most of the time the thermodynamic stability of a given system is calculated using DFT simulations. For large systems that are beyond the capability of DFT simulations at the present time, empirical MD can give reasonable result with much less computational effort. GW and Bethe-Salpeter calculations are used to model the adsorption of monolayer material based solar cells. TDDFT is employed to calculate light adsorption and excited state relaxation dynamics for STF.

8.4.2.2 Codes and Algorithms

The LAMMPS code allows us to carry out empirical force field simulations and calculate thermal conductivities of graphene layers. It also enables us to do a fast pre-screening for our high-throughput search of novel STF materials. DFT methods as implemented in the VASP code are extensively used across all of our research topics to calculate ground state energy and structural stabilities. GW with Bethe-Salpeter equation calculations are also carried out using the VASP code. DFT and TDDFT as implemented in real space in the code Octopus are used to calculate adsorption spectra and excited-state dynamics of composite systems involving organic molecules and carbon nanostructures. The Quantum-Espresso code and Gaussian code are used to calculate phonon modes and dynamic stabilities of those systems.

8.4.3 HPC Resources Used Today

8.4.3.1 Computational Hours

We used a total of 15.4M hours at NERSC in 2013. We also used a total of 2.6M SUs on Stampede, an NSF HPC resource. We have three local clusters in our group with a total of 2,000 cores that we use to run smaller calculations.

8.4.3.2 Parallelism

For LAMMPS simulations, we typically use 120 cores at a time, depending on the size of the system. A VASP calculation of relaxation and ground state energies typically runs on 96 cores. GW and Bethe-Salpeter calculation usually takes around 240 cores per job due to the

amount of memory requested. Phonon calculations using Quantum-Espresso and Gaussian typically runs on 96 cores. Octopus typically needs around 300 cores to do TDDFT propagations.

The scaling for VASP calculations for our systems is good up to 300 cores. LAMMPS does scale well with larger systems up to 10,000 of cores. With production runs for the size of our system, we estimate the maximum number of cores we can use will be around 1,000. Octopus parallelizes well up to 10,000 cores.

Usually we are using less than the maximum number of cores we could possibly use to reduce the queuing time of each job and increase the overall production rate. Also the parallelization depends strongly on the size of the system we are simulating and for smaller systems it is more reasonable to use less number of cores. For the high-throughput search of STF material, we utilize the "thruput" queue on Hopper. We are able to run 500 calculations concurrently each time.

Since we have a wide range of research subjects, the scaling needs are broad. For TDDFT simulations, High-throughput calculations, and large-scale empirical MD simulations, parallelizing on many cores is critical to solve the problem within a reasonable time frame.

8.4.3.3 Scratch Data

The maximum amount of scratch data generated during one batch of jobs would be around 1TB. The total amount of scratch space we use is around 10TB.

8.4.3.4 Shared Data

We have a project folder g2e. We had 7 TB stored at the end of 2013.

8.4.3.5 Archival Data Storage

We used 7TB of archival storage in 2013.

8.4.4 HPC Requirements in 2017

8.4.4.1 Computational Hours Needed

To achieve our scientific goals for 2017, we expect to need 800M hours.

To model larger functionalized materials for solar cell and thermal electric applications, the simulation system must be built on large super cells. If we double the size of the simulation cell in 2D planes, the number of electrons to be simulated will be 4 times more than what we currently have. As DFT calculation expenses increase with the cubic-order of number of electrons, 4 times increase in number of electrons will result in 64 times more mathematical operations in DFT simulations.

A single production job with a 1,000 atoms functionalized structure will take 48 hours using 1,000 cores to finish, taking 4-times per-core speed-up after 4 years into account. To investigate 50 structures for solar cell applications, it would need 100*48*500 = 2.4M hours for the DFT calculations. The GW and Bethe-Salpeter calculations will need 1.0M hours for a single large-sized (200 atoms) system. The total estimated resources that we will need are 52.4M hours for investigating 50 structures.

Similar analysis will be true also for thermo-electric application simulations with larger functionalization groups. With a system size of 1,000 atoms, relaxation expect to take 300 steps with 5 hours per step on 1000 cores, which will be 300*5*1000=1.5M hours. Density of States calculation will take 24 hours to run on 4,000 cores, which sums up to 1.5M+24*4000=1.6M per each structure. To calculate 12 samples, it will need a total of 12*1.6=19.2M hours.

To explore the photophysics of solar thermal fuels with large templates such as high index nanotubes, the size of the system will be double the size of the current molecular system. The Casida approach will be scaling with the sixth order of number of electrons due to matrix diagonalization. A calculation with double the number of electrons will take 2^6, or 64 times more computational resource to finish. Each structure will use 100 hours on 64,000 cores to finish the Casida TDDFT calculations. Because of the sixth order scaling of the TDDFT calculations, the DFT part will be negligible comparing to that. Therefore, for each structure it will take 100*64000 =6.4M hours. To calculate 5 such structures will need 6.4*5 = 32M hours.

To sum that up, the total amount of resource required to accomplish the aforementioned goals in 2017 will be 52.4M+19.2M+32M = 103.6 M. Here we have assumed a four-fold increase in single-processor performance, which means our requirement is 416 M hours in units of 2013 NERSC MPP hours.

8.4.4.2 Parallelism

We expect to be running on 1,000 cores for typical DFT calculations and up to 64,000 cores for TDDFT simulations. Octopus has already demonstrated good parallelization over more than 16,000 cores and therefore can be expected to scale well.

8.4.4.3 I/O

We expect to have around 5~10 times increase of I/O due to the increase of number of electrons. Currently the largest I/O bandwidth request comes from Quantum-Espresso calculations of phonon modes, which is around 5 GB/sec per single job. Therefore we expect to have maximum I/O bandwidth request of around 50GB/sec by 2017.

8.4.4.4 Scratch Data

Based on similar reasons as above, we expect 5~10 times increase in scratch files, which will be 50TB by 2017.

8.4.4.5 Shared Data

We estimate a 5~10 times increase in data generation rate at the present level, which will be around 70TB by the year 2017.

8.4.4.6 Archival Data Storage

With similar reasons, we expect to use around 70TB of archival data storage space.

8.4.4.7 Memory Required

With a good parallelization over more cores, we expect the memory requirement will be similar to what we currently have. However, considering that each core will be roughly 4 times faster after 4 years, we will get 4 times more calculations done per core. Therefore the

memory needed per core will be 4 times more than what we need at the present, which will be around 8GB in 2017.

8.4.4.8 Emerging Technologies and Programming Models

We have noticed that the Quantum-Espresso code we use requires a relatively high I/O bandwidth to calculate the phonon modes of a mid-size system (~200 atoms). On many other HPC resources, we cannot run more than one of such calculations at the same time as it is pushing the system I/O bandwidth limit. The high performance scratch file architectures at NERSC enabled us with such calculations.

8.4.4.9 Software Applications and Tools

Besides the above-mentioned codes, we will also need the following software and libraries:

Intel C++ and Fortran compilers, MKL libraries, fatwa libraries, ScaLAPACK libraries, GNUplot, VMD, SIESTA, Java, Python.

8.4.4.10 HPC Services

We would like to get the following consulting services from NERSC by 2017:

Code compilation, database server host (for high-throughput libraries generated), web server host (same as above), cloud sync service.

8.4.4.11 Time to Solution and Throughput

As we have demonstrated above, as long as the parallelization over cores is large enough, there would not be problem with throughput and time to solution.

8.4.4.12 Data Intensive Needs

By 2017 we expect to have a large library of candidate STF molecules generated from our High-throughput simulations. It would be more convenient by then to host the STF candidate material library on the NERSC computers with external accessible database and web interface. In this way we could integrate data-generation, data-mining and new material discovery together at the same place.

8.4.4.13 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	15.4 M	416 M
Typical number of cores* used for production runs	96	1,000
Maximum number of cores* that can be used for production runs	960	64,000
Data read and written per run	2 TB	10 TB
Maximum I/O bandwidth	5 GB/sec	50 GB/sec
Percent of runtime for I/O	0.1%	0.1%
Scratch File System space	10 TB	50 TB
Shared filesystem space	7 TB	70 TB
Archival data	7 TB	70 TB
Memory per core **	2 GB	8 GB
Aggregate memory	1.5 TB	50 TB

* "Conventional" cores

** Changed from per node to per core.

9 Chemical Sciences Case Studies

9.1 Combustion of Alternative Fuels for Transportation Systems – Fundamental Investigation using Direct Numerical Simulations

Principal Investigator: Jacqueline H. Chen (Sandia National Laboratories)
NERSC Repository: mp241

9.1.1 Project Description

Over 70% of the 86 million barrels of crude oil that are consumed in this nation each day are used in internal combustion engines. The nation spends about 1 billion dollars a day on imported oil. Accompanying the tremendous oil consumption is the undesirable emissions – nitric oxides, particulates, and CO_2 production. To mitigate the negative environmental and health implications, there is legislation that mandates reductions in fuel usage per kilometer by 50% in new vehicles by 2030 and greenhouse gases by 80% by 2050. Although these dates may seem to be far off, the time required to bring new vehicle technologies to market and to become widely adopted is lengthy.

Hence, the urgent need for a concerted effort to develop non-petroleum-based fuels and their efficient, clean utilization in transportation is warranted by concerns over energy sustainability, energy security, and global warming. Drastic changes in the fuel constituents and operational characteristics of automobiles and trucks are needed over the next few decades as the world transitions away from petroleum-derived transportation fuels. Conventional empirical approaches to developing new engines and certifying new fuels have only led to incremental improvements, and as such they cannot meet these enormous challenges in a timely, cost-effective manner. Achieving the required high rate of innovation will require computer-aided design, as is currently used to design the aerodynamically efficient wings of airplanes and the molecules in ozone-friendly refrigerants. The diversity of alternative fuels and the corresponding variation in their physical and chemical properties, coupled with simultaneous changes in automotive design/control strategies needed to improve efficiency and reduce emissions, pose immense technical challenges.

A central challenge is predicting combustion rates and emissions in novel low temperature compression ignition engines. Compression ignition engines have much higher efficiencies than spark ignited gasoline engines (only 20% efficiency) with the potential to increase by as much as 50% if key technological challenges can be overcome. Current diesel engines suffer from high nitric oxide and particulate emissions requiring expensive after-treatments. To reduce emissions while capitalizing on the high fuel efficiency of compression ignition engines constrains the thermo-chemical space that advanced engines can operate in, *i.e.* they much burn overall fuel-lean, dilute and at lower temperatures than conventional diesel engines. Combustion in this new environment is governed by previously unexplored regimes of mixed-mode combustion involving strong coupling between turbulent mixing and chemistry characterized by intermittent phenomena such as auto-ignition and extinction in stratified mixtures burning near flammability limits. The

new combustion regimes are poorly understood, and there is a dearth of predictive models for engine design operating in these regimes. Basic research in this area is underscored in the Department of Energy Basic Energy Sciences workshop report [i] on "Basic Energy Needs for Clean and Efficient Combustion of 21st Century Transportation Fuels" which identified a single overarching grand challenge: to develop a *"validated, predictive, multi-scale, combustion modeling capability to optimize the design and operation of evolving fuels in advanced engines for transportation applications."*

This project addresses this challenge through first principles direct numerical simulation of turbulent reactive flows on petascale machines. Fortunately, recent advances in chemical kinetics of alternative fuels for transportation by the DOE Combustion Energy Frontier Research Center, petascale turbulent reactive flow direct numerical simulation software – S3D [Chen+09] and LMC [Day+00] – and high-performance computing suggest that first-principles-based predictive tools for optimum integration of energy conversion/control methodologies and new fuel compositions are possible. In particular we will perform a suite of canonical target problems that are aimed at elucidating mixing, ignition and combustion characteristics of low temperature engines burning alternative fuels using state-of-the-art DNS tools: uniform mesh high-order DNS and block structured low-Mach adaptive mesh refinement. In consultation with key stakeholders in the automotive industry (e.g. Bosch Corporation and Cummins Corporation) and from the DOE Combustion Energy Research Frontier Research Center addressing similar research issues, we plan to perform a set of four direct numerical simulation target problems that together will address the foremost design challenges for fuel efficient, low emissions internal combustion engines burning alternative transportation fuels:

- Ignition characteristics of alternative fuels (oxygenated bio-fuels) and cetane additives in homogeneous charge compression ignition engine environments
- Effect of turbulent mixing and multi-stage autoignition on low-temperature diesel combustion (lifted flame stabilization)
- Combustion instability of lean hydrocarbon fuels (modes and mechanisms of instability and their affect on turbulent burning velocity)
- Influence of thermal, composition and reactivity stratification due to staged injection in Reaction Controlled Compression Ignition (RCCI) on flame dynamics and autoignition.

9.1.2 Overview and Context

The advent of petascale computing applied to direct numerical simulation (DNS) of turbulent combustion has transformed our ability to interrogate fine-grained 'turbulence-chemistry' interactions in canonical and laboratory configurations. In particular, three-dimensional DNS, at moderate Reynolds numbers and with complex chemistry, is providing unprecedented levels of detail to isolate and reveal fundamental causal relationships between turbulence, mixing and reaction. This information is leading to new physical insight, providing benchmark data for assessing model assumptions, suggesting new closure hypotheses, and providing interpretation of statistics obtained from lower-dimensional laser-based optical measurements.

In this research program we have developed and applied a massively parallel three-dimensional **(**DNS) code, S3D, to building-block, laboratory scale flows that reveal fundamental turbulence-chemistry interactions in combustion. The simulation benchmarks are designed to expose and emphasize the role of particular phenomena in turbulent combustion. The simulations address fundamental issues associated with chemistry-turbulence interactions that underlie practical energy conversion devices. Some of the interactions that have been studied in detail include: extinction and re-ignition (Lignell *et al*. 2011 Yang *et al*. 2013), premixed and stratified flame propagation and structure in intense shear driven turbulence (Hawkes *et al.* 2012, Lyra *et al.* 2013), lifted flame stabilization in autoignitive coflowing jet flames (Yoo *et al.* 2011, Yoo *et al.* 2010, Lu *et al.* 2010, and Luo *et al.* 2012) and reactive jets in crossflow (Grout *et al.* 2011, Grout *et al.* 2012, Kolla *et al.* 2012), and flame propagation in boundary layers (Gruber and Chen 2010; Gruber *et al.* 2012).

In addition to the new understanding provided by these simulations, the resultant DNS data are increasingly used to develop and validate predictive mixing and combustion models required in coarse-grained engineering Reynolds-Averaged Navier Stokes (RANS) and large-eddy (LES) simulations. Both *a priori* and *a posteriori* evaluation of key modeling assumptions in RANS and LES have been performed based on benchmark DNS data, and some recent modeling citations to these collaborations are summarized in Table 1.

DNS Benchmark/LES Modeling Issues	**Scalar Flux Modeling**	**Combustion and Mixing Models**	**Scalar dissipation rate and scalar variance modeling**	**Flame Wrinkling Models**
Lifted C_2H_4 Jet Flame in Hot Coflow		Yang *et al* 2013; Knudsen *et al* 2012	Kaul *et al*. 2013; Knudsen *et al* 2012	
H_2/Air Transverse Jet Flame	Kolla *et al* 2012	Lee et al. 2012		
H_2/Air Premixed Jet Flame in Shear Turbulence, **CH_4 Premixed Jet Flame**		Richardson *et al*. 2010; Richardson and Chen, 2012		Hawkes *et al* 2012; Chatakonda *et al* 2012a; Chatakonda *et al.* 2012b
H_2/Air Premixed Flame Boundary Layer Flashback	Raman *et al*. 2013		Raman *et al.* 2013	
H2/Air and C2H4/air Non-premixed Slot Jet Flame Extinction & Re-ignition		LES/PDF Yang *et al* 2013; ODT Punati *et al* 2011; LEM/ISAT Sen *et al* 2010.		

Table 1. DNS-Based Model Development

9.1.2.1 Scientific Objectives for 2017

We propose to perform DNS in laboratory configurations at elevated pressure to enable exploration of chemistry-turbulence interactions – autoignition in stratified mixtures and with mixed mode combustion (*e.g.,* partially premixed flames propagating in autoignitive mixtures) - relevant to fuel efficient, low emissions advanced low temperature engine

concepts burning alternative bio-derived fuels. The two main engine concepts can be broadly categorized as homogeneous charge compression ignition (HCCI) and low-temperature diesel combustion (LTC) engines.

Several classes of target DNS problems will be addressed in 2017:

1) Two canonical DNS configurations for Homogeneous Charge Compression Ignition (HCCI) and Spark-Assisted Compression Ignition (SACI) are proposed. The configurations correspond to a three-dimensional volume of lean fuel-air mixture undergoing isentropic compression and expansion to mimic the mixing and combustion processes in the bulk gases in an engine cylinder during portions of the compression and expansion strokes. We plan to study both HCCI and SACI in three-dimensional DNS in a canonical configuration with isotropic turbulence and temperature stratification with pure ethanol, and ethanol with a cetane improver, EHN. The balance of flame propagation and spontaneous autoignition will be studied along with their influence on ignition timing and pressure rise rate. Moreover, the effectiveness of the cetane additive on modulating the reactivity of alcohol fuels, e.g. ethanol, and on NO generation will be quantified.
2) Direct numerical simulations of lifted dimethyl ether flames will be performed in a three-dimensional configuration at elevated pressure (40 atm) with a diluted dimethyl ether impulsive fuel jet issuing into quiescent air heated to between 900-1200K. We will perform a parametric study by varying the co-flow temperature and diluent. This will allow better understanding of the effect of turbulent mixing processes in an impulsive jet that establish the partially-premixed mixture conditions for multi-stage autoignition, and the role of stable low-temperature auto-ignition intermediate species in stabilizing a lifted diesel jet flame. The DNS will be performed with LMC, an adaptive mesh low-Mach code developed at LBNL.
3) The third target problem is a planar or spherically expanding turbulent premixed flame brush at moderate pressures (greater than 5-10 atm), providing theoretical insight into premixed flame front instabilities under low-temperature, high-pressure conditions typical for transportation environments. At these conditions there exist strong 'turbulence-chemistry' interaction due to the overlap of finite-rate ignition chemistry and local mixing rates, and added complexities associated with thermal/composition stratification that affect the turbulent burning velocity. In both SACI and HCCI with high levels of thermal stratification, understanding and predicting the turbulent premixed burning velocity at high pressure is important. Moreover, the prevalence of intrinsic flame instabilities at high pressure complicates the ability to provide combustion control. At high pressure, experimental differentiation of the aforementioned factors on premixed flame propagation still remains extremely difficult. Carefully designed direct numerical simulations will enable detailed comparisons of flame instabilities and flame-turbulence interaction with experimental data performed in spherically expanding high-pressure flames. We propose to study the interaction of turbulence with two

types of intrinsic premixed flame instabilities: diffusive-thermal pulsating instability and Darrieus-Landau (DL) hydrodynamic instability.

9.1.3 Computational Strategies (now and in 2017)

9.1.3.1 Approach

Direct numerical simulations

Combustion, as it occurs in most practical devices including internal combustion engines, is a highly coupled multi-physics problem spanning a broad range of length and time scales and involving a large number of degrees of freedom. Invariably, the flow is turbulent with a large Reynolds number and the environments are extreme with rapid compression rates and elevated pressures of up to 60 atmospheres involving chemical reactions between hundreds of chemical species. The broad requirements of maximizing fuel efficiency and minimizing emissions under a wide range of operating conditions pose many challenges for the combustion process, requiring novel design concepts and fuels. The design practice in industry typically involves simulating the processes at the largest scales along with a low-order modeling of the coupled physics at the finer scales. This calls on accumulated expertise and can be reliable for conventional fuels, but alternative fuels require fresh consideration and characterization.

Our methodology, first principles-based direct numerical simulations, instead simulates the physics at the finest continuum scales to study the targeted phenomena piecemeal in each canonical configuration. Each of the constitutive physical process: fluid dynamics, thermodynamics, finite rate chemical kinetics and molecular transport are fully represented mathematically at the continuum scale level and the governing equations are solved using high-order accurate numerical methods. While such simulations are computationally expensive, they provide information rich in detail at a level of fidelity sometimes inaccessible even to experiments. The benchmark data, while providing fundamental insight, are invaluable in developing and assessing models of engineering utility and improving their predictive capability. Leadership class computing resources are uniquely positioned to enable simulations of this scale and have a potentially high impact in driving the next generation of advanced combustion technologies using alternative fuels.

9.1.3.2 Codes and Algorithms

The aforementioned simulations will be performed using two research codes; S3D, developed at Sandia National Labs (SNL) and LMC, developed at Lawrence Berkeley National Labs (LBNL). S3D, described in detail in ref. [Chen+09], is a compressible fluid flow solver uniquely tailored for performing gas-phase reacting flow simulations. It solves the conservation equations for mass, momentum, energy and species concentrations in their finite difference form on a fixed mesh rectangular Cartesian domain. An explicit eighth-order accurate central difference scheme with a nine-point stencil is used for spatial derivatives along with a tenth-order filter for damping spurious high-frequency noise [Kennedy+94]. The time integration is performed using an explicit fourth-order accurate low-storage six-stage Runge-Kutta (RK) scheme [Kennedy+00]. Detailed evaluation of thermodynamic quantities, molecular transport coefficients and chemical kinetic reaction rates is incorporated in S3D by interfacing with modified versions of Sandia's CHEMKIN suite of routines. Characteristic boundary condition treatment is applied at the domain

boundaries with special attention being paid to chemical reactions and transverse variation of flow properties on the boundaries [Poinsot+92, Sutherland+03, Yoo+05].

S3D is written primarily in Fortran 90 and fully parallelized using the Message Passing Interface (MPI) programming environment. The computational domain is represented using a fixed regular grid and a static MPI domain decomposition is used whereby each MPI rank maps to a corresponding portion in the rectangular Cartesian domain. This results in all MPI ranks handling the same number of grid points yielding a near-ideal computational load balance. Owing to the explicit numerics the solution loop in S3D involves no all-to-all communication with virtually all communication being between nearest neighbors. Furthermore, the communication is made partially asynchronous with computation to hide some of its cost, resulting in excellent parallel performance. S3D has demonstrated parallel scalability on petascale machines at both OLCF (*jaguarpf*-XT5) and NERSC (*Hopper*-XE6) with near optimal weak scaling up to 150,000 cores on both machines, see Figure 3 below.

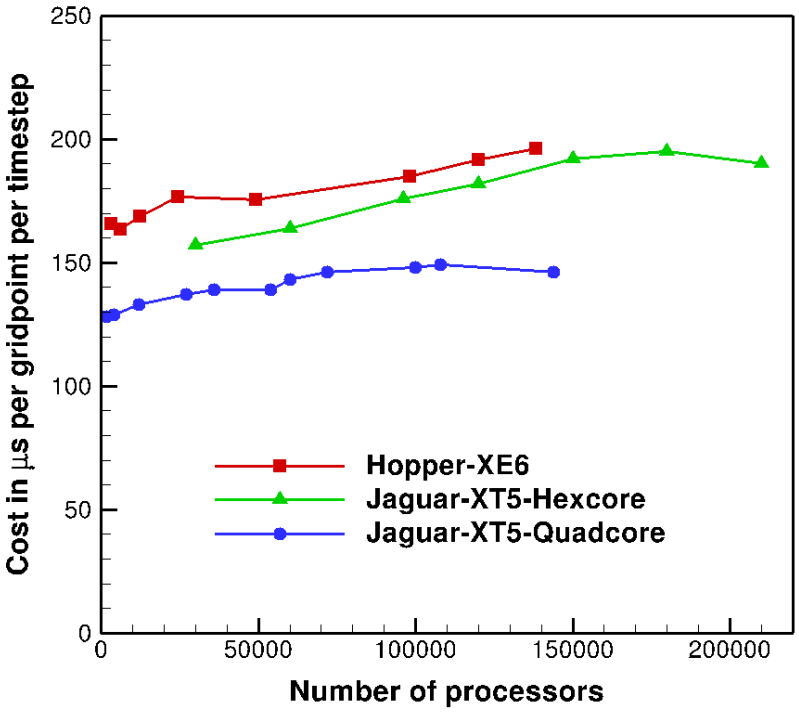


Figure 3. Weak scaling of S3D on Cray XT5 and XE6 machines.

While S3D is attractive from an algorithmic simplicity and parallel scalability standpoint, the explicit time integration dictates that the grid sizes and time steps in S3D simulations be small enough to resolve the smallest length and time scales of the problem. For this reason S3D is ideally suited for problems in which the constitutive physical processes, the fluid dynamics and chemical reactions for example, have time and length scales that are comparable. When there are large disparities in length scales between the flames and turbulence, for example, in some combustion regimes at high pressure, more suitable approaches such as adaptive mesh refinement (AMR) should be considered.

The second computational tool we will use for this project is the low Mach number adaptive mesh refinement code, LMC, developed at LBNL [Day+00]. Similar to S3D, the LMC code integrates the multi-species Navier-Stokes equations with models for detailed chemical kinetics and transport. However, LMC is based on a low Mach number formulation that exploits the natural separation of scales between the fluid velocity and acoustic wave propagation in this low speed problem, removing acoustics and the need to evolve them, from the analytic description of the system entirely. Bulk compressibility effects due to chemical reaction and thermal conduction, remain in the description but appear as a global constraint on the evolution of the velocity field. A predictor-corrector procedure is used to integrate the low Mach number model. The intermediate velocity field that results is then decomposed using a density-weighted projection to extract the component satisfying the global constraint. The projection step involves the solution of a self-adjoint, variable coefficient elliptic equation. For the species conservation equations a splitting method is used that incorporates a stiff ODE integration technique to handle the disparate time scales associated with detailed kinetics. Time evolution of the overall fractional step scheme is constrained by the fluid velocity rather than the acoustic wave speed, increasing the maximum time step by one to two orders of magnitude for many low Mach number combustion applications.

The LMC code also includes block-structured adaptive mesh refinement. In this approach, regions to be refined are organized into rectangular patches, with several hundred to several thousand gridpoints per patch. One is thus able to use rectangular grid methods described above to advance the solution in time; furthermore, the overhead in managing the irregular data structures is amortized over relatively large amounts of floating-point work. Error estimation and refinement criteria are used to dynamically adjust the refinement as the computation proceeds.

The adaptive projection framework uses a hybrid parallelization strategy based on MPI for coarse-grained parallelism and OpenMP for fine-grained parallelism. The code is written in a software framework, BoxLib that handles data distribution and communication for distributing work to computational nodes. OpenMP is used within the physics modules to distribute the work among the different cores within a node. A dynamic load balancing algorithm accommodates the changing workload as regions of refinement are created and destroyed during the computation. For combustion applications the load-balancing problem is complicated by the heterogeneous workloads associated with chemical kinetics. The hybrid implementation has been shown to scale efficiently to more than 50K cores. For problems in which the low Mach number approximation is valid, the combination of adaptive mesh refinement with a low Mach number formulation can result in savings of as much as two orders of magnitude in computational cost compared to a non-adaptive compressible formulation.

9.1.4 HPC Resources Used Today

9.1.4.1 Computational Hours

We used 73.4 million hours at NERSC in 2013, much of it coming from early user time on Edison.

9.1.4.2 Parallelism

The number of cores typically used per simulation is determined by the resolution and size requirements of the investigated target problems. These have been in the range of 20-100K cores.

S3D has demonstrated parallel scalability on petascale machines at both OLCF (jaguarpf-XT5) and NERSC (Hopper-XE6) with near optimal weak scaling up to 150,000 cores on both machines.

Chances of node failure are greater at larger core counts and the wall time of our simulations are long (12/24hours). Therefore, we aim to run at slightly smaller core counts for longer intervals.

In addition we perform periodic I/O collecting the state for statistics accumulation and none of the I/O strategies work well at larger core counts.

We do not use high throughput computing.

Our problems are governed by weak scaling. Our need for capability computing is due to the large number of grid points required to resolve high pressure, high Reynolds number (large dynamic range of flow and flame scales) parameter space of engines and gas turbines with detailed chemistry (10-100 transported species continuity equations per grid). The need for leadership resources is due to the large number of grid points required to resolve high pressure, high Reynolds number (large dynamic range of flow and flame scales) parameter space of engines and gas turbines with detailed chemistry (10-100 transported species continuity equations per grid).

9.1.4.3 Scratch Data

During the simulations restart and checkpoint files are typically stored which are used to monitor the run, to perform data analysis and visualization a posteriori, resulting in temporary disk space requirements of ~300 TB.

9.1.4.4 Shared Data

Our project(s) has a permanent space (mp241) used to store and share data with collaborating members of mp241 or different repositories. The fact that the project directory is not a Lustre file system and it is GPFS limits the IO performance, and hence we don't use it actively for analysis and post processing.

9.1.4.5 Archival Data Storage

The project has 828 TB of data stored in HPSS in 2013.

9.1.5 HPC Requirements in 2017

9.1.5.1 Computational Hours Needed

The computational estimate for the target problems we aim to simulate in CY 2017 is 500 M hours.

We expect that we will probably use OLCF Titan through INCITE and/or ALCC awards.

We propose to perform DNS in canonical configurations to enable exploration of chemistry-turbulence interactions – autoignition in stratified mixtures and with mixed mode combustion (e.g. partially premixed flames propagating in autoignitive mixtures) - relevant to fuel efficient, low emissions advanced low temperature engine concepts burning alternative bio-derived fuels. The need for an increase in capability computing cycles is mainly due to our goal of achieving DNS in relevant aero-thermo-chemical regimes of internal combustion and gas turbine engines. That is, to attain higher Reynolds number (wider dynamic range of inertial to viscous flow scales), higher pressure (30-60 atm), and more complex fuels (bio-fuels with 50-100 transported species) requires a significantly larger number of grid points. Our domain sizes will necessarily increase to accommodate the integral scale of engines and we will have to simulate longer, i.e. larger number of time steps, to capture the relevant ignition phenomena and to attain converged statistics associated with intermittent combustion events. In 2013, our larger grid counts are of the order of 6-7 billion. We expect by 2017 to increase this by an order of magnitude or more. DNS of turbulent combustion is limited by weak scaling performance.

9.1.5.2 Parallelism

We estimate to perform a range of simulations in CY 2017. The number of cores needed is in the range of 100-400K. The maximum we could probably use is about 200K cores in 2013, possibly 400K in 2017 (don't know haven't tested this).

Because we aim to perform multiple simulations for a number of target problems to complete a parametric study, it is typical that we may need to run 2-3 jobs concurrently. We will also be running two executables of S3D: S3D-DNS and S3D-LES lockstep, and hence, required several executables to share and pass information through memory at every sub-stage of our explicit RK method. We currently use MPI-communicator to facilitate this communication, and may explore other staging methods like ADIOS for this purpose. This works on Titan; however, our experience shows machines like Edison cannot support multiple executables on the same node. I can also anticipate additional concurrent executables if we include topological region segmentation running concurrently with the solver.

We do not use high throughput computing.

9.1.5.3 I/O

The estimated I/O size per run is 50-300 TB and the total size of data from the planned simulations may include about 400-800 TB of raw data. The effective I/O bandwidth is 15 GB/s, which corresponds to 5% of the total run time.

9.1.5.4 Scratch Data

The target problems planned for 2017 will have temporary disk space requirements 50-150 TB. We aim to perform a suite of canonical target problems that are aimed at elucidating mixing, ignition and combustion characteristics of low temperature engines burning alternative fuels at high pressures. These calculations impose stringent resolution requirements and involve larger chemical models that require the transport of more species than our currently performed runs, thus increasing our scratch space requirements.

9.1.5.5 Shared Data

As more computational time becomes available we will be able to tackle larger problems both in terms of Reynolds numbers and chemistry complexity. As the size of problems increases the range of scales of scales increases and this is reflected also in our data footprint at all stages of the simulation, (scratch, project, HPSS)

We typically need to share DNS data that is generated with other users in repo mp241 for analysis and visualization purposes. This may include about 100 terabytes of raw data. There will be several 2D and 3D production runs and parametric studies that our collaborators will need access to for post processing.

9.1.5.6 Archival Data Storage

The estimate for the DNS data generated is about 800 terabytes of raw data essential for analyses and visualization. Our 2013 usage was 828 TB and we estimate needing on the order of 10X, or 8.3 PB in 2017.

We store the minimum information so that we can recover the state and collect statistics but as the targeted problem sizes grow so that in every step of the simulations the storage requirements increase. We are also investigating data reduction and compression strategies to reduce the amount of data we store.

9.1.5.7 Memory Required

S3D problem sizes are not restricted by the available memory per node but rather by the processing speed and by the width of the stencils used for the derivative computations. We are memory bandwidth bound typically in our computations.

9.1.5.8 Emerging Technologies and Programming Models

S3D is being optimized for heterogeneous architectures including CPU/GPU and MIC systems. We have completed one refactorization of the code already in order to be able to run on hybrid CPU/GPU architectures (Titan) using OpenACC directives. We are actively seeking alternative strategies with Intel and Cray. We also have an ongoing effort with ExaCT codesign to explore the Legion dynamic runtime and Domain Specific Languages (DSLs) for taking advantage of the increasingly parallel and heterogeneous architectures available for high performance computing. Legion/S3D is a domain specific language that allows application developers to express computations in a high-level language while the details of extracting parallelism and mapping the computation onto a target architecture, such as a GPU, are left to the DSL compiler and dynamic runtime. Legion/S3D is able to identify fine-grain task and data parallelism through task dependency graphs, autotuning, and dynamic scheduling of these tasks. We hope to have comparisons of this new code with the OpenACC S3D by end of 2013.

Hybrid implementations are becoming more critical as processors continue to increase in number of cores and/or threads. It is especially important to have an optimized thread parallel implementation that will continue to scale as threads/cores increase in future architectures. SNL and LBNL have an ongoing effort with Intel Corporation to optimize the thread parallelism and vectorization of their combustion codes. While LMC makes efficient use of the MPI/OpenMP* programming environment, the current production code base of S3D uses only MPI. A hybrid parallelization of S3D utilizing both MPI and OpenMP* is

currently in progress. From September 2012 to December 2012 engineers from Intel, Sandia, and NREL (National Renewable Energy Laboratory) optimized S3D to gain a 1.5x speedup on a platform based on Intel® Xeon® processor E5-2600 product family by employing optimizations targeted at vectorization (intra-register parallelization) and thread parallelism. Sandia, Intel, and NREL are continuing efforts in 2013 and 2014 to find ways to further improve concurrency to enable even more thread parallel optimizations on Intel MIC architectures. These efforts are particularly directed towards making optimal use of Edison and future Intel machines. Any assistance from NERSC towards this goal would be beneficial. The current team is led by Antonio Valles of Intel, Weiqun Zhang from LBNL, H. Kolla and J. Chen from Sandia and R. Grout from NREL.

9.1.5.9 Software Applications and Tools

We need MATLAB, ParaView, VisIt, VISUS from Pascucci, ADIOS, HDF5.

9.1.5.10 HPC Services

We typically need account support and consulting support to help compile our code. Training and support services, and possibly assistance in setting up a computational combustion gateway/portal for community access to simulation data and software tools archived at NERSC.

9.1.5.11 Time to Solution and Throughput

Data storage at every step: scratch, project, HPSS storage and I/O bandwidth and user and project quotas. We will require large compute allocations and a queuing system that is favorable towards large runs requiring long times.

9.1.5.12 Data Intensive Needs

As the problem sizes increase, data transfer between scratch and HPSS and project and scratch will be a bottleneck and it will be an important part of all steps of the workflow to minimize any potential bottlenecks.

We are satisfied with NERSC's HPSS system. The separate queue for transfer of data to the archival storage system is a very good feature of NERS**C.**

We do not have a data management plan.

9.1.5.13 What Else?

Our experience has been that queue policies at NERSC currently seem to favor very small users by allowing flooding of the queues with a large number (~50) small jobs concurrently which prevents large jobs from running. A 20K core job on Edison recently has been in queue for 2-3 weeks before it starts running.

The HPC features that are most important to us include reliability, availability, job turn around, performance.

9.1.5.14 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	73.4 M	500 M
Typical number of cores* used for production runs	20-100K	100-400K
Maximum number of cores* that can be used for production runs	150 K	400K
Data read and written per run	50-100TB	50-200TB
Maximum I/O bandwidth	15GB/sec	15 GB/sec
Percent of runtime for I/O	5%	5%
Scratch File System space	100TB	300 TB
Shared filesystem space	8.5 TB	100 TB
Archival data	828 TB	8,300 TB

* "Conventional cores"

9.2 Rational Catalyst Design for Energy Production

Principal Investigator: Andreas Heyden, University of South Carolina
Worksheet Author: Muhammad Faheem, University of South Carolina
NERSC Repository: m1065

9.2.1 Project Description

9.2.1.1 Overview and Context

Our research primarily focuses on developing a molecular understanding of heterogeneous catalysis at solid-liquid interfaces to elucidate the specific role of a liquid environment on the activity and selectivity of transition-metal catalysts. Despite significant advances in computer algorithms and increasing availability of computational resources, molecular simulations of such large and complex systems remain challenging. Our general approach is to develop multi-scale, mixed-resolution modeling techniques that combine the accuracy of *ab initio* quantum mechanical (QM) methods with the efficiency of classical molecular dynamics (MD) and continuum models to provide a reliable energetic description of the complex system at a fraction of the cost. We have previously demonstrated the effectiveness of an implicit solvation scheme (iSMS) based on integration of planewave density functional theory (DFT) calculations with continuum solvation models for rapid computation of reaction free energies of processes occurring at solid-liquid interfaces. Recently, we have developed and validated an explicit solvation scheme (eSMS) based on integration of planewave DFT calculations with classical MD simulations through free energy perturbation (FEP) methods. Our plan is to apply these novel computational techniques to rationalize the design of heterogeneous catalysts with superior activity, selectivity, and stability for various biomass conversion processes to fuels and value-added chemicals.

Our experience has shown that the computational cost of simulating a reaction with iSMS is 2-3 times that of standard planewave DFT calculations, whereas using eSMS is about 2 orders of magnitude more expensive. Similar trends have been observed regarding data generated and stored for application of these techniques. With the current generation of computational resources, iSMS can be routinely applied for solvent screening and to study complete reaction mechanisms. Application of eSMS, on the other hand, is currently affordable only for a small subset of the overall reaction network.

9.2.1.2 Scientific Objectives for 2017

We envision a 4-step procedure for simulating heterogeneously catalyzed reactions in liquid environments where (1) standard planewave DFT calculations are performed for the entire reaction mechanism, (2) iSMS serves as a pre-screening tool for solvents, process conditions, and elementary steps that are most sensitive to the presence of a liquid environment, (3) eSMS is used for refinement of reaction free energies and free energy barriers for previously identified elementary steps and for generation of meaningful solvation structures, and (4) *ab initio* QM calculations for the solvated model are performed for rate-controlling steps (i.e., our procedure is similar to step 3; however, some water molecules are now treated at the QM level). Our rationale for this strategy is to successively improve the description of the effect of a complex liquid environment on reaction equilibrium and kinetics in conjunction with microkinetic modeling to balance the use of expensive computational techniques with affordability. With usefulness of iSMS and eSMS demonstrated for simulating reactions at solid-liquid interfaces, and with computational

power expected to increase by an order of magnitude in the next 5 years, both techniques should get a wider user base and help create a scientific basis for the rational catalyst design of liquid phase processes.

9.2.2 Computational Strategies (now and in 2017)

9.2.2.1 Approach

Computational tasks associated with our project can be divided into 3 main categories:

1. DFT calculations with planewave and Gaussian-type orbital (GTO) basis: These calculations currently account for almost 100% of our usage of NERSC resources and are performed with commercially available DFT codes (TURBOMOLE, VASP).
2. MD simulations: These calculations are performed using the DLPOLY program and currently account for less than 1% of our usage of NERSC resources. This is mainly because our focus up to this point has been modification and integration of source codes of various programs to better communicate with each other for our QM/MM methodologies. With these developments now complete and rigorously tested, MD simulations may account for about 10% of our usage of NERSC resources.
3. High-throughput calculations: Our methodologies, especially eSMS, require a large number of relatively small but completely independent tasks to be performed, and thus are excellent candidates for bundling and high-throughput computing. We have developed FORTRAN programs and BASH scripts for interfacing of various DFT and MD codes, automatic input generation and output processing. These codes have been extensively tested and are currently in use for production runs at Stampede (an XSEDE resource). Recently, we have started moving some of these calculations to NERSC resources.

9.2.2.2 Codes and Algorithms

VASP: Standard planewave DFT code (≈70% of our usage on NERSC). Most relevant algorithms include Fast Fourier Transform and matrix diagonalization. We have performed minor modifications in the source code to enable better communication with our QM/MM methodologies.

TURBOMOLE: Commercially available GTO-based quantum chemistry code (≈30% of our usage on NERSC). Most relevant algorithm for our project is the Periodic Electrostatic Embedded Cluster method (PEECM). Scripts included with TURBOMOLE have been modified to serve as drivers for geometry optimization and free energy perturbation in our QM/MM methodologies.

DLPOLY: Classical MD code (<1% of our usage on NERSC). This share is expected to grow considerably with more frequent application of the eSMS technique. The source code has been locally modified to include new force field functional forms and to interface with the driver scripts from TURBOMOLE.

9.2.3 HPC Resources Used Today

9.2.3.1 Computational Hours

Period	Resource	Core-Hours used (million)
January 2013 – December 2013	NERSC (Carver/Hopper)	3.5 (allocation)
January 2013 – December 2013	NERSC (Edison)	≈1.0 (estimated)
October 2012 – September 2013	XSEDE (various resources)	3.0
October 2012 – September 2013	PNNL (Chinook)	3.0 (estimated)
January 2013 – December 2013	USC (local resources)	>1.0 (estimated, no formal accounting)

9.2.3.2 Parallelism

VASP	16-32 (Carver); 24-48 (Hopper/Edison)
TURBOMOLE	8-32 (Carver); 24-48 (Hopper); 24 (Edison)
DLPOLY	8-32 (Carver). We note that the use of DLPOLY in eSMS constitutes an embarrassingly parallel problem, and we prefer to run an ensemble of tasks with small core counts.
VASP	64 (Carver); 120 (Hopper)
TURBOMOLE	64 (Carver); 96 (Hopper)
DLPOLY	32 (Carver)

Note: We have listed the maximum number of cores that our group has used. The code(s) may be utilizable on even larger core counts.

Our choice of the typical number of cores for both TURBOMOLE and VASP is based on scaling tests for a typical model system in our project and intended to strike a balance between performance gain, time-to-solution, and waiting time in queue. Currently DLPOLY accounts for a very small fraction of our usage, and is essentially always run in conjunction with TURBOMOLE: we simply distribute an ensemble of DLPOLY tasks on the same cores as used by TURBOMOLE.

We have extensively used such job ensembles on Stampede (an XSEDE resource), typically with 8-16 tasks per job (maximum used=50). We are currently in the process of moving similar calculations to NERSC resources, and expect to use a similar setup.

In general, strong scaling is preferable for our project. The bottleneck in our QM/MM methodology comes from the QM calculations that do not scale well on large core counts. MM calculations in our project can be split in as many independent parallel streams as practically possible. It would be highly desirable to reduce the time-to-solution for QM calculations using large core-counts and further increase the number of parallel MM threads.

9.2.3.3 Scratch Data

We typically use less than 100 GB for temporary disk space.

9.2.3.4 Shared Data

We do not have a project directory at NERSC.

9.2.3.5 Archival Data Storage

Not Applicable

9.2.4 HPC Requirements in 2017

9.2.4.1 Computational Hours Needed

We will need more than 20 million hours at NERSC in 2017. We expect to receive yearly allocations on various XSEDE resources. The primary factor driving the need for more hours is the need to do a large number of computations on more complex systems. Within our fields of study (catalysis and materials/surface science), complexity and accuracy of DFT are huge issues that dramatically limit our ability to understand and design new materials/catalysts for, e.g., future bio-refineries. At the same time, the MGI initiative demands of us that we start designing more materials on a computer. While new/improved methods and codes will be essential, our area of science still needs for the foreseeable future the help of the computing revolution that doubles CPU power per year.

9.2.4.2 Parallelism

We do not expect a significant change on a per-task basis, either in terms of the maximum that could be used or the concurrency we will actually use in our runs. We typically have about 20 jobs running concurrently for each scientist in the group.

We have experience using up to 50 tasks per job. Considering the specific requirements and limitations of our method, this number may be as large as a few hundred.

9.2.4.3 I/O

QM calculations	≈1 GB per job
MM calculations	≈10-50 GB per job
Hybrid QM/MM calculations	≈1-2 TB per job

Note: The numbers listed here are from our recent jobs on Stampede. We do not expect them to change significantly in future.

QM calculations	≈1 GB total data, only written at the end.
MM calculations	≈100 MB/minute/job write. No read.
Hybrid QM/MM calculations	≈2 GB/minute/job (total for all simultaneous tasks).

	Currently read/write ratio is ≈2-3. In planned modifications to our code, we envision it to be almost exclusively read.

We would need I/O to be less than about 10% (for hybrid QM/MM jobs) to be successful.

9.2.4.4 Scratch Data

In 2017 we will need 50-100 TB of "scratch" disk space. The growth relative to today is again due to the need to do a larger number of computations.

9.2.4.5 Shared Data

Not Applicable

9.2.4.6 Archival Data Storage

Not Applicable

9.2.4.7 Memory Required

Per-node memory available on current generation of NERSC systems is sufficient for our jobs for the foreseeable future.

9.2.4.8 Emerging Technologies and Programming Models

Not Applicable

We are not working on modifying any code for accelerator-based computing. However, if GPU versions are available (e.g., there have been multiple efforts for VASP), we can quickly switch.

9.2.4.9 Software Applications and Tools

Libraries: MKL (or equivalent); FFTW

COMPILERS: FORTRAN; C

9.2.4.10 HPC Services

General problem resolution.

9.2.4.11 Time to Solution and Throughput

Availability of longer queues: geometry optimizations in our project are very expensive, and we need multiple submissions to a 48-hour queue to achieve convergence. The ability to run more simultaneous jobs than currently allowed in the longer queues would be highly beneficial for our project.

9.2.4.12 Data Intensive Needs

No special needs.

All large files produced in our project are required only for temporary storage, and if needed again, most of them can be generated quickly. With purge policies of 6-8 weeks on SCRATCH, moving such files to archival storage is not required. Our data management plan is based on local storage of all input but only important output files. We periodically backup our local storage system (2 copies). In addition, after completion of each subtask of the project, we create a DVD of all relevant data.

9.2.4.13 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	4.3 M	20 M
Typical number of cores* used for production runs	24	24
Maximum number of cores* that can be used for production runs	120	120
Data read and written per run	Max 2 TB	Max 10 TB
Maximum I/O bandwidth	2 GB/min/job	2 GB/min/job
Percent of runtime for I/O	10%	10%
Scratch File System space	2 TB	100 TB
Shared filesystem space	N/A	N/A
Archival data	N/A	N/A
Memory per node	64 GB	64 GB
Aggregate memory	TB	TB

* "Conventional" cores

9.2.5 Additional Storage and I/O Remarks

During MM simulations, we generate a single trajectory file (≈10-50 GB) per job. Before starting hybrid QM/MM calculations, we split this file into multiple parts (by re-reading/writing the entire file). Each of these parts then serves as input to an independent stream of eSMS calculations (typically using 2-4 cores per stream).

For example, considering one node of Hopper, we might split the original trajectory file (say 48 GB) into 12 parts (4 GB each), and then use 12 parallel threads (2 cores each). Within each thread, only the first core handles file reading/writing (serial I/O). However, there are 12 such file-handling cores per node (essentially distributed I/O). The rationale for splitting the original file is that these calculations must be repeated at each optimization step (≈15-

20 steps are usually achieved in a 48-hour queue; hence a total of 1-2 TB of data read/written per job). Within each optimization step, most of the time is spent on QM calculations with ≈1 GB data written only at the end of each step. A profile of I/O bandwidth usage of such a job would show almost zero disc activity for the most part, with periodic peaks of large disc activity (each spanning a few minutes).

9.3 Condensed Phase Studies with CP2K

Principal Investigator: Christopher J. Mundy (Pacific Northwest National Laboratory)
Case Study Co-Author: Sotiris Xantheas (Pacific Northwest National Laboratory)
NERSC Repository: m452

9.3.1 Project Description

9.3.1.1 Overview and Context

Our research centers on the use of quantum density functional theory (DFT) and understanding aqueous systems in the condensed phase. DFT, in principle, can explain the detailed molecular structure of solutes in complex heterogeneous environments. We use this information in conjunction with experiment and theory to elucidate mechanisms of ion transport to interfaces. Our studies include, but are not limited to, complex chemical reactions at interfaces, and bulk and interfacial solvation of simple and complex ions. To this end, we have performed some of the largest simulations to date explaining the complex solvation of simple and complex ions in the vicinity of the air-water interface. These studies have pointed to the shortcomings of empirical based models and point to the importance of how a precise determining of local structure about complex solutes can lead to emergent phenomena, such as self-assembly, at larger scales.

9.3.1.2 Scientific Objectives for 2017

At present, we are limited to the simulation of defects in the dilute limit. Our goal for the future will be to study concentrated electrolytes in both bulk and interfacial geometries at a suitable scale to examine the potential of mean force between objects immersed as a function of ionic strength, pH, and electrolyte composition.

9.3.2 Computational Strategies (now and in 2017)

9.3.2.1 Approach

We use electronic structure based potentials of interaction in conjunction with statistical mechanical methods to understand the free energetics of interactions between solutes or complex objects (comprised as an assemble of solutes) as a function of the composition of the solvent.

9.3.2.2 Codes and Algorithms

We use CP2K (www.cp2k.sourceforge.net). CP2k performs atomistic and molecular simulations of solid state, liquid, molecular and biological systems. It provides a general framework for different methods such as, e.g., density functional theory (DFT) using a mixed Gaussian and plane waves approach (GPW), and classical pair and many-body potentials.

We use computational algorithms based in statistical mechanics (e.g., umbrella sampling).

9.3.3 HPC Resources Used Today

9.3.3.1 Computational Hours

We used 6.4 million hours at NERSC in 2013.

9.3.3.2 Parallelism

We typically use between 1,000-2,000 cores per simulation. The maximum is probably around 64,000. Our production runs are application dependent. Typically runs of 1,000-10,000 cores are sufficient for the science we are performing.

Both strong scaling and weak scaling are important to us. We are currently doing the biggest aqueous systems (350 water molecules) for the longest simulation times. What system size we study is tightly coupled to both the level of theory that we utilize and the kind of statistical sampling.

9.3.3.3 Scratch Data

Current scratch space is sufficient for our runs.

9.3.3.4 Shared Data

We use the project directory m452, which currently has about 2.7 TB stored in it. We use this to share data between our collaborators and co-workers.

9.3.3.5 Archival Data Storage

We back up regularly and have 27 TB stored in 2013.

9.3.4 HPC Requirements in 2017

9.3.4.1 Computational Hours Needed

We anticipate we will need about 18 million CPU hours. The reason is that we have more sophisticated quantum mechanical algorithms (e.g. better accuracy) in conjunction with statistical sampling.

9.3.4.2 Parallelism

We anticipate using about 10,000 cores per run; maximum, 100,000 cores. Up to 20 jobs utilizing 10,000 cores might be run concurrently for the statistical mechanical sampling such as umbrella sampling.

9.3.4.3 I/O

CP2K is not an I/O limited code.

9.3.4.4 Scratch Data

I do not anticipate needing a significant upgrade of scratch space.

9.3.4.5 Shared Data

We estimate we will need 5 TB of shared storage space in 2017.

9.3.4.6 Archival Data Storage

We estimate needing to store 10 terabytes/year, which on top of the 27 TB we have stored in 2013, would give us about 70 TB in 2017.

9.3.4.7 Memory Required

16 GB/node is required.

9.3.4.8 Emerging Technologies and Programming Models

CP2K is GPU ready and we have worked directly with vendors (e.g., NVidia) to optimize our code.

9.3.4.9 Software Applications and Tools

We rely on gfortran to build our codes.

9.3.4.10 HPC Services

We do not anticipate a need for special services. Current services seem excellent.

9.3.4.11 Time to Solution and Throughput

No special needs here.

9.3.4.12 Data Intensive Needs

We are very satisfied with NERSC HPSS.

9.3.4.13 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	6.4 M	18 M
Typical number of cores* used for production runs	2,000	10,000
Maximum number of cores* that can be used for production runs	64,000	100,000
Archival storage	27 TB	70 TB
Shared project data space	2.7 TB	5 TB

* "Conventional" cores

9.3.4.14 Additional Storage and I/O Remarks

CP2K applications are not I/O limited.

9.4 Accurate Scalable Calculations for the Ground and Excited States of Complex Molecular Assemblies

Principal Investigator: Sotiris S. Xantheas (Pacific Northwest National Laboratory)
Worksheet Authors: Additional input has been obtained from Drs. Edoardo Apra and Karol Kowalski (EMSL, PNNL)
NERSC Repository: m1513

9.4.1 Project Description

9.4.1.1 Overview and Context

The objective of this research effort aims toward developing a comprehensive, molecular-level understanding of the collective phenomena associated with intermolecular interactions occurring in guest/host molecular systems and aqueous environments. The motivation of the present work stems from the desire to establish the key elements that describe the structural and associated spectral features of simple ions in a variety of hydrogen bonded environments such as bulk water, aqueous interfaces and aqueous hydrates.

Simple model systems including small molecules of complex electronic structure as well as aqueous clusters offer a starting point in this process by providing the testbed for validating new approaches for analyzing the electronic structure as well as the nature of interactions and the magnitude of collective phenomena at the molecular level. For instance, high level first-principles electronic structure calculations of the structures, energetics, and vibrational spectra of aqueous neutral and ionic clusters provide useful information needed to assess the accuracy of reduced representations of intermolecular interactions, such as classical potentials used to model the macroscopic structural and thermodynamic properties of those systems. The database of accurate cluster structures, binding energies, and vibrational spectra can, furthermore, aid in the development of new density functionals, which are appropriate for studying the underlying interactions.

Of particular importance is the understanding of the factors controlling the affinity and selectivity of several molecular hosts to a variety of guest molecules with a particular emphasis on energy applications. The molecular level details of the various prototype guest/host systems, as probed experimentally via spectroscopic techniques and obtained theoretically by various levels of electronic structure theory, play an important role in the assessment of the accuracy of the latter. This information is subsequently used to model the guest/host interactions in complex molecular hosts such as hydrate lattices.

Representative applications include the modeling of liquid water and ice, aqueous ion solvation and applications, including the structure of clathrate hydrates and the interaction of host molecules with those guest networks. The detailed molecular-scale account of aqueous systems provided by these studies are relevant to Department of Energy programs in contaminant fate and transport and waste processing as well as hydrogen storage.

9.4.1.2 Scientific Objectives for 2017

The project aims to

1) Obtain high level benchmark results for the interaction of complex molecular systems such as medium size water clusters $(H_2O)_n$, $10 > n > 25$, accommodation of guest molecules (such as H_2, CH_4, CO_2) inside molecular hosts (i.e. hydrate lattices, crown ethers, nanotubes)
2) Explore the significance of multi-reference character for systems related to light harvesting, photovoltaics and catalytic processes
3) Push the state-of-the-art in scalable, accurate electronic structure calculations by combining expertise in domain science and computer science with specific targets to *efficiently* scale on hundreds of thousands of cores

9.4.2 Computational Strategies (now and in 2017)

9.4.2.1 Approach

The strategy is based on a hierarchical approach. Low and medium level calculations are run either on local resources or at NERSC utilizing a low core count (i.e. up to 500 cores). High-end calculations require large computational resources in excess of 100,000 cores.

As regards the methodological approach, the second order Moller-Plesset perturbation theory (MP2) level (scaling as N^5, where N is the size of the system), in conjunction with medium basis sets is used as the starting point in the calculation. Accurate energetics are obtained by expanding the basis sets, thus estimating the Complete Basis Set (CBS) limit. For some systems this is adequate, however it is oftentimes required to expand the level of theory to include additional electron correlation. This is done at the Coupled Cluster level that includes full single and double and a perturbative estimate of triples excitations [CCSD(T), scaling N^7]. Excited states are treated with the Equation-of-Motion approach [EOM]CCSD(T)], whereas multi-reference (MR) systems are treated via MR-CCSD(T).

The target by 2017 would be 120 atoms, 400 correlated electrons and 1,500 basis functions via the most accurate and most expensive method [MR-CCSD(T)].

9.4.2.2 Codes and Algorithms

The main code used to perform the calculations is the NWChem suite of electronic structure codes (developers: Apra, Kowalski). It is based on the use of the Global Arrays for the parallelization and ScaLAPACK for parallel linear algebra. The main kernel for (T) in CCSD(T) is matrix multiply.

For the high-end calculations the number of FLOPS is $\sim 10^{18}$, whereas the number of intermediate triples amplitudes: $\sim n_o^3 * n_u^3$; for example, for $(H_2O)_{24}$ / aug-cc-pVTZ : $\sim 10^{16}$. The approach has already shown to scale to the full length of petascale hardware (Jaguar (2009), Blue Waters etc.)

As a result of PNNL's investments under the "Extreme Scale Initiative" (Krishnamoorthy, Kowalski), it has been recently efficiently implemented on GPU technology. Preliminary tests on ORNL's Titan (2013) show ~6x speedup. Finally, the multi-Reference

implementation [MR-CCSD(T)] based on three levels of parallelism (reference level for MR part, task level within each reference and GPU for (T) contribution) is currently under development.

9.4.3 HPC Resources Used Today

9.4.3.1 Computational Hours

We used 13 million hours at NERSC in 2013. Other local (PNNL) resources such as PIC and Cascade (EMSL) are also used.

9.4.3.2 Parallelism

The code as it stands now can use an excess of 200,000+ cores for high-end production runs., but we typically use 500 – 20,000 cores today at NERSC. We note that NERSC scheduling does not appear to prioritize large jobs.

Our calculations process through different stages [HF, CCSD, CCSD(T)]. Each has a different level of efficiency, with the last part being about 95% of the total calculation.

We do not usually compute in High Throughput Computing mode, but some applications (such as the numerical calculation of second derivatives with CCSD(T)) can benefit from it.

Both strong scaling and weak scaling are important to us; however, strong scaling is the hallmark of the NWChem application.

9.4.3.3 Scratch Data

Scratch I/O is usually minimal (less than 10 TB total). Checkpoint and restart capabilities need to store intermediate amplitudes effectively up to 5 TB.

9.4.3.4 Shared Data

At the moment the project does not have a NERSC project directory, as there was no need for it. As more people are included in the project, the need might arise to create one.

9.4.3.5 Archival Data Storage

So far, we have minimal (< 1 TB) needs for HPSS storage.

9.4.4 HPC Requirements in 2017

9.4.4.1 Computational Hours Needed

We have an estimated need of 500M core hours. We arrived at this estimate using data obtained from previous INCITE awards, which were in excess of 100M hours.

Additional computational resources (~50M hours) are anticipated from local PNNL (PIC, Cascade @EMSL) hardware that will be obtained via a proposal submission process.

The primary factors driving the need for additional resources by two orders of magnitude are

(1) The need to investigate larger systems that go beyond model and approach realistic ones (i.e., the full light harvesting system including the antenna and the resonance energy transfer component),
(2) The need to assess the accuracy of electronic structure theories (such as local methods, Density Functionals, QMC) for larger systems by establishing "the benchmark" via high-end calculations,
(3) The need to strive for higher accuracy in the underlying theories.

9.4.4.2 Parallelism

Typically 250,000 cores could be used with the ability to use in excess of 1,000,000 if available and assuming that software required for the parallelization (Global Arrays) is tuned accordingly. We do not have a strong need to run more than one job concurrently but if the resources are available 3-5 jobs can be run concurrently. We will not use High Throughput Computing.

9.4.4.3 I/O

Scratch I/O is usually minimal (less than 200 TB total). Checkpoint and restart capabilities need to store intermediate amplitudes in CCSD(T) effectively up to 100 TB. We estimate needing 1 TB / sec I/O bandwidth.

By doing asynchronous I/O we are able to keep the time devoted to I/O down to a negligible amount, < 1 %.

9.4.4.4 Scratch Data

We estimate needing 1 PB in 2017. The primary cause of the growth in scratch space needs is the size of the system we simulate, future modifications of the code and implementation of new theories.

9.4.4.5 Shared Data

We estimate that our needs would be accommodated by a 1 TB quota in the NERSC /project file system

9.4.4.6 Archival Data Storage

We estimate that we'll need to archive about 200 TB to HPSS in 2017. Growth is caused by having more users and more runs or projects.

9.4.4.7 Memory Required

The current algorithms used can adopt to use as much memory per node as available. An increase of 5x in memory will be extremely beneficial.

9.4.4.8 Emerging Technologies and Programming Models

The path towards this has already started. Currently there are efforts to extend those implementations to heterogeneous architectures (GPU/Intel MIC). Speedup currently achieved is ~6x for numerically intensive parts of the calculation [(T) in CCSD(T)].

9.4.4.9 Software Applications and Tools

We will need RDMA extensions in MPI; efficient parallel linear algebra for heterogeneous architectures; efficient libraries for parallel I/O (not as crucial).

9.4.4.10 HPC Services

Consulting and account support services will be crucial. NERSC already has an excellent track record regarding those services. Most of the data are transferred back to local resources to be analyzed.

A very important issue is related to the various levels of fault tolerance, especially the ones that are difficult to detect and warn.

9.4.4.11 Time to Solution and Throughput

Implementing a policy around job scheduling to reduce the wait time in the queue for very large jobs would be extremely helpful.

9.4.4.12 Data Intensive Needs

No additional needs. HPSS is very good and we are very satisfied with the archival storage. We do not, as of this time, have a data management plan for our project.

9.4.4.13 What Else?

As NERSC moves to larger machines, the issue of fault tolerance becomes very important. HPC features that or most importance to us are size, speed, memory, and network.

9.4.4.14 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	13 M	500 M
Typical number of cores* used for production runs	500 – 20,000	250,000+
Maximum number of cores* that can be used for production runs	200,000+	1,000,000+
Data read and written per run	TB	< 200 TB
Maximum I/O bandwidth	GB/sec	1 TB/sec
Percent of runtime for I/O	0.1%	< 1%
Scratch File System space	10 TB	1 PB
Shared filesystem space	0 TB	1 TB
Archival data	622 GB	200 TB
Memory per node	GB	GB
Aggregate memory	TB	TB

* "Conventional" cores

9.5 Molecular Dynamics of PNIPAM Agglomerates and Composite Architectures

Principal Investigator: Sanket A. Deshmukh (Argonne National Laboratory)
Worksheet Authors: Derrick C. Mancini, Subramanian Sankaranarayanan, Ganesh Kamath, (Argonne National Laboratory)
NERSC Repositories: (1) m1528, Agglomeration dynamics in thermo-sensitive polymers across the lower critical solution temperature (Deshmukh is PI)
(2) m1524, Molecular dynamics simulation of PNIPAM-coated gold nanoparticles (Mancini is PI)

9.5.1 Project Description

9.5.1.1 Overview and Context

Our group's research is focused on the large-scale atomic-level modeling of temperature-sensitive polymers and direct comparison of our simulation results with the results of ongoing experimental work at the Advanced Photon Source (APS) division of Argonne National Laboratory (ANL). Mainly, we are conducting all-atom molecular dynamics (MD) simulations of Poly(n-isopropylacrylamide) (PNIPAM) in aqueous media across their lower critical solution temperature (LCST).[1-4] PNIPAM represents an important class of thermo-sensitive polymers that undergoes a coil-to-globule transition above the LCST around 32°C.[5] This coil-to-globule transition is also of great importance in a number of practical applications including energy storage and conversion, drug delivery, medical diagnostics, tissue engineering, electrophoresis, separation, and enhanced oil recovery.[6, 7] Brush structures of PNIPAM consist of the flexible macromolecules anchored by a special end group to a substrate at sufficiently large grafting density, such that different PNIPAM chains overlap. Such brush structures have potential applications as chemical valves.[8,9] For example, tuning the LCST of PNIPAM close to human body temperature *via* copolymerization can enable development of controlled drug delivery system.[10, 11]

Understanding the mechanism, thermodynamics, and kinetics of the conformational transformations of linear polymer chains is a fundamental problem in the field of polymer science.[12, 13] For thermo-sensitive ionic polymers, the change in conformation with temperature is the fundamental basis of thermal sensitivity. Specifically, for aqueous PNIPAM solutions, observations of phase separation or phase change at LCST occurs as a macroscopic manifestation of the coil-to-globule transition followed by aggregation.[14] Key to understanding this coil-to-globule transition in PNIPAM are the effect of the local structure of the surrounding medium, the effect of the interaction of surrounding medium with the polymer on the dynamics of agglomeration and the coil-to-globule conformational transition, and the exact mechanism of agglomeration of PNIPAM oligomers and high-molecular-weight PNIPAM chains. In the case of PNIPAM architectures such as brush structures, effect of grafting density and chain length of PNIPAM, the effect of hydrophobic/hydrophilic nature of substrate, and themorphology of substrate (sphere vs. planar) on the coil-to-globule transition are not very well understood. Using current experimental techniques it is very difficult to probe the conformation and extent of entanglement of PNIPAM chains, both below and above the LCST, in such brush structures. Even when x-ray, neutron, or dynamic light scattering measurements are used to study these transitions, it is necessary to have sufficiently accurate detailed models to correctly fit the data and interpret the results.

Our team has successfully studied the nature of these conformational transitions in oligomers of PNIPAM through the LCST by coupling high performance computing (HPC) with MD simulations. We have probed the local structure of the surrounding medium and its interactions with the polymer and on the conformational transitions, and thereby on the functional properties (e.g. diffusion) of the short-chain-length oligomers. Our results of all-atom MD simulations of PNIPAM consisting of 30-monomer units suggested that proximal water plays a key role in determining LCST behavior of PNIPAM. Additionally, the structure of proximal water also dramatically changes during this coil-to-globule transition of PNIPAM through the LCST. Today, very few theoretical studies, however, have been carried out on simulations of thermo-sensitive polymer architectures in presence of explicit solvent.[9, 12, 15-17] Recently, we have successfully carried out all-atom MD simulations of PNIPAM grafted brush structures (See Figure 1 (b)) in water. Depending upon the size of the nanoparticle and chain length of the PNIPAM chains, the size of these systems varied from *3 to 9 million* atoms. Initial results of our simulations suggest that PNIPAM chains are in a coil-like-state and entangled below the LCST. Above the LCST, however, the PNIPAM chains transform into a globule-like-state and locally collapse and agglomerate to result in a transformed morphology of the entire structure. Additionally, the structure of the proximal water dramatically changes as the PNIPAM chains undergo these coil-to-globule transitions.

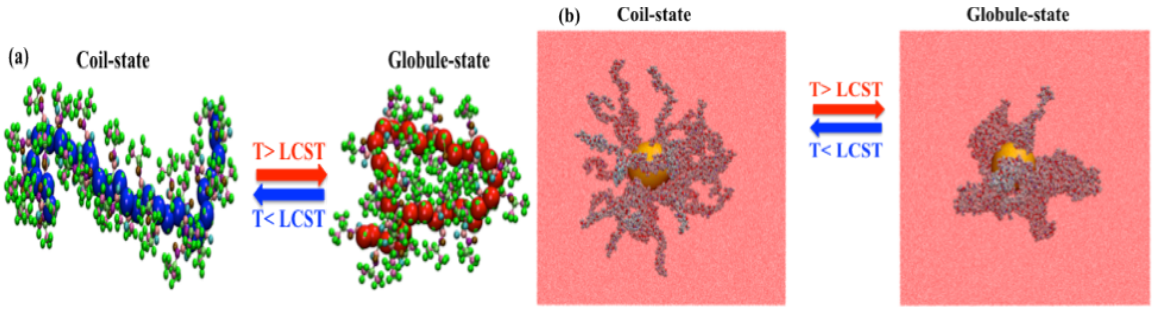


Figure 1: (a) Coil-to-globule transition of 30-mer of PNIPAM (water molecules are not shown for clarity) (b) Coil-to-globule transition of 60-mer of PNIPAM grafted on a gold nanoparticle in presence of water molecules (System consisting of ~9 million atoms). Gold nanoparticles and water are shown in yellow and red, respectively.

9.5.1.2 Scientific Objectives for 2017

In the case of MD simulations of PNIPAM brush structures: currently, we are probing effect of shape, size and nature of substrate on the coil-to-globule transition of PNIPAM and structure of water. Our long-term goals (year 2017) are to answer following questions: 1) What is the effect of chain length and grafting density polymer chains and size of nanoparticles on the morphology and geometry of the agglomerated or self-assembled structures? 2) What is the effect of increase in hydrophilicity or hydrophobicity of PNIPAM through copolymerization on the coil-to-globule transition of PNIPAM and self-assembly of nanoparticles? 3) How can the morphology be controlled by adjusting the hydrophilicity or hydrophobicity of nanoparticles? 4) What is the role of proximal water in driving the self-assembly of these polymer brush structures both below and above the LCST? 5) What are the dynamic properties and mechanism of nanoparticle-nanoparticle interactions mediated by the polymer brush structures? 6) What is the atomistically-derived driving force for the self-assembly of these polymer brush structures? And 7) How can we utilize the atomistically-derived interactions of these structures to create mesoscale coarse-grained models of their behavior that can be scaled up to model larger assemblies over longer times? To this end, we propose to carry out all-atom simulations of multiple polymer brush

structures by strategically placing them in vicinity of each other. This will allow us to extract information from atomic-level models of these brush structure, which will be utilized to build coarse-grained models of these brush structures. The overall aim of our coarse-graining approach is to provide a simple model that is computationally fast and easy to use, and can give significant insights on the structural and dynamics processes during the self-assembly of these polymer brush structures. Moreover, the results of these simulations can be used to construct models of the behavior of these assemblies for direct comparison to scattering measurements and microscopies.

In the case of agglomeration of PNIPAM chains both below and above the LCST, we propose to study the effect of various end groups and copolymers on the agglomeration of PNIPAM, as well as the effects of tacticity. Additionally, we also plan to study the effect of various salt ions, ionic liquids, and solvent mixtures (methanol-water, ethanol-water etc.) on the agglomeration of PNIPAM.

9.5.2 Computational Strategies (now and in 2017)

9.5.2.1 Approach

To study both agglomerations of PNIPAM as well as larger PNIPAM architectures (brush structures, nanogels, hydrogels etc.), we currently employ all-atom molecular dynamics (MD). In the case of PNIPAM brush structures, goal is to understand the effect of temperature, grafting density, and substrate used for grafting of the PNIPAM chains on the structural and dynamical properties of polymer brush structures of PNIPAM. The simulations of PNIPAM chains grafter on gold nanoparticles of ~6, ~10, and ~15 nm diameter with grafting densities in the range of 0.05 and 0.4 chains/nm^2 are performed at 278 K and 310 K to study the effect of temperature on the polymer structure and evaluate the structural phase transition in PNIPAM. Our simulations of single chain PNIPAM suggests that to observe a clear coil-to-globule transitions in PNIPAM MD simulations must be carried out for ~20 ns. Hence, we conduct simulations up to ~25-30 ns with a time step of 1fs. The atomic trajectories (atom positions, temperature, pressure, velocities etc.) will be accumulated and stored every 1ps. Trajectory files obtained from these simulations will be analysed for various dynamical properties.

In the case of agglomeration of PNIPAM chains We initiated our study with two chains of PNIPAM chains consisting of 100 monomer units (100-mer) followed by insertion of water molecules in a simulation cell. This system was relaxed for ~50 ns which was followed by random insertion of a third chain of 100-mer both at 275 K and 325 K. Again the system was relaxed for ~30 ns both at 275 K and 325 K. This procedure was carried out till we have 5 chains of 100-mer.

In the future, to study both agglomeration and polymer architectures, we plan to use meso-scale models, which will allow us to study systems consisting of billions of atoms and up to timescales of microseconds. This will allow the direct comparison of our simulation results with the experimental results. In addition, we also propose to study in a similar way the copolymers of PNIPAM as well as other ionic polymers, particularly of potential interest to energy and environmental problems.

9.5.2.2 Codes and Algorithms

In this study to conduct MD simulations, we employ the NAMD simulation package with CHARMM force fields, which is designed for parallel computation. Full and efficient treatment of electrostatic and van der Waals interactions are provided via the (O(NLogN) Particle Mesh Ewald algorithm. NAMD includes a rich set of MD features, such as multiple time stepping, constraints, coarse-grain force-fields, and dissipative dynamics. Table 1 gives the scaling data for NAMD using CHARMM force-field for a PNIPAM-water system consisting of ~3M atom on HOPPER.

Table 1. Scaling study for a brush structure of PNIPAM in water molecules (~3 million atoms) on Hopper.

Strong scaling of PNIPAM-water system for 3 ns with time-step of 1fs.		
Cores	Time (minutes)	Efficiency
3072	~2160	1.0
6144	~1080	1.0
9216	~1075	0.5
12228	~1070	0.5

9.5.3 HPC Resources Used Today

9.5.3.1 Computational Hours

Our projects m1528 and m1524 used 13 M computing hours at NERSC. In addition to the NERSC computing facility we are getting support from the Argonne Leadership Computing Facility (ALCF) at the Argonne National Laboratory (ANL).

9.5.3.2 Parallelism

The number of cores we use on Hopper varies from 4,096-8,192, depending on our system size. For example, to conduct simulations of systems consisting of ~400 K atoms we utilize 4,096 cores; to conduct simulations with a million atom systems, we utilize 8,192 cores.

We mainly use NAMD with CHARMM force field to conduct our simulations. NAMD can be scaled up to ~32 K cores.

The system sizes of our simulations vary between ~400 K atoms to ~10 M atoms. For these system sizes NAMD scales better on the core range of 4096-8192 for currently used machines.

We do not have any computation in High Throughput Mode.

For our project, both strong and weak scaling are equally important.

9.5.3.3 Scratch Data

The system sizes of our simulations vary from 400 K atom to 10 M atoms. To study coil-to-globule transition in PNIPAM, we need to conduct simulations for ~30 ns or more. To access the structural and dynamical properties of PNIPAM and proximal water, we need to store

our trajectories every 1 ps. Given the system size, our trajectory files can be in the range of 100-500 GB/simulation.

9.5.3.4 Shared Data

Our project has been allocated a project directory named "pnipammd." The primary reason for this space is to store intermediate-term archive data generated from our simulations.

9.5.3.5 Archival Data Storage

The directory was allocated recently. We have not stored any data as of now but we expect to store ~4-5 TB of data by the end of 2013.

9.5.4 HPC Requirements in 2017

9.5.4.1 Computational Hours Needed

Based on the current scaling data for the software on the conventional machines and given the large-scale of simulation models of polymer brush structures and various agglomeration studies of PNIPAM copolymers that will be treated in the proposed work, we expect that a large amount of time will be required, and therefore would request approximately **~500,000,000 core hours** of CPU time on the conventional machine.

We expect to receive computing support from ALCF computing facility at the ANL for related projects.

In the next stage of our research plan we propose to develop computational models of different polymer architectures at multiple levels (atomic- to meso-scale) to study the conformational transformations and agglomeration behavior and their role in self-assembly. Given the large-scale simulation models of self-assembly of polymer brush structures that will be treated in the proposed work, we expect that a large amount of time will be required.

9.5.4.2 Parallelism

We expect to use ~32 to 64 K computing cores in 2017.

Currently NAMD shows strong scaling up to ~32 K cores. We expect we can use at least 32 K cores by 2017.

We expect to run at least two jobs concurrently. The aim of our project, in 2017, is studying the self-assembly of nanoparticles composites below and above the LCST of PNIPAM. Hence, we expect to run simulations for at least at two temperatures, below and above the LCST of PNIPAM.

9.5.4.3 I/O

See the sections immediately below.

9.5.4.4 Scratch Data

Currently, we need temporary disk space of ~4-5 TB for our simulations. In the future, we expect to utilize 8-10 TB of disk space, as the system size of our simulations will be much larger than what we are currently simulating. Additionally, to compute various structural

and dynamical properties of materials and solvent we will be storing trajectory files every pico-second. This will lead to generation of trajectory files of ~2-3 TB/simulations. As we will be running more than 1 simulation concurrently, we expect temporary disk space of 8-10 TB.

As mentioned earlier, we propose to simulate self-assembly of polymer nanoparticles. In this study we propose to use meso-scale models and to study realistic system sizes of billions of atoms, which will allow us the direct comparison of our experimental results with the results of our simulations. In the initial stages of these simulations all-atom systems consisting of millions of atoms will be equilibrated for ~5-6 ns. Coarse-graining of the simulated system will follow this and further simulations will be conducted for 5-6 microseconds to capture the dynamics of self-assembly of these brush structures. As the end of simulation run of 5-6 microseconds, coarse-grained model will be transformed back to all-atom model and simulations will be conducted for another ~10 ns to retain and study the atomic level structure of these polymer brushes. These factors are what lead to the increase in scratch data required.

9.5.4.5 Shared Data

We expect requirements for NERSC project directory space of ~20 TB.

We propose to study the more realistic system sizes for self-assembly of PNIPAM coated nanoparticles, which will allow us direct comparison of results with our experimental results. The systems will consist of millions of atoms and the trajectories generated will cover simulation runs of microseconds. The size of each trajectory will be in 2-3 TB and we will be running at least 5-6 such simulations under different conditions. These factors are what lead to the increase in project data required.

9.5.4.6 Archival Data Storage

We would be storing all the data generated from our simulations. We estimate 20 TB of data to be generated in 2017. Increase in the system size and simulation time is causing the growth in archival storage.

9.5.4.7 Memory Required

NAMD has traditionally used less than 100 MB of memory even for systems of 100,000 atoms. With the reintroduction of pair lists in NAMD v2.5, however, memory usage for a 100,000-atom system with a 12-Å cutoff can approach 300 MB, and will grow with the cube of the cutoff. This extra memory is distributed across processors during a parallel run. We expect to use the 12-Å cutoff for our systems; hence, we estimate the 300 to 500 MB memory for our systems.

9.5.4.8 Emerging Technologies and Programming Models

The NAMD code is ready to be used on GPUs. We are successfully using NAMD on the GPUs on the "Carbon" cluster at the Center for Nanoscale Material, Argonne National Laboratory. Preliminary observations suggest that NAMD is faster on GPUs than traditional CPU cores.

9.5.4.9 Software Applications and Tools

We will need following software and libraries:

MD software: NAMD, LAMMPS

Libraries: MPI, FFTW

Compilers: FORTRAN, C, C++, PYTHON

Visualization software: VMD, RASMOL

9.5.4.10 HPC Services

We need consulting and account support, data analytics, and visualization.

9.5.4.11 Time to Solution and Throughput

N/A

9.5.4.12 Data Intensive Needs

While intermediate-term archival storage is being made on NERSC systems, as important is efficient file transfer from NERSC to local data systems for analysis and long-term archival storage. We are satisfied with NERSC's HPSS system.

We already have proposed budget to procure storage in 2017 to provide for our project's data management plan.

9.5.4.13 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	13 M	500 M
Typical number of cores* used for production runs	4,096 – 8,192	~32K - 64K
Maximum number of cores* that can be used for production runs	~32K	~32K - 64K
Data read and written per run	~1.0 – 1.5 TB	~2 - 3 TB
Scratch File System space	~4 - 5 TB	~15 - 20 TB
Shared filesystem space	0 TB	20 TB
Archival data	0 TB	20 TB
Memory per node	2-3 GB	4-5 GB

* "Conventional" cores

References

[1] Sanket A. Deshmukh, Subramanian K.R.S. Sankaranarayanan, Kamlesh Suthar, and D. C. Mancini, Submitted to Journal of American Chemical Society (2011).

[2] G. Kamath, S. A. Deshmukh, G. A. Baker, D. C. Mancini, and S. K. R. S. Sankaranarayanan, Phys. Chem. Chem. Phys **15**, 12667 (2013).

[3] S. A. Deshmukh, Z. Li, G. Kamath, K. J. Suthar, S. K. R. S. Sankaranarayanan, and D. C. Mancini, Polymer **54**, 210 (2013).

[4] S. A. Deshmukh, S. K. R. S. Sankaranarayanan, and D. C. Mancini, The Journal of Physical Chemistry C **116**, 5501 (2012).

[5] H. G. Schild, Progress in Polymer Science **17**, 163 (1992).

[6] Martien A. Cohen Stuart, et al., Nature Materials **9**, 101 (2010).

[7] A. Lendlein; and V. P. Shastri, Advanced Materials **22**, 3344 (2010).

[8] T. Yoshinobu, O. Kohji, Y. Shinpei, G. Atsushi, and F. Takeshi, Advances in Polymer Science **197**, 1 (2006).

[9] D. I. Dimitrov, A. Milchev, and K. Binder, The Journal of Chemical Physics **27**, 084905 (2007).

[10] J. E. Chung, M. Yokoyama, M. Yamato, T. Aoyagi, Y. Sakurai, and T. Okano, Journal of Controlled Release **62**, 115 (2002).

[11] A. S. Hoffman, Journal of Controlled Release **6**, 297 (1987).

[12] Robert M. Briber, Xiaodu Liu, and B. J. Bauer, Science **268**, 395 (1995).

[13] Eva Harth, Brooke Van Horn, Victor Y. Lee, David S. Germack, Chad P. Gonzales, Robert D. Miller, and C. J. Hawker, Journal of American Chemical Society **124**, 8653 (2002).

[14] Alexander D. MacKerell Jr., Nilesh Banavali, and N. Foloppe, Biopolymers **56**, 257 (2000).

[15] P. Y. Lai and K. Binder, The Journal of Chemical Physics **95**, 9288 (1991).

[16] J. M. D. Lane, A. E. Ismail, M. Chandross, C. D. Lorenz, and G. S. Grest, Physical Review E **79**, 050501 (2009).

[17] M. Murat and G. S. Grest, Physical Review Letters **63**, 1074 (1989).

9.6 Sampling Diffusive Dynamics on Long Timescales, and Simulating the Coupled Dynamics of Electrons and Nuclei

Principal Investigator: Thomas Miller (California Institute of Technology)
NERSC Repository: m822

9.6.1 Project Description

9.6.1.1 Overview and Context

The goal of this research is to develop and employ new theoretical and computational methods for understanding the dynamics of complex systems. We are focused on two main areas of research that are of fundamental interest to the DOE-BES mission: (i) coupled electronic and nuclear dynamics in enzymes and photo-catalysts and (ii) long-timescale dynamics in protein-transport processes involving transmembrane channels. A critical aspect of this research is the development of simulation algorithms to leverage the massively parallel computational systems. The support of NERSC computer resources is critical in our efforts to understand and design of chemical processes that are critical for solar energy conversion, enzyme catalysis, and biomolecular transport. This project is of direct relevance to the DOE basic energy science (BES) mission.

All aspects of our work utilize NERSC resources – a critical resource for our progress. Available computational resources limit many applications, so we need more hardware access, increased numbers of available processors and increased processor speed. Also, NERSC support is important for our efforts.

Public software is a less critical bottleneck for us at this time, but the currently available codes are essential to our work.

9.6.1.2 Scientific Objectives for 2017

By 2017 we expect to develop and utilize new methods to achieve these two focuses.

9.6.2 Computational Strategies (now and in 2017)

9.6.2.1 Approach

We approach these problems computationally at a high level by using quantized molecular dynamics (MD), classical MD, and course-grained simulation methods.

9.6.2.2 Codes and Algorithms

The codes we use are either developed in-house or are modified versions of existing programs, such as NAMD, GROMACS, DL_POLY, or AMBER. These codes are characterized by these algorithms: molecular dynamics (ODE) integration and FFTs. Our biggest computational challenges are FLOP availability and queue times. Parallel scaling is typically limited by potential energy surface evaluation, number of DOFs, and number of weakly coupled trajectories.

By 2017 we expect to make increased use of GPUs but we also expect increased computational cost (and parallel scalability) of potential energy surface evaluations. We also expect to simulate large systems for longer times.

9.6.3 HPC Resources Used Today

9.6.3.1 Computational Hours

We use Carver, Edison, and Hopper at NERSC and used about 21.4 M hours during 2013. We also used about 9M hours on Jaguar at OLCF and about 8M at ALCF (on Intrepid and Vespa).

9.6.3.2 Parallelism

We typically use 500-2,000 processors per set of jobs (tight parallelization for trajectories, weak among them), which take about 6-12 hours per trajectory, and we do hundreds of such runs per year. Memory usage per core is modest: 50 - 100 MB / core.

9.6.3.3 Scratch Data

Data read or written during our runs is modest: 1-5 GB.

9.6.3.4 Shared Data

We have only 2 GB stored in /project today, but we estimate needing 4 TB of shared space in 2017.

9.6.3.5 Archival Data Storage

We have 33 TB stored in the NERSC HPSS Archive system in 2013. Based on past growth rates, we expect to have about 75 TB stored in 2017.

9.6.4 HPC Requirements in 2017

9.6.4.1 Computational Hours Needed

We estimate needing 150 M hours, due to increasing demands of potential energy surface (PES) evaluations.

The primary driving factor is the need to make advances in the following areas:

- Quantitative prediction of biological and catalytic processes, design of new catalysts;
- Theory-guided enhancement of integral membrane protein expression;
- Progress towards the computational design of electrolyte and electrode materials.

9.6.4.2 Parallelism

We expect a 10-to-20-fold increase in concurrency due to greater PES parallelizability with no other major changes. We do not use high throughput computing. We expect no changes to the amount of data read or written, no changes to the software we require, and expect to need no more than about 1 GB/core of memory. We ask that NERSC recognize the value of ensemble calculations, while also encouraging efficient parallelization.

9.6.4.3 Emerging Technologies and Programming Models

We have expanded utilization of GPUs in many/most simulation studies (not at NERSC). To date we have prepared for manycore by:

- Utilization of local (Caltech-based) GPU machines
- Utilization of GPU implementations of classical MD packages
- Working with CS groups at Caltech and Pomona Colleges to develop efficient GPU versions of the coarse-grained simulations

We are already planning to also develop and implement GPU versions of existing codes.

To be successful on many-core systems we will need help with efficient implementation and scaling tests.

9.6.4.4 Other Comments

Please don't make the queues too short.

9.6.4.5 Requirements Summary

	Used at NERSC in 2013	**Needed at NERSC in 2017**
Computational Hours	21.4 M	150 M
Typical number of cores* used for production runs	500-2,000	5,000-40,000
Maximum number of cores* that can be used for production runs	2,000	40,000
Shared filesystem space	2 GB	4 TB
Archival data	33 TB	75 TB
Memory per node	50-100 MB/core	

* "Conventional" cores

10 Geoscience Case Studies

10.1 Large Scale Geophysical Inversion & Imaging

Principal Investigator: Gregory Newman (Lawrence Berkeley National Laboratory)
NERSC Repository: m372

10.1.1 Project Description

10.1.1.1 Overview and Context

Geophysical Imaging – The Basics: Geophysical imaging involves mapping the physical attributes of the Earth's subsurface, such as electrical conductivity (resistivity), seismic velocity, mass density, fluid saturation, etc. This requires thousands of simulations, each of which solves partial differential equations (PDE) that describe the physical fields of interest. Examples of such fields arising in geophysical prospecting include gravity, electrostatic, electromagnetic (EM), acoustic and elastic waves. From those solutions, the attributes of interest can be derived.

The solution of the corresponding imaging problem typically seeks to minimize the errors between the observed and simulated field, typically in a least squares statistical sense based upon an L_2 norm; sometimes large outliers in the noise of the observations can produce significant bias in the solution and an L_1 norm may be preferred to mitigate the overstated influence of outliers. The size of imaging and data volumes required for a 3D imaging experiment is considerable. Millions of unknowns are needed to describe the 3D distribution of the subsurface attributes and data density exceeding millions of observations may be necessary to provide sufficient spatial aperture necessary for adequate sensitivity to the subsurface. In seismic imaging, data volumes can easily exceed several terabytes arising from 1000's of sources and 1000's of detectors per source.

Our research objectives will be to further develop and apply three-dimensional (3D) geophysical imaging methods, incorporating electromagnetic, and extended to gravity data and seismic data under a joint geophysical imaging framework. Our approach is to use as much rigor as possible and to avoid approximations, which sacrifice solution accuracy for speed. The need for a full solution to the geophysical imaging problem has been critical since we are now focusing on joint imaging problems in complex geological terrains, where fast 2D and 3D approximate methods are unsatisfactory. Moreover, to better model complex geological formations, we are now implementing a finite element module for EM field simulation and imaging, in which the mesh can now conform to these geological interfaces. Our goal is to image the subsurface geophysical properties (electrical conductivity, density, and seismic velocity and other elastic attributes) in three spatial dimensions with sufficient spatial resolution to advance energy resource exploitation and address the DOE and Office of Science mission in energy security. Our work also addresses the DOE mission in remediating the nation's waste legacy issues and DOE sponsored projects for protection and safeguarding of the subsurface environment and water resources. An outstanding feature of our work is the use of massively parallel (MP) and GPU computers to create realistic imaging solutions that cannot yet be achieved with serial machines or modest size PC clusters. Up scaling the imaging process to tens of thousands of processors and beyond so that one imaging (inversion) experiment of a field data set can be carried out in days rather than months, as is now the case, will continue to be a priority.

10.1.1.2 Scientific Objectives for 2017

By 2017 we will have completed implementation of a massively parallel 3D elastic wave field imaging code along with joint seismic-electromagnetic imaging capabilities for characterizing subsurface elastic attributes, which include mass density, compression and shear velocity, bulk and shear modulus and electrical resistivity. Demonstration of the imaging codes on complex 3D test models and data sets will demonstrate their capabilities and potential. Because the 3D elastic and joint imaging codes will come with considerable computational expense, the use of faster solvers will be required. Development of new scalable algebraic multi-grid solvers and preconditioners in complex arithmetic will be essential to meet our scientific objectives.

10.1.2 Computational Strategies (now and in 2017)

10.1.2.1 Approach

Practical solutions of the 3D imaging problem are based upon non-linear least squares optimization principles that uses some variant of Newton, Gauss Newton, non-linear/steepest decent techniques. They are often termed "deterministic imaging" because of their reliance on a good starting model. With large scale model parameterizations of the attributes and the data volumes needed for a viable 3D imaging experiment, it is best to avoid direct formulation of the Hessian and Jacobean in the solution of a deterministic optimization problem. These operators are dense and the cost of direct formulation is considerable, even on a distributed computational platform. Instead, it is advisable to only compute the action of these operators on a vector. Because the inversion process is non-unique and unstable additional steps are required to stabilize or regularize the problem. These include constraints on any acceptable models based on a priori information, and a means to appraise what features in these models are necessary to explain the data. As a natural consequence, one solution to the inverse problem is insufficient. Rather multiple solutions are needed using different assumptions on the background geology, regularization, data noise, and starting models used to launch the non-linear inversion process.

High performance computing (HPC) resources are essential for realization of 3D geophysical imaging experiments within an acceptable timeframe. Massively parallel (MP), multiple instruction multiple data (MIMD) machines have been the standard platform for high performance computation for nearly the last 24 years. These machines have dedicated access to 10's to 100,000's of compute tasks. The smallest MIMD machines (clusters) are typically a few tens to hundreds of compute tasks, while the largest machines are to be found in super-computing centers around the world, such as the National Energy Research Scientific Computing (NERSC) center. Large and small MIMD machines rely on a dedicated backbone for communication amongst the computing cores. The standard Message Passing Interface (MPI) is used for inter-processor communication. MPI provides portability so MP software can be run across a range of MIMD platforms, including dedicated distributed machines and/or distributed network of machines. The recent arrival of graphics processing units (GPUs) is allowing new alternatives for solving geophysical imaging problems on HPC platforms. Recent implementations of geophysical imaging software on GPUs shows encouraging cost versus performance metrics compared to MIMD architectures; cost includes not only money but time. In the last few years GPUs have gained considerable favor in the oil and gas industry for seismic imaging.

Computational approaches for 3D simulation of geophysical wave fields (i.e. seismic and electromagnetic) will favor finite difference and finite element implementations because of their flexibility in modeling complex geologies. Successful approaches in dealing with large-scale model parameterizations and data volumes exploit multiple levels of parallelization. At one level of parallelization, a domain decomposition of the attributes (the model) is made across a subset of computational tasks, called a local processor group, and is distributed across multiple copies of these groups. Because the data calculations are independent on each group, results are embarrassingly parallel and scale with the number of groups employed. Implicit solver methods take advantage of this processor topology, where iterative Krylov solvers are ideally suited for such problems, since each calculation is independent of the others. A similar strategy can be used for solving wave propagation problems with explicit methods that exploit some type of time stepping scheme. We have previously demonstrated the efficiencies that can be gained with such distributed computations where in one experiment we employed 32,768 computing tasks (compute cores) on the IBM Blue Gene Machine to solve a 3D EM imaging problem. While this application was demonstrated in 2008, at the time it required enormous resources to execute, and clearly demonstrated that 3D imaging problem could be solved in days, rather than weeks on more modest size clusters. It offered important verification that with industrial size imaging problems one could exploit fine and course gain parallelism to achieve solutions on a time scale acceptable to energy exploration companies.

Recently, there has been interest in multi-frontal direct solvers in large-scale geophysical field simulations, which can easily be adapted to many-core architectures. Popular, parallel multi-front solvers libraries include MUMPS, Super LU, and PARDISO, among others. While there is much appeal in exploiting these parallel solvers, large model parameterizations and corresponding meshes that arise in 3D geophysical imaging applications will in general limit their applicability to modest size problems.

We believe our main computational strategy in 2017 will focus on finite element solutions and associated imaging using the elastic-dynamic equations arising in seismology, along with Maxwell's equations and the associated DC equations on unstructured meshes. The linear systems that result from unstructured meshes are large, sparse, and highly ill conditioned. Very efficient and scalable iterative solutions to these linear systems will be needed. In this context, it is interesting to note that algebraic multi-grid (AMG) solvers have not received much attention for the types of geophysical field simulation problems that are of interest to us. AMG is reportedly the optimal iterative solution method that can be applied to sparse linear systems. While AMG (as well as geometric multigrid solvers) have been successfully developed for real systems, DC type problems for example, application to complex, and complex-symmetric systems have proven more difficult and elusive. One possible reason may be attempts to use AMG libraries designed for real systems on complex systems, which are expressed in terms of equivalent real forms (ERF). Interestingly, Freund in 1992 warned that Krylov solvers designed for real systems, such as GMRES, were not very effective when applied to complex symmetric linear systems expressed in ERF. Freund claimed that it is better to solve such systems directly in the complex domain, and developed the QMR and the transpose-free QMR methods for that task. One reason for poor performance on ERF is the resulting complex eigenvalue distribution, which is folded about the real axis, and this folding could cause problems with the Krylov iteration. From my own experience, a similar problem is observed when AMG, designed for real systems, is applied to complex-symmetric systems. Development of robust AMG solvers formulated specifically

for complex linear systems (symmetric and non symmetric) is clearly needed and should be made available for production MP scientific computation by 2017.

In anticipation of enhanced NERSC computing resources approaching 100's of Petaflops, we will embark to solve geophysical simulation and imaging problems, and corresponding data volumes at unprecedented size and scale, exceeding one billion grid points and data sets exceeding tens of millions of measurements. For elastic wave propagation and imaging, problems of this size will be especially challenging, but the computational resources are expected to be available to tackle the problem. We therefore anticipate using on the order of 900,000,000 core hours in 2017.

10.1.2.2 Codes and Algorithms

Code EMGeo: Simulates and Inverts electromagnetic fields for subsurface conductivity in three spatial dimensions; uses finite difference approximations, preconditioned iterative Krylov solvers, explicit time-stepping methods, and preconditioned gradient optimization methodologies. Techniques now extended to gravitational fields and seismic wave field propagation simulation and imaging problems, concurrently. Geophysical attributes that can now be imaged included seismic velocities, mass density, shear modulus, bulk modulus and electrical conductivity.

10.1.3 HPC Resources Used in 2013

10.1.3.1 Computational Hours

We used 29 million hours at NERSC in 2013.

10.1.3.2 Parallelism

We currently use 5,000 to 20,000 compute cores at NERSC. The maximum number of cores that EMGeo can use for production runs today is 153,216 compute cores, which comprises all the compute cores on Hopper. Because shared allocation of machine resources, we use far fewer compute nodes because requests for more resources that would limit our production due to excessive queue wait times.

We do not currently use High Throughput Computing mode.

Our problems exhibit strong scaling. We have a problem of a given size and we need to use parallel computing to solve it in an acceptable timeframe.

10.1.3.3 Scratch Data

We need approximately 2 Terabytes of temporary space.

10.1.3.4 Shared Data

We do not currently have such a NERSC project directory.

10.1.3.5 Archival Data Storage

Our project did not take advantage of the NERSC archival data storage facility.

10.1.4 HPC Requirements in 2017

10.1.4.1 Computational Hours Needed

At a minimum, we would need 300,000,000 core hours in 2017 (30X our 2013 allocation) and could use 900,000,000 (30X our actual 2013 usage) to achieve our research goals.

The size of imaging and data volumes required for a 3D imaging experiment will expand considerably. Millions of unknowns will be needed to describe the 3D distribution of the subsurface attributes and millions of observations will be exploited to provide sufficient spatial aperture for adequate sensitivity to the subsurface. In seismic imaging, data volumes can easily exceed several terabytes arising from 1000's of sources and 1000's of detectors per source.

We do not expect to receive any significant allocations from sources other than NERSC.

10.1.4.2 Parallelism

We expect to use 20,000 to 50,000 cores per run.

Over 1,000,000 compute cores could easily be exploited for the largest seismic data sets.

We might need several jobs running concurrently.

10.1.4.3 I/O

We expect to write about 10 Terabytes per run in 2017. Depending upon the application, approximately 16 to 160 Gigabytes of data would be written to scratch for checkpointing.

We would be willing to devote only 10 percent of the total runtime to I/O but our problems are not typically I/O constrained. They are CPU constrained.

10.1.4.4 Scratch Data

For the largest computational problems, we estimate needing 2 to 10 Terabytes temporary disk .

10.1.4.5 Shared Data

We do not anticipate the need of anything beyond the NERSC default project space of 1 TB per project.

10.1.4.6 Archival Data Storage

We do not anticipate the need to store a significant amount of archival data for 2017.

10.1.4.7 Memory Required

This will depend on the number of compute core per node. The current Hopper system has 24 compute cores per node with 32 GB of memory per node on most of its nodes (6,000); 300 nodes have double that 64 GB. Expanding the memory to 100 GB per node by 2017 would be beneficial.

10.1.4.8 Emerging Technologies and Programming Models

While we have taken advantage of GPU architectures, we still favor the CPU-MPI interconnect for performance. A very strong case would have to be made to for us to engage in any significant reprogramming of our software to take advantage of specialized and heterogeneous computing environments. We conducted direct comparisons of our applications on GPU - CPU/MPI configurations. Our finding was that the speedups from GPUs, while impressive, were still not optimal; where for the largest applications, CPU/MPI configurations would out-preform the computation of a single GPU. We did not attempt the computations across multiple GPUs because of inefficiencies in message passing between GPUs. Clever reprogramming of our application on a heterogeneous MPI/CPU/GPU system could solve this problem, provided we see that we can scale up the size of our computations and are prudent in how the computational work is distributed across the machine. Once such a system is there that will scale to the levels we need, we will take advantage of it to optimize performance of our codes.

Regarding OpenMP, the current view is that we would not see that much benefit. However we can still investigate this further to see if better performance can be achieved.

In sum, we will engage in significant reprogramming of our software to take advantage of specialized and heterogeneous computing environments when there is a clear benefit to performance in our computations.

10.1.4.9 Software Applications and Tools

We need support for Fortran 90 and 95, MPI, C, C++, CUDA and Open CL for GPU technologies.

10.1.4.10 HPC Services

We do not know yet what services we'll require in 2017.

10.1.4.11 Time to Solution and Throughput

Time to solution is a big deal for us. We will need faster and bigger computing machines to advance our work. We would like to see a 10x speed up in our time to solutions. Any combination of increase in processing power, number of compute core and speed, would be ideal.

10.1.4.12 Data Intensive Needs

We do not at this time envisage additional needs, beyond an increase in computational power, as described above, that can be applied to our problems. This will allow us to expand the size of the data sets we would treat.

So far we have not used NERSC's HPSS system, so we cannot comment on it.

We do not at this time have a data management plan in place for our project.

10.1.4.13 Additional Comments

At this point no, our main concern is for NERSC to expand the size and speed of the computational systems. Faster job turnaround is the most important and useful service that NERSC could provide.

10.1.4.14 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	29 M	900 M
Typical number of cores* used for production runs	18,000	50,000
Maximum number of cores* that can be used for production runs	Greater than 10,000	Greater than 300,000
Percent of runtime for I/O	< 10 percent	< 10 Percent
Scratch File System space	2 TB	20 TB
Shared filesystem space	16 GB	1 TB
Archival data	172 GB	10 TB
Memory per node	32 GB	100 GB
Aggregate memory	modest estimate 10 TB	modest estimate 30TB

* "Conventional" cores

10.2 Computational Studies in Molecular Geochemistry

Principal Investigator: Andrew R. Felmy (Pacific Northwest National Laboratory)
Additional Worksheet Author: Eric J. Bylaska (Pacific Northwest National Laboratory)
NERSC Repository: mp119

10.2.1 Project Description

10.2.1.1 Overview and Context

The chemical complexity and time scale of the problems likely to be encountered in new energy strategies (e.g., CO_2 sequestration, nuclear waste storage, energy storage materials) will require a much larger dependence on molecular simulation for interpretation, impact assessment and design [1]. The difficulty of simulating these processes is greatly increased by the sensitivity of the processes at the macroscopic scale to the atomic scale; the unusual/unexpected bonding behaviors of the materials; complex, defected and extreme temperature and pressure environments likely to be encountered; and the requirements that simulations be parameter free as possible and extremely reliable. The well-developed tools of quantum chemistry and physics have been shown to approach the accuracy required. However, despite the continuous effort being put into improving their accuracy and efficiency, these tools will be of little value to condensed matter problems without dramatic improvements in techniques to traverse and sample the high-dimensional phase space needed to span the $\sim 10^{12}$ time scale differences between molecular simulation and chemical events. Current methods for exploring phase space are imperfect; e.g., explicit time integrators for non-linear differential equations are not parallelized and require small time steps, implicit time integrators show significant energy drift, free energy methods need very large numbers of iterations to converge even simple processes, and search methods for complex processes require appropriate order parameters that are often unknown. New methods for time integration, efficient exploration of phase space, and choosing order parameters will be needed.

The NERSC project Computational Studies in Molecular Geochemistry supports the Basic Energy Sciences (BES) Geosciences Research Program managed by PNNL. The BES research program at PNNL supports research in mineral electron conduction/transfer, mineral and water film nucleation and growth, and the development of new theoretical approaches for predicting interfacial reactivity. Our effort over the next three years will focus on developing improved computational models to better describe reaction paths in the nucleation and growth of minerals.

[1] DOE BES Basic Research Needs Reports (http://science.energy.gov/bes/news-and-resources/reports/basic-research-needs/)

10.2.1.2 Scientific Objectives for 2017

Macroscopically observable mineral assemblages are often not the result of equilibrium processes but instead are determined by the dynamics of mineral nucleation and growth. Hence the mineral assemblages that we observe are often dependent upon the initial molecular or microscopic events that occur early, in the nucleation of a phase, or later, in the difference in growth rates between the different mineral phases. In that regard, we often observe minerals whose presence is determined by complex reaction paths that begin with the initial steps in mineral nucleation. As an example, in our current research we are

concerned with the nucleation and growth of magnesium carbonate minerals because of their possible impact on the long-term sequestration of carbon dioxide in the subsurface. In such cases, the maximum mass of carbon dioxide that would be sequestrated is related to the thermodynamic stability of the phases. The more stable magnesium carbonate phase, the greater mass of carbon dioxide is sequestered. Unfortunately, thermodynamic stability does not turn out to be a good predictor of the magnesium carbonate phases that actually form. Instead the exact conditions of mineral nucleation and subsequent rates of growth determine the magnesium carbonate phases that are experimentally observed (see Figure 1). As a result, to accurately predict the minerals that form in geochemical systems we must develop the capability to predict the reaction paths that different phases follow from their initial nucleation through their subsequent growth phases until they obtain macroscopically observable size. Unfortunately, this goal is impossible to achieve entirely by experiment owing to the wide range of temperatures, pressures, and solution phase conditions that may occur.

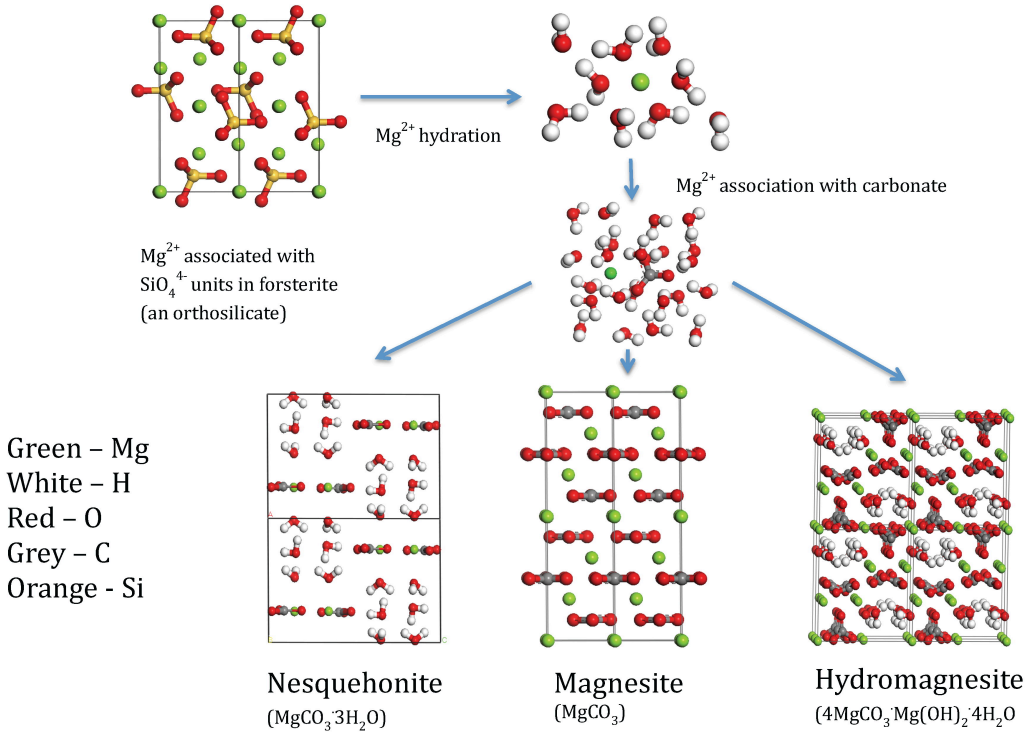


Figure 1. Possible reaction paths in the transformation of an orthosilicate mineral (forsterite) to possible magnesium carbonate minerals in the presence of CO_2. The reaction path involves the dissolution of Mg^{2+} from the forsterite surface, hydration and/or carbonation of the Mg^{2+} ion in solution, followed by the precipitation of magnesium carbonate mineral phases depending on the dynamics of the transformation.

Hence our major goal over the next three years is to develop the capability to follow reaction paths for mineral nucleation and growth. Such a capability would have a wide impact on the entire field of geosciences and in fact the entire field of chemical sciences.

10.2.2 Computational Strategies (now and in 2017)

10.2.2.1 Approach

Development of new petascale and cloud (distributed computing) based computational tools required to treat the highly correlated nature and long time scales encountered in geochemical problems. The two major bottlenecks in direct simulation of geochemical processes are the need to accurately represent the atomic level forces in the system and the very long time scales that may be encountered (e.g., in chemical reactions). Goal 1: The development of the theory, numerical methods and implementations required to treat complex chemical processes in the interface region. These would include highly scalable density functional (DFT) methods of solutions to electronic structure problems, highly scalable hybrid density functional and LDA+U methods and the development of methods that are capable of treating the highly correlated nature of the electrons in metal oxide materials (DMFT). Goal 2: (i) The exploration of new simulation methods that increase the time scale of dynamical simulations by replacing the usual sequential methods of molecular dynamics by parallel implementation of forward in time integrations. These methods have the potential to utilize information from a less precise physical model simulation to accelerate the performance of complex models (e.g., accelerate the MD performance of high level 1st principle methods). (ii) The development of sampling methods that would support the efficient exploration of complex many-body potential landscapes.

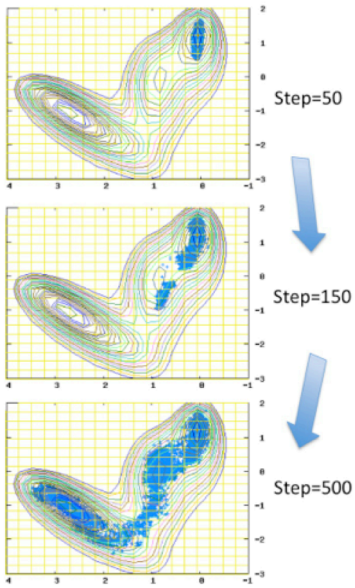


Figure 2. Sampling of the Meuller surface diffusion model. Note that the reaction paths and the three minima have been sampled by 500 steps of our algorithm.

To facilitate the efficient exploration of phase space with first principles simulations we are pursuing a three-pronged approach. (i) We are developing new parallel in time integration methods[2]. These methods utilize information from a less precise physical model to accelerate the performance of complex models. These methods work well across low cost networks and they can be used to expand scaling of simulations with limited scaling. Further, we will investigate new time-parallelization schemes that are particularly designed to resolve the multiple time-scales in a hierarchal fashion as in multigrid solvers. (ii) We are developing sampling methods that support the efficient exploration of complex many-body potential landscapes. Promising classes of techniques that bias the process by replicating walkers making progress and discontinuing walkers that do not are being developed (Fig.2). (iii) We are pursuing dimensional reduction and uncertainty quantification methods, such as principal component analysis/principal orthogonal decomposition and machine-learning techniques to extract reliable order parameters for free energy simulations and rate determination. New "event driven" parallel programming models that efficiently create and destroy terascale/petascale simulations will be used to schedule the collection of simulations resulting from (i)-(iii). These tools will be applied to BES Geochemical related problems (e.g., processes at defected transition/actinide metal oxide surfaces and CO_2 acidity in nanohydration environments).

[2] Eric J. Bylaska, Jonathan Q. Weare, and John H. Weare. "Extending molecular simulation time scales: Parallel in time integrations for high-level quantum chemistry and complex force representations." The Journal of Chemical Physics 139 (2013): 074114.

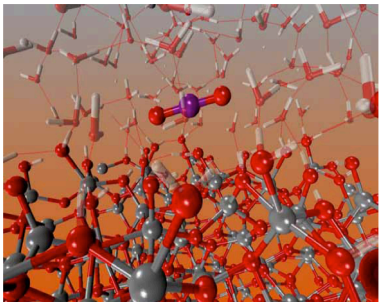
10.2.2.2 Codes and Algorithms

This project contains within it a subtask that emphasizes the development of new computational tools for first-principle, parameter-free simulations of complex geochemical processes with application to metal ion and surface hydration. These new computational capabilities have been added to NWChem, allowing such capabilities to be accessed and used by the wider geochemical and scientific community.

NWChem: Delivering High-Performance Computational Chemistry

NWChem aims to provide its users with computational chemistry tools that are scalable both in their ability to treat large scientific computational chemistry problems efficiently, and in their use of available parallel computing resources from high-performance parallel supercomputers to conventional workstation clusters.

NWChem software can handle



- Biomolecules, nanostructures, and solid-state
- From quantum to classical, and all combinations
- Ground and excited-states
- Gaussian basis functions or plane-waves
- Scaling from one to thousands of processors
- Properties and relativistic effects

NWChem is actively developed by a consortium of developers and maintained by the EMSL (www.emsl.pnl.gov) located at the Pacific Northwest National Laboratory (PNNL) in Washington State. Researchers interested in contributing to NWChem should review the Developers page. The code is distributed as open-source under the terms of the Educational Community License version 2.0 (ECL 2.0).

The current version of NWChem is version 6.3.

The NWChem development strategy is focused on providing new and essential scientific capabilities to its users in the areas of kinetics and dynamics of chemical transformations, chemistry at interfaces and in the condensed phase, and enabling innovative and integrated research at EMSL. At the same time continued development is needed to enable NWChem to effectively utilize architectures of tens of Petaflops and beyond.

10.2.3 HPC Resources Used Today

10.2.3.1 Computational Hours

In 2013, we used 2.2M core hours on the Hopper computer system.

10.2.3.2 Parallelism

A typical simulation on this project will use between 240 and 1,200 cores. Our primary codes that we use are the ab initio molecular dynamics and band structure functionality contained in NWChem. This code has been demonstrated to run to at least 100K cores at NERSC[3].

A very conservative number of cores per job is used because the amount of core hours available to the project is limited.

The project does not currently run high throughput jobs at NERSC. However, we anticipate some parts of the project, e.g., *ab initio* thermodynamics methods used by A. Chaka, will run these types of jobs in the next few years. In addition, we are also running more advanced free energy sampling methods in combination with ab initio molecular dynamics and molecular dynamics that will require multiple instances running at the same time.

Predictive molecular simulations with today's computational molecular methods are beyond strong scaling ("Strong scaling hard?") because as the system size becomes larger the time spanned by the simulation needs to be longer to adequately sample the system.

10.2.3.3 Scratch Data

Our *ab initio* molecular dynamic simulations typically use 1-10 GB per run.

10.2.3.4 Shared Data

The project did not have a NERSC project directory in 2013.

10.2.3.5 Archival Data Storage

Up to a 1 Petabyte of data has been stored in 2013 on the archives located at NERSC and EMSL at PNNL from simulations at NERSC. (NOTE: NERSC repo mp119 had less than 1 TB stored in the NERSC archival storage system in 2013.)

10.2.4 HPC Requirements in 2017

10.2.4.1 Computational Hours Needed

We anticipate that we will need at least a factor of 10 increase in core hours from 2013 to 2017, primarily to run new types of free energy and parallel in time simulations.

10.2.4.2 Parallelism

We anticipate that we will need at least a factor of 10 increase in the number of cores used concurrently in 2017 vs. 2013 (i.e. 2,400 to 12,000 cores). We have already developed algorithms that efficiently scale to 100,000 cores using non-trivial parallelization [3]. In addition, it is expected that future simulations will make extensive use of methods that can be trivially parallelized such as single sweep free energy methods [4] and parallel in time simulations [5],

---

[3] http://www.nwchem-sw.org/index.php/Benchmarks#Parallel_performance_of_Ab_initio_Molecular_Dynamics_using_plane_waves

We anticipate that our *ab initio* molecular dynamics codes will be demonstrated to run on over 500K cores at NERSC in 2017. Fault tolerance issues and machine sizes primarily limit the scaling of the current ab initio molecular dynamics codes. Performance analyses of our current algorithms suggest that they will scale to 1M cores [3].

[3] http://www.nwchem-sw.org/index.php/Benchmarks.

[4] Luca Maragliano, Eric Vanden-Eijnden, "Single-Sweep Methods for Free Energy Calculations", http://arxiv.org/abs/0712.2531, (2007).

[5] Eric J. Bylaska, et al. "Hard scaling challenges for ab initio molecular dynamics capabilities in NWChem: Using 100,000 CPUs per second." *Journal of Physics: Conference Series*. Vol. 180. No. 1. IOP Publishing, 2009; Eric J. Bylaska, et al. "Parallel implementation of G-point pseudopotential plane-wave DFT with exact exchange." *Journal of Computational Chemistry* 32.1 (2011): 54-69.

10.2.4.3 I/O

In 2017 our *ab initio* molecular dynamic simulations typically read and write a modest 1-50 GB per run. At a rate of 1Gb/s this suggests a reasonable bandwidth of <100 GB/s for parallel I/O.

10.2.4.4 Scratch Data

In 2017 our *ab initio* molecular dynamic simulations typically use 1-50 GB per run. The increase in storage from 2013 is the result of running larger simulations.

10.2.4.5 Shared Data

The project currently does not have a NERSC project directory.

10.2.4.6 Archival Data Storage

We will need 10s of petabytes of storage in 2017. The increase in storage from 2013 is the result of running larger simulations. Currently, we store the majority of our data in the archive at EMSL, PNNL. We hope to continue to use local storage at PNNL into 2017, depending on network bandwidth. The majority of this data will be stored at the archive at PNNL. Given the expected increases in computational power in the next five years we estimate that a typical electronic structure simulation in 2017 will use between 2-100 terabytes to store the wavefunctions (e.g., 2,048 orbitals with a 512^3 grid → 2.2 terabytes for O(N^3) algorithm, 16K orbitals with 2048^3 grid → 263 terabytes for O(N^2) algorithm). The major bottleneck limiting the systems sizes in our electronic structure simulations is that the computational cost scales as O(N^3) where N=number of atoms. It is expected in the next five years that the basic algorithms used in our electronic structure algorithms will transform from being O(N^3) to O(N^2)-O(N).

10.2.4.7 Memory Required

We anticipate that our *ab initio* molecular dynamics and band structure codes will need at least 1 Gb per core in 2017. For each simulation this will result in at least 1 Tb of aggregate memory.

10.2.4.8 Emerging Technologies and Programming Models

We have been porting our *ab initio* molecular dynamics and band structure codes as well as other NWChem functionality to use GPUs and MICs. We anticipate that the port to the Knights Corner MIC coprocessor using the offloading model will be finished in 2014. We also anticipate that further developments targeted towards the Knights Landing processors contained in the next NERSC system, Cori, will take continue to 2015-2016.

In addition to these hardware-driven changes to programs, we also starting to use more flexible event driven programming models, such as run-time systems like CometCloud (M. Parashar, Rutgers), that will allow the efficient scheduling of the creation and destruction of large numbers of (terascale/petascale) simulations resulting from advanced free energy and parallel in time simulations. NERSC should be working with their users and computer scientists who are using event driven programming models to make sure that these new classes of simulations will be able to run on NERSC computers.

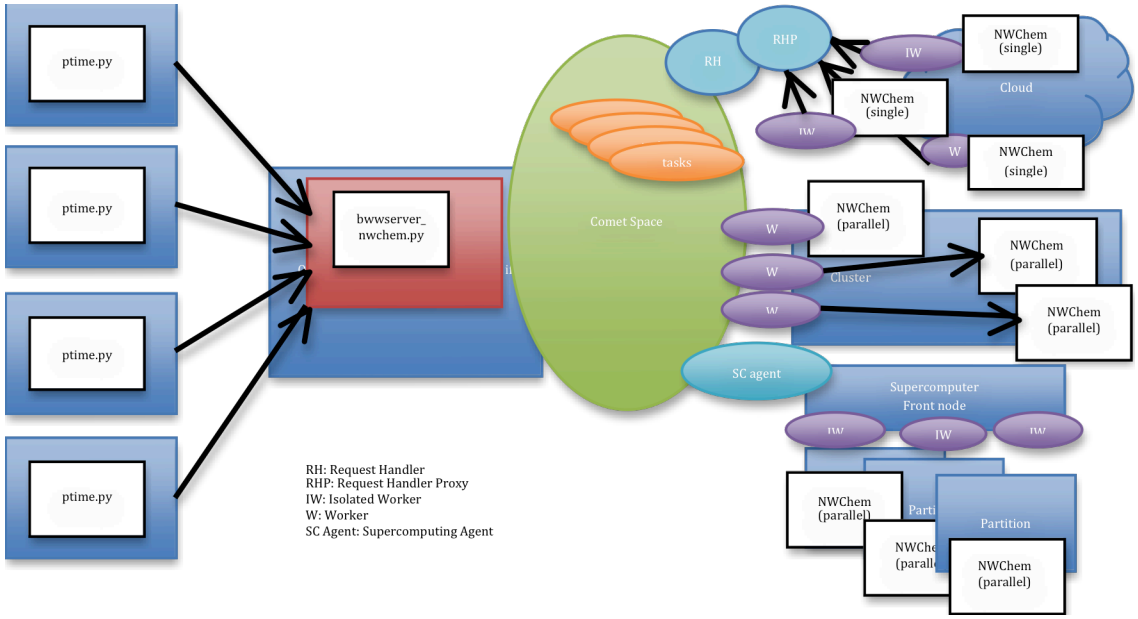


Figure 3. Proposed computational structure of parallel in time simulation on CometCloud running over cloud and HPC computer systems.

10.2.4.9 Software Applications and Tools

NWChem and our group's various molecular dynamics and electronic structure programs are required. In addition, we need C, C++, Fortran, Python compilers and interpreters, BLAS, LAPACK, FFT, and math libraries, as well as MPI, Global Array and possibly UPC programming models. Also, new types of event driven programming models, like CometCloud.

10.2.4.10 HPC Services

We would like NERSC to support complex simulations with multiple instances that may be running across multiple sites (e.g., Figure 3).

10.2.4.11 Data Intensive Needs

The current HPSS system currently meets our needs. Larger quotas on user disk space would be helpful for compiling and maintaining large software installations like NWChem.

10.2.4.12 Requirements Summary

	Used at NERSC in 2013	**Needed at NERSC in 2017**
Computational Hours	2.2 M	22 M
Typical number of cores* used for production runs	240-1,200	2,400-12,000
Maximum number of cores* that can be used for production runs	100K	500K
Data read and written per run	1-10 GB	1-50 GB
Maximum I/O bandwidth	1 GB/sec	100 GB/sec
Percent of runtime for I/O	<1%	<1%
Scratch File System space	< 1TB	<1 TB
Shared filesystem space	0 TB	1 PB
Archival data	449 GB	10 PB
Memory per node	2 GB	2-4 GB
Aggregate memory	1 TB	10 TB

* "Conventional" cores

10.3 Direct Numerical Simulation of Poisson-Nernst-Planck Equation in Charged Clays

Principal Investigator: Carl Steefel (Lawrence Berkeley National Laboratory)
Additional Worksheet Author: David Trebotich (Lawrence Berkeley National Laboratory)
NERSC Repositories: m1516; m1792 (PI: Trebotich via ASCR)

10.3.1 Project Description

10.3.1.1 Overview and Context

Engineered clay barriers have remarkable macro-scale properties such as high swelling pressure (Gonçalvès et al., 2007), very low permeability (Mammar et al., 2001), semi-permeable membrane properties (Malusis et al., 2003), and a strong coupling between geochemical, mechanical, and osmotic properties (Malusis and Shackelford, 2004; Gonçalvès et al., 2007). These properties are thought to arise from the distinct geochemical, transport, and mechanical properties of the interlayer nanopores of swelling clay minerals such as Na-montmorillonite and other smectites (Gonçalvès et al., 2007). This makes them ideal as a backfill for nuclear waste repositories, which is an important reason why the U.S. Department of Energy is interested in their behavior. The need to understand the behavior of clay-rich rock is also important for the problem of shale gas, since shales include a high proportion of clay in their mineralogical makeup.

In compacted smectite-rich media, most of the pore space is located in slit-shaped interlayer nanopores with the width of a few statistical water monolayers (Kozaki et al., 2001; Bourg et al., 2006). The complex microstructure of these clay barriers (Cebula et al., 1979; Melkior et al., 2009) can be approximated with a conceptual model on which all pores are identical slit-shaped pores between parallel negatively-charged smectite surfaces. With this model, the width of the interlayer pores can be derived from the dry bulk density of the smectite, and properties such as the anion exclusion volume or the swelling pressure of the clay barrier can be predicted by solving the Poisson-Boltzmann equation in the space between the parallel, negatively charged clay particles (Gonçalvès et al., 2007; Tachi et al., 2010). Solving the Poisson-Boltzmann equation, however, is not a minor exercise carried out in the context of a general multicomponent framework, in addition to requiring a fine discretization to capture the chemical and electrostatic gradients in the vicinity of the charged clay surfaces. This is why we are also pursuing a mean electrostatic approach, which makes it easier to consider larger length scales for transport (e.g., Tournassat and Appelo, 2011). The Poisson-Boltzmann approach, or even the Mean Electrostatic (or Donnan equilibrium approach) that is based on it, however, potentially provides a more mechanistic treatment of swelling pressure and anion exclusion than is possible with the largely empirical approaches employed by Gens (2010) and Guimaraes et al (2013).

In this work, we focus on developing a conceptual model and applying it to the self-diffusion of water and "hard" acids and bases (alkali metals, chloride). To develop our model, we apply a combination of micro-continuum scale models based on the Poisson-Boltzmann (PB) and Poisson-Nernst-Planck (PNP) equations to elucidate the coupling between EDL phenomena and molecular diffusion in clay nanopores. Specifically, we propose to use the Poisson-Boltzmann equation, and the Mean Electrostatic Model based on it, to develop mechanistic descriptions of clay swelling pressure. Rather than being empirically based on cation exchange capacity (which incorporates both inner sphere and outer sphere sorption,

as well as Electrical Double Layer (EDL ions), this will consider the overlap of electrical double layers in the context of Stern layer (inner sphere) sorption.

Year	Problem dimen-sion	Domain size	Grid-points	Spatial resolution	Problem time scale	HPC resources	Problem DOF	Data storage per plot file
2013	2D	1m	2B	O(1 μm)	O(hours)	50M hours	25 variables	1TB
	3D	1cm	2B	O(1 μm)	O(minutes)	100M hours	25 variables	1TB
2017	2D	10cm	20B	O(1 nm)	O(hours)	500M hours	5 variables	10TB
	3D	1cm	20B	O(1 nm)	O(minutes)	1B hours	5 variables	10TB

10.3.1.2 Scientific Objectives for 2017

The primary objective of the nano-continuum modelling work will be to develop a Poisson-Boltzmann description of the electrical double layer (EDL) within the context of a multicomponent reactive transport code that accounts for Nernst-Planck (electrochemical migration) effects. The general multicomponent software framework of the Poisson-Nernst-Planck (PNL) equations will be applied to the problem of compacted clays (used as backfill for geological nuclear waste repositories) and clay-rich rocks like shale. We are not aware of an implementation of the Poisson-Boltzmann equation in the context of a true multicomponent reaction-diffusion solver, with the possible exception of the work by Leroy et al (2006), although this development resulted in largely hard coded software. In addition, they did not include a Nernst-Planck treatment of ion diffusion. The Poisson-Nernst-Planck (PNP) equations will be calibrated based on molecular modelling and laboratory diffusion experimental results. The PNP modelling results will be compared with results from the Mean Electrostatic model.

Our plan is to bring the full capabilities of Chombo to bear on the problem, coupling it to CrunchEDL to resolve both solute transport and electrostatic effects at the nanoscale. This requires higher resolution than we have considered to date (nanometers), since the electrical double layer (EDL) needs to be resolved near the charged clay surfaces. A typical EDL thickness is on the order of nanometers.

10.3.2 Computational Strategies (now and in 2017)

10.3.2.1 Approach

Our current approach is two fold:

1) Micro-continuum modeling without the use of HPC to capture the diffusion of ions through compacted clays (CrunchEDL). CrunchEDL has capabilities for both a Mean Electrostatic (or Donnan Equilibrium) approach, and we have developed preliminary capabilities for solving the Poisson-Boltzmann equation.

2) Pore scale modeling without electrostatic effects based on a combination of the Chombo Computational Fluid Dynamics (CFD), Embedded Boundary Methods (EB), and Adaptive Mesh Refinement (AMR) that are coupled to the general purpose geochemical simulator CrunchFlow. Chombo-Crunch, in contrast, currently does not handle the electrostatic effects, but solves the Navier-Stokes equation together with molecular diffusion and general (bio)geochemical reactions, resolving mineral-water interfaces at high resolution.
3) In addition, we have a collaboration with an environmental firm AMPHOS to develop a coupled model for flow, transport, and electrostatic effects in charged, compacted clays based on integration of the codes CrunchEDL and Comsol. This CrunchEDL-Comsol capability, however, will not scale past about 150 cores, but is useful as a prototype of what we expect to do at the HPC level with the Chombo-CrunchEDL coupling described below.

Our plan for 2017 is to bring the full capabilities of Chombo to bear on the problem, coupling it to CrunchEDL to resolve both solute transport and electrostatic effects at the nanoscale. This means solving the combination of the Poisson-Boltzmann equation and the Nernst-Planck equation for lateral migration of ions through the clay. This requires higher resolution than we have considered to date, since the electrical double layer (EDL) needs to be resolved near the charged clay surfaces. We will then have a micro/nano-continuum model that includes the possibility of both Navier-Stokes or Stokes flow and electrochemical and ion transport through compacted clay. This will allow us to consider osmotically-driven flow and swelling pressure in clays, as well as electrostatically modified molecular diffusion.

10.3.2.2 Codes and Algorithms

- Chombo flow and transport solvers using adaptive, finite volume methods to treat complex clay geometries.
- New algorithm for PNP equation based on high performance scalable elliptic solvers in Chombo coupled to flow and transport.
- Coupled to CrunchEDL via operator splitting to include the Coulombic (electrostatic) effects on transport and reaction in the compacted clays.

10.3.3 HPC Resources Used Today

10.3.3.1 Computational Hours

We used 103M at NERSC in 2013 using three repositories: m1792, which is an ALCC-allocated ASCR project; m1411, an ERCAP-allocated ASCR project; and m1516, an ASCR project allocated from the NERSC director's reserve. We will also use about 800M at ALCF (Mira) as part of an ALCC award.

10.3.3.2 Parallelism

Today, our runs use 48,152-96,304 cores. The code has been run on up to 131,072 cores. Weak scaling is most important for our project. Our sweet spot for load balancing is 1 grid box per core. In 3D a grid box is 32^3 grid cells; in 2D, it is 256^2. In order to perform strong scaling on a transport problem in a realistic medium we would lose the sweet spot

optimization, both in memory and performance. As it is, we scale very efficiently up to 128K cores in weak scaling studies. High-throughput computing is of no importance.

10.3.3.3 Scratch Data

We require about 80TB of scratch storage today.

10.3.3.4 Shared Data

m1792 has a project directory at NERSC used for file sharing between computational mathematicians and applications scientists.

10.3.3.5 Archival Data Storage

The three repos combined used 744 TB of storage in the NERSC HPSS system in 2013.

10.3.4 HPC Requirements in 2017

10.3.4.1 Computational Hours Needed

We currently need about 100 M hours on a machine like Edison. In 2017, we could use 1B hours on a new architecture machine (Trinity, NERSC 8?). We have an INCITE proposal currently under review for 2014 and we currently have an ALCC award. The primary factor driving our increased requirement is the need to use much higher resolution (factor of 1000, nanoscale) for solving PNP equations in charged clays compared to current pore scale simulations at microscale resolution.

10.3.4.2 Parallelism

We will probably use about 131,072 cores in production and we estimate being able to scale to about 393,216.

We will probably need to have two jobs running concurrently —one for a benchmark blank scaling run, one for a science production run.

10.3.4.3 I/O

A full series of 36-hour production runs dumping checkpoint and plot files in hdf5 format every 100 timesteps will create about 1.5PB. Typically we run for 100,000 time steps for steady state solutions at the pore scale. And each plot file is about 1TB, checkpoint is about .5 TB.

Our I/O generally consists of 1-TB plot files and 0.5-TB checkpoint files. We would be will to devote no more than about 2% of the total runtime to I/O (and it generally is about that level now).

10.3.4.4 Scratch Data

We will need 100TB of scratch space in 2017. The primary driver for this is our need to do runs using higher resolution.

10.3.4.5 Shared Data

We will also need 100TB of project space in 2017. The primary driver for this is, again, our need to do runs using higher resolution.

10.3.4.6 Archival Data Storage

We estimate having to archive 10 PB of data. The primary driver for this is, again, our need to do runs using higher resolution.

10.3.4.7 Memory Required

We require 2GB per node.

10.3.4.8 Emerging Technologies and Programming Models

We are not currently thread safe. We will port our code to the new thread safe Chombo4 libraries.

10.3.4.9 Software Applications and Tools

We need PETSC and VisIt (or something comparable).

10.3.4.10 HPC Services

We will need consulting or account support and support for data analytics and visualization.

10.3.4.11 Time to Solution and Throughput

As mentioned above, we will need a large quota for scratch space in order to have an efficient workflow for doing post-processing of a run, balanced with moving to archival storage.

10.3.4.12 Data Intensive Needs

We are generally satisfied with NERSC's HPSS system but would like to see a speedup in transfer rates. It currently takes several hours to move 1TB from scratch to HPSS using hsi.

We already have a data management plan in place now that includes archival storage.

10.3.4.13 Additional Comments

The most important factors for us are (wall)clock time and memory bandwidth. For example, Edison is 2.5 times faster than Hopper for our code. We are memory bandwidth limited.

10.3.4.14 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours	96 M	1,000 M
Typical number of cores* used for production runs	49,152-96,304	131,072-393,216
Maximum number of cores* that can be used for production runs	131,072	393,216
Data read and written per run	1,000 TB	10,000 TB
Percent of runtime for I/O	<2%	<2%
Scratch File System space	80 TB	100 TB
Shared filesystem space	2 TB	100 TB
Archival data	744 TB	10,000 TB
Memory per node	2 GB	2 GB

* "Conventional" cores

10.4 Imaging and Calibration of Mantle Structure at Global and Regional Scales Using Full-Waveform Seismic Tomography

Principal Investigator: Barbara Romanowicz (UC Berkeley)
Worksheet Author: Scott French (UC Berkeley)
NERSC Repository: m554

10.4.1 Summary and Scientific Objectives

Our current research is focused on imaging the interior of the earth at global and regional scales using seismic waves. We aim to improve understanding of the interior structure of the earth's mantle and the dynamic processes that drive a broad range of observed phenomena, including plate tectonics and hotspot volcanism. To achieve this goal, we employ advanced methods that combine: (a) direct inversion of full seismic waveforms (seismograms), allowing us to take advantage of the rich information content thereof; and (b) numerical simulations of the seismic wavefield that accurately reflect the physics of wave propagation in the real earth.

Full-waveform seismic inversion of this type is both computation and data intensive, and would not be feasible without HPC and coupled data storage resources. While these numerical simulation techniques are still quite new to global seismology, we are already using millions of CPU core hours per year for production runs and expect this to only increase. Further, we note that having a collocated high-performance scratch file system available is very important, particularly for accommodating the large quantities of intermediate data produced when processing simulation output in combination with the observed seismograms (which together drive the underlying seismic-imaging optimization problem).

10.4.2 Scientific Objectives for 2017

In the broadest sense, we do not expect our underlying scientific goals for 2017 to vary sharply from those given in Section 9.4.1 – imaging of the earth's interior using highly accurate numerical simulations of the seismic wavefield. However, we expect that the resolution of our imaging will increase considerably, coupled with more computationally heavy modeling of higher-frequency seismic waves, allowing us to make more detailed inferences regarding the interior dynamics of the earth. Indeed, our current research not only reveals details of mantle structure never before seen, but also lays the groundwork for future high-resolution imaging by providing a starting model that is well-constrained at longer wavelengths.

10.4.3 Computational Strategies

10.4.3.1 Approach

Our work is focused on imaging the interior structure of the earth using full seismic waveforms. This is an inverse problem, in which we seek a model of the earth that accurately predicts the seismograms observed for real earthquakes. Our approach to solving this problem involves simulating the propagation of seismic waves through the earth for many (hundreds) of individual earthquakes using a high-order ("spectral") finite

element method. The results of these simulations are used to update our estimate of the earth's interior structure, and the process is repeated until convergence.

10.4.3.2 Codes and Algorithms

Our primary codes are spectral finite element method (SEM) solvers, employing an explicit time stepping scheme and a matrix-free formulation, which are specialized for either global or regional (continent) scale simulations, respectively (both are parallelized with MPI). Other codes, which use a much smaller fraction of our total allocation, are responsible for processing and assimilation of simulation output, namely: (a) calculation of partial derivatives for updating our seismic model using normal-mode coupling theory (MPI+OpenMP); and (b) optimization of the model, by combining our simulation output, the observed waveform data, and the partial derivatives that relate the two (MPI+ScaLAPACK for large-scale dense linear algebra).

10.4.4 HPC Resources Used Today

10.4.4.1 Computational Hours

We used 3 million hours at NERSC in 2013. NERSC is currently the only site where we have significant resources available.

10.4.4.2 Parallelism

The finest unit of simulation for our SEM computations is the individual earthquake, which typically requires only 150-300 compute cores. However, we have many earthquakes to simulate (hundreds), and because there are no data dependencies between earthquake simulations we routinely merge these computations into concurrent 2-3k core production runs. Other codes described in Section 9.4.3.2 above need not be run on a per-earthquake basis (instead processing simulation output from the entire set of earthquakes at once) and typically use roughly 500 compute cores.

For the problem sizes relevant to our current research, which dictates the discretization of our finite-element mesh, our SEM simulations would not scale well to larger core counts than those listed above. Indeed, measurements of parallel efficiency, together with turnaround time, guided those specific choices of core counts. However, as noted above, these simulations can be aggregated into much larger production runs.

We do not use what would typically be termed a high-throughput computing mode (though we do merge simulations, 10-20 depending on core count, into single production runs as noted above).

For our project, the problem size (the finite-element mesh) is fixed, or is changed only very rarely. Thus, strong scaling is more important.

10.4.4.3 Scratch Data

In general, we are able to comfortably stay below the current 5TB limit on Hopper local scratch. In the final phase of each inversion iteration, where simulation output is consumed and the model is updated, we can episodically approach 3-4TB of scratch utilization. We approach similar utilization levels during experimental SEM simulations that require heavier use of checkpointing, though these are rare. In addition, we occasionally use the

global scratch file system to make simulation output available to the data transfer nodes, but typically only for sets of files on the order of 100GB in size.

10.4.4.4 Shared Data

We currently use our project directory only lightly, storing <100 GB of shared data (preprocessed seismic waveforms).

10.4.4.5 Archival Data Storage

We have approximately 100GB stored on the NERSC HPSS data archive now.

10.4.5 HPC Requirements in 2017

10.4.5.1 Computational Hours Needed

Given compute costs evaluated on current "conventional" machines, we anticipate a CY 2017 allocation request of at least 25M hours. This value represents a conservative lower bound on the number of hours that will be required in order to meet our scientific goals and remain productive in 2017 (see below). We do not currently plan to receive significant allocations elsewhere.

The primary factor behind the growth in hours is the need to simulate higher-frequency seismic waves in order to attain better resolution. This, in turn, necessitates a refined finite-element mesh, which drives up the cost of each earthquake simulation (as approximately frequency^3 for our particular application and configuration). Further, we anticipate that by 2017 we will have incorporated adjoint-state computations into our waveform inversion approach, which will improve upon the waveform gradients currently approximated using mode-coupling theory. This approach introduces an additional factor-of-two increase in the number of simulations required and is currently under active development. Thus, assuming 3M-hour present-day resource needs, ≥1.5x increase in seismic wave frequency by 2017, an overall 2x cost increase when using adjoint methods, and an additional 20% overhead due to both simulation-output assimilation calculations and validation runs on future platforms, we arrive at 25M hours minimum allocation size.

10.4.5.2 Parallelism

We anticipate that our primary codes (the SEM solvers) will typically use 300-500 compute cores per simulation. Our simulation output assimilation codes will likely use somewhere in the 3,000 – 10,000 core range, but this is not well constrained at the moment (it depends on future choices of parameterization for our seismic model, which is subject to change).

We anticipate the above values for the SEM to again be near the optimal tradeoff between parallel efficiency and run duration by analogy with the current per-core workload (though this weak-scaling argument clearly ignores changes in scaling behavior on future platforms).

We will typically have more than one job running concurrently because our simulations will remain "discretized" by earthquake, with hundreds of simulations to perform. However, there are no dependencies between these runs, so they can be scheduled individually or can trivially be merged into larger production runs, depending upon available resources and I/O load. The latter has not currently been found to be a problem, though future simulations

may require heavier checkpointing – particularly the adjoint-state computations referred to in Section 5.1, which accumulate checkpoint snapshots as the run progresses for subsequent reconstruction of the wavefield in reversed time.

We do not anticipate adopting a high-throughput computing mode.

10.4.5.3 I/O

We anticipate that checkpointing will occupy on the order of 40-50 GB for "typical" simulations (which use checkpoints only for crash recovery) and at least 500 GB for adjoint-state simulations (which require time histories of checkpoints). Specialized (rare) adjoint simulations may require up to 1 TB of scratch storage for these purposes. In addition, we expect the computations that process and assimilate simulation output to produce on the order of 10-20 TB of intermediate data (discretized in files approximately 0.5-1 TB in size).

Aggregate I/O bandwidth for our current large production runs on Hopper (multiple SEM simulations occupying up to 3,000 CPU cores) typically peaks near 20 GB/s and is associated with distributed read of files produced by our finite-element mesher. We anticipate these files to grow in size by a least 6x in simulations for CY 2017. Assuming the same total number of CPU cores per run (an underestimate) and holding the current read-time fixed leads to a lower-bound estimate of 120 GB/s. For adjoint-state SEM simulations producing or consuming large volumes of checkpoint data, we know that access will occur in small chunks, which we expect to be similar in size to the mesh files, suggesting similar I/O requirements. Assimilation of simulation output should produces files of 0.5-1TB in size, but occurs so rarely that it does not factor significantly into our bandwidth requirements (i.e. we are willing to trade off with read/write time).

For the majority of our SEM simulations, we do not anticipate I/O time to be a significant fraction of our total run times, since they are characterized by infrequent, distributed read or write of 10's of GB of data. Thus, anything larger than a few percent, and certainly over 5%, would seem excessively large, with the exception of our simulation assimilation codes (for these, our expectations of percentage runtime will have to be relaxed).

10.4.5.4 Scratch Data

Combining the estimated lower bound on checkpoint footprint for adjoint-state simulations of 500 GB (Section 5.3), with a reasonable estimate of 20 simulations in some stage of completion at once (possibly aggregated into fewer runs), yields an estimate of 10 TB+ of actively occupied scratch space or 12 TB+ with a modest 20% overdesign. Solution output (seismic waveforms) will likely remain quite small (<1GB per simulation). Intermediate data produced in assimilating simulation output is anticipated to peak on the order of 10-20 TB (Section 5.3) at which point the checkpoint files will already have been removed. This estimate does not include the specialized large-I/O SEM simulations (1 TB of checkpoints) referred to above, as it is not clear at this time how to often these will occur (though likely not often).

Growth in scratch space is primarily due to the checkpoint files produced by our SEM runs and the intermediate data produced during solution assimilation, both of which grow as our

mesh is refined at higher frequencies. The former also grows with heavier use of checkpointing in the adjoint-state SEM simulations.

10.4.5.5 Shared Data

We have a small data store in the NERSC global file system, /project. We do not anticipate significantly expanding usage there.

10.4.5.6 Archival Data Storage

We do not anticipate our on-site archival storage needs to grow significantly – likely only to a 500 GB or less.

10.4.5.7 Memory Required

We anticipate our SEM solvers to continue to require on the order of 1-2GB per process/core. Thus, any per-node memory configuration satisfying this requirement should be acceptable (albeit, an assessment limited to HPC architectures similar to present-day "conventional" ones). Also, though it represents a significantly smaller fraction of our allocation usage, the MPI/OpenMP parallelized partial-derivative estimation and solution assimilation code (which assembles large, dense matrices for later factorization) would benefit from having from more memory per node than is currently possible on Hopper – on the order of 100 GB or more (potentially on designated "large-memory" nodes).

10.4.5.8 Emerging Technologies and Programming Models

Our codes are not currently ready for advanced architectures. There has been some success in the seismic-modeling community on porting these types of matrix-free finite element calculations to heterogeneous environments – specifically those hosting GPUs (currently in development for ORNL Titan, in particular). We are currently investigating how the structure of our particular SEM implementation would lend itself to such an effort (perhaps merging with the community code when GPU support is fully integrated in the latter), and assessing how community experiences in porting to GPUs will also inform efforts to support other architectures – e.g. Intel MICs). However, guidance from NERSC as to what hardware technologies to anticipate and, associated with that, what level of generality will be supported at the compiler level (e.g. CUDA Fortran, OpenACC, etc.), would be valuable.

10.4.5.9 Software Applications and Tools

Our primary SEM codes are written in Fortran 90, while the others are in a mix of Fortran 90, C (ANSI and C99), and C++. This is unlikely to change in the next 4 years. Our codes are largely parallelized with MPI and/or OpenMP, but also more recently incorporate UPC-like PGAS extensions to C++ (collectively known as UPC++, with similar dependencies to UPC – e.g. GASNet). We expect this to change to some degree as NERSC adopts new technologies for heterogeneous computing (hopefully exposed at a more abstract, portable level – e.g. directive-based as opposed to writing pure CUDA). For our languages of choice, we have the most experience with the PGI and GNU compilers and would like to see them remain available on future NERSC resources. Our library needs are rather narrow and we again do not expect them to change: an optimized BLAS and FFT, as well as ScaLAPACK. Finally, we anticipate that we will continue to perform model analysis offsite (though processing and assimilation of solution output will remain onsite).

10.4.5.10 HPC Services

Although it is difficult to say definitively, we anticipate that assistance from NERSC consulting will be valuable in further tuning our codes for many-core or heterogeneous computing environments (assuming that we will already have a working production implementation thereof in 4 years).

10.4.5.11 Time to Solution and Throughput

We do not anticipate any special needs in this regard. However, as is the case now, our workload will be episodic in nature – we run many (hundreds) of simulations, assimilate the results by incrementally updating our seismic model, and repeat until convergence. These iterations can be delayed by extensive offsite analysis of the model, as well as efforts to expand our dataset to improve model resolution. This workload runs somewhat contrary to the mode of continuous simulation often encouraged by the allocation reduction schedule (though we have not recently incurred such reductions).

10.4.5.12 Data Intensive Needs

Given the storage needs estimated above, it would seem that our current workflow would scale to the problem sizes we anticipate for CY 2017 with fairly modest (2-4x) scratch quota increases.

We are satisfied with NERSC's HPSS system, though we are not heavy users of HPSS.

Because (a) both our data and the results of our simulations (seismic waveforms) occupy so little space, and (b) we are a single research group performing the majority of our analyses offsite, we do not have a formal data management plan at NERSC. That said, we do keep careful records of data quality and provenance, and are investigating further formalizing these efforts. At present, we do store simulation output for archival purposes in HPSS.

10.4.5.13 Additional Comments

Two concerns that we would like to draw attention to, as noted above, are: (a) the desire for guidance in planning which technologies for many-core / heterogeneous computing to have in mind as we adapt our codes; and (b) ensuring that the irregular, episodic nature of our workflow is understood from an allocation management perspective.

For our use case, in which we are managing many production simulations in the low-to-mid hundreds of CPU cores in size – thousands of simulations per year by 2017 – one of the most import features of an HPC system is the ability to predict and reason about wall-clock run duration. Thus, we are particularly interested in features that mitigate nondeterministic variation in run duration, including the availability of a high-performance scratch file system providing uniform high I/O bandwidth and, to a lesser extent for our use case, a scheduler that considers (interconnect) locality when mapping tasks to compute nodes. The former is particularly important for our adjoint-state SEM simulations, while the latter we expect not to be a significant factor on future systems (an expectation supported by recent benchmarks reported for NERSC Edison). Our run-management abilities would be further aided by functionality for monitoring node health, memory availability, I/O congestion, etc. at runtime, possibly allowing our runs to save state and cleanly shut down in unexpected circumstances. NERSC does an excellent job providing a uniform and high-availability environment for computations, but with the large number of runs that we anticipate for 2017, we foresee runtime monitoring becoming a useful feature if it is available.

Aside from maintaining highly reliable and available HPC systems, one of the most useful services provided by NERSC is their extensive documentation of best practices for a wide range of topics (e.g., compilation and optimization, runtime tuning, IO considerations, etc.). Further, the inferences drawn in these documents are often backed up with empirical benchmarks, increasing their value for the specific systems considered.

10.4.5.14 Requirements Summary

	Used at NERSC in 2013	**Needed at NERSC in 2017**
Computational Hours	3 M	25 M
Typical number of cores* used for production runs	SEM: 150-300 Other: 500	SEM: 300-500 Other: 3K-10K
Maximum number of cores* that can be used for production runs	Same (though can be aggregated)	Same (though can be aggregated)
Data read and written per run	SEM: 30 GB Other: 100 GB	SEM: 0.5 TB Other: 0.5-1 TB
Maximum I/O bandwidth	20 GB/sec (aggregate)	120 GB/s (approx.)
Percent of runtime for I/O	<<5%	<<5%
Scratch File System space	~0.1 TB	~0.5 TB
Archival Storage	81 GB	500 GB
Shared filesystem space	49 GB	100 GB

11 Scientific User Facility Case Studies

11.1 Introduction

NERSC is playing an ever-increasing role in the science produced by users of the BES user facilities that offer unique and powerful technical tools for a wide variety of 21st-century energy-related scientific disciplines. NERSC involvement in research associated with user facilities comes about in two ways. First, the volume of data being transferred to NERSC has increased dramatically, as experimental facilities are inundated with large quantities of scientific data. Over the past few years, more data has been transferred to NERSC than away from NERSC – an unprecedented paradigm shift for a supercomputing center. Scientists expect rapid and simplified access to the data and NERSC computational scientists, along with ESnet, have helped solve challenges in extreme data organization, distribution, long-term storage and real-time computational analysis. Second, there has been increasing close interplay between theory and experiment, as scientists use simulation to rapidly help understand results obtained at X-ray and neutron sources.

The following case study is characteristic of the needs of large BES light and neutron source user facilities. NERSC is also presently engaged in test collaborations with beamline scientists at the Stanford Synchrotron Radiation Lightsource (SSRL), a part of the SLAC National Accelerator Laboratory (SLAC); the Linac Coherent Light Source (LCLS), which is also located at SLAC; the Advanced Photon Source (APS) at Argonne National Laboratory; and the National Synchrotron Light Source (NSLS) at Brookhaven National Laboratory, which is one of the most prolific scientific user facilities in the world.

11.2 Advanced Light Source

Principal Investigator: Michael Banda (Lawrence Berkeley National Laboratory)
NERSC Repository: ALS

11.2.1 Project Description

11.2.1.1 Overview and Context

The Advanced Light Source (ALS) is a national user facility open to scientists from academic, industrial, and government laboratories. The ALS is a third-generation synchrotron radiation source optimized for high brightness at soft x-ray and ultraviolet photon energies using undulator and bend magnet sources. It also provides outstanding performance in the hard x-ray region, using wiggler and superconducting bend-magnet sources, to serve the needs for complementary tools at a single facility and the user community needs for capacity. There are 37 user beam lines at ALS that support about 2,000 users per year who study materials for basic research, energy sciences and drug design. Modern experiments and sources increasingly require very high data rate detector and data acquisition systems. These systems can rapidly produce data files too large for users to copy onto portable storage devices. Further, those files will be too large to move by conventional methods to the users' home institutions.

This facility allocation is intended to support the data and computational needs of a diverse user community. It does not have a singular scientific focus. This ALS allocation will explore high performance data movement storage and analysis for the ALS users. Such a scientific data and analysis portal will be applicable to other photon and neutron user facilities. In conjunction, simulations and real time data analysis to inform beam and end station parameters will be explored. Finally, the ALS allocation will produce advanced analysis applications that take advantage of the high performance computational and storage resources at NERSC. All of the tools developed at ALS/NERSC will be available to the larger light source community.

11.2.1.2 Scientific Objectives for 2017

ALS Users are asking for faster detectors, but are not prepared for the consequences because data rates and volumes overwhelm them. Most users do not have the background to use high performance computers. The ALS/NERSC data portal project allows users to take advantage of high performance computers to overcome their data intensive challenges.

Scientific challenges addressed by light sources are diverse. The science questions come from a wide variety of fields, including biology, material science, physics, earth science and archeology. While the science challenges are diverse the techniques are connected by common data themes (Real Space, Reciprocal Space, and Spectroscopy). Facility users are routinely able to generate tens of thousands to hundreds of thousands of images in a few days of running. Using current analysis techniques, software, and resources only a small percentage of usable data are analyzed.

Recent improvements in detector resolution and speed and in source luminosity are yielding unprecedented data rates at national light sources supported by the Scientific User Facilities Division of BES. Over 2 PB per year are expected from some facilities in 2014. Those data rates already exceed the capabilities of data analysis approaches and computing

resources utilized in the past, and will continue to outpace Moore's law scaling for the foreseeable future. The consensus within light source scientific user communities is that scientific insight and discovery at BES facilities are now being limited by computational and computing capabilities much more than by detector or accelerator technology.

The scientific data portal project at ALS is designed to develop a suite of tools to provide ALS users - to leading edge data management of scale: data analysis and simulation tools that are beyond the typical resources of a facility user. This effort involves collaboration at Berkeley Lab with NERSC, ESnet, and computational scientists. The portal project has achieved a significant milestone. It has demonstrated the ability to seamlessly move large data sets from one beamline to the NERSC computing environment where they can be stored and analyzed by users. The ALS does not have the local resources to handle data of scale. When complete, the automated portal will allow ALS users to manage and share their experimental and simulation data of scale, analyze those experimental data, and view real time results during their ALS beam times. By 2017 we expect that several high data rate beamlines will be interacting with NERSC. We also expect to pilot the ability of NERSC to extend the science portal to beamlines from other BES user facilities.

This collaborative effort of Office of Science resources (BES and ASCR) will revolutionize the user experience at BES facilities and begin to break down the barrier to scientific discovery imposed by the challenge of large data sets.

11.2.2 Computational Strategies (now and in 2017)

11.2.2.1 Approach

The problems addressed by modern day light sources are extremely diverse. The science questions range from biology, material science, physics to earth science and archeology. While the science challenges are diverse the techniques are connected by common themes.

Real Space: X-ray imaging beamlines have been generating TBs of raw image data per month, and with the development of new and faster cameras data rates will double or triple in the near future. As the data rate and volume have increased, the need for on-site analysis has increase. The imaging data analysis involves reconstruction of large 3D images, segmentation of the images into sub-regions, discrimination of multiphase solid materials, identification of microstructures, calculation of statistical correlation functions (e.g. surface, pore-size) and extraction of channel networks. As part of the analysis, visual representations of structures are often mandatory. For example, during modeling of structures, a theoretical optimum may not lead to the optimum problem solution, and scientific visualization tools can be indispensable in such cases. The amount of data demands both automation and re-factorization of algorithms into software that can use multiple cores, several nodes to solve a problem that we can characterize as a high performance analytics (data centric). Issues of disk I/O, efficient threading and distributed communications will play a major role on scaling algorithms to provide the necessary tools.

Reciprocal Space: Just as in imaging beamline, scattering beamlines have seen a revolution with respect to flux and detectors. High brilliance beams, efficient x-ray focusing optics and fast 2D area detector technology allow the collection of thousands of diffraction patterns occupying terabytes of disk space. The need for real time analysis also stems from the need to modify experiments on the fly when the next action depends on the outcome of the

previous scan. Additional computational needs for the light sources involve the integration of modeling and simulation as tools accessible to the users in real time. Clever visualization tools are also needed for displaying multidimensional data in a practical way. Techniques such as coherent diffractive imaging, nano-crystallography and ptychography under development at light sources are also the fruit of advances in reconstruction techniques. In these techniques, the need for algorithms capable of solving large scale ill-conditioned, underdetermined, noisy inverse problems has never been so clear. The vast majority of data in diffractive imaging is almost never looked at. Reconstructions fail, often.

Spectroscopy: The data rates for spectroscopy are traditionally not as high as the previous described techniques, but just as before it relies very heavy on usually very complicated analysis algorithms and simulations.

Computational Tools: In each of these light source data themes, domain scientists are routinely able to generate 10,000's to 100,000's of images in a few days of running. Such data volumes cannot be analyzed individually, but rather must rely upon automated methods that translate Materials Science Descriptions to input for modeling and simulation, and that quantitatively compare output of simulation with beam line data. To maximize both the functionality and robustness of these kinds of end-to-end analysis systems, a community-wide, open-source project must be initiated, and nurtured at the agency level (i.e., DOE-BES). These same kind of large scale, automated, and customized systems have been developed and deployed for other science communities. Although none are directly adoptable by BES light source scientists, many principles, approaches, and lessons learned are directly applicable. These tools must possess the following essential features in order for them to be widely adopted in the materials science community.

1. Ease-of-use and extensibility: These features both ensure widespread adoption within a scientific community and maximize the reusability of software components developed by research team by others. As an example, a graphical modeling interface similar to those used by solids modeling programs would provide researchers a natural method of describing the microscopic structure of material samples, and a common format for input to simulations.

2. Deployment of advanced algorithms using state-of-the-art computer hardware: In order to develop the fastest algorithms and robust codes, we need to exploit and leverage the resources provided by the ASCR office, including the expertise in Applied Math, Computer Science and the High Performance Facilities, such as NERSC. In particular, parallelization of these algorithms on multiple CPUs, graphical processor units (GPUs), and hybrid CPU/GPU multicore architectures will dramatically decrease the analysis time by more than several orders of magnitude while simultaneously permitting larger data sets to be treated.

3. Quantitative, interactive visualization: Visualization tools which allow quantitative comparison of simulation and experiment, including whole-image comparison or feature extraction, will aid both large-scale processing, and improve the quantity, quality, and reproducibility of scientific results.

4. Leveraging of advanced computer technologies: As enabling computer technologies (such as data I/O and formats, ontologies, FFT libraries, etc.) and architectures (such as GPUs, multi-core, or heterogeneous architectures) evolve and improve, a common

framework must allow for graceful evolution to accommodate and take advantage of the latest improvements while insulating users from the underlying details. This provides two advantages: The perturbative effects of such changes to scientists' research are minimized and the advantages are more quickly and widely available.

11.2.2.2 Codes and Algorithms

Code Name: ***Fiji***

Description: Tomographic reconstruction
Machines: Edison, Hopper, Carver
Languages: C, C++, Java, Python, shell script
Libraries: HDF
Performance Limits: I/O-disk speed input

Code Name: ***GridRec***

Description: Tomographic reconstruction using gridding method
Numerical Techniques: Spectral Methods
Machines: Hopper, Carver
Languages: C, C++
Libraries: FFTW, HDF

Code Name: ***HipGISAXS***

Description: A massively-parallel, high-performance GISAXS simulation code. Includes optimization algorithms and reverse Monte Carlo algorithm to solve the inverse problems (structure fitting).
Numerical Techniques: Spectral Methods, Structured Grids
Machines: Edison, Hopper, Dirac
Planned Processors: 1- 65,536+
Languages C++, Open MP

Code Name: ***ImageRec***

Description: Tomographic reconstruction using filtered back projection method
Numerical Techniques: Spectral Methods
Machines: Hopper, Carver
Planned Processors: 2-63
Languages: C language
Libraries: FFTW, HDF

Code Name: ***MBIR***

Description: Model-Based Image Reconstruction, First a model is developed for image formation in tomography along with a prior model to formulate the tomographic reconstruction as a MAP estimation problem. The formulation also accounts for certain missing measurements like the transmission measurement, offsets and noise variance, treating them as nuisance parameters in the MAP estimation framework. We adapt the Iterative Coordinate Descent (ICD) algorithm to our application to develop an efficient method to minimize the corresponding MAP cost function. Reconstructions of simulated as well as experimental data sets, show results that are superior to FBP and SIRT reconstructions, significantly suppressing artifacts and enhancing contrast.

Numerical Techniques: Sparse LE
Machines: Hopper, Carver
Planned Processors: 2-63,64-511
Languages: C, C++
Libraries: FFTW, HDF,MPI

Code Name: ***QuantCT***
Description: Semi-automatic image filtering and segmentation
Machines: Hopper, Carver
Planned Processors: 1 (serial), 2-63
Languages: C, Java

Code Name; ***XMAScluster***
Description: Processing of synchrotron Laue x-ray micro-diffraction for grain orientation and strain/stress mapping.
Planned Processors: 512-4095
Number Serial Jobs: 1000
Languages: Fortran90
Libraries: LAPACK, custom

11.2.3 HPC Resources Used Today

11.2.3.1 Computational Hours

In 2012, the ALS Repo had 1,000,000 hours awarded but ended up using 2,00,000 hours. This was the first year of the ALS facility allocation. In 2013 the ALS facility allocation was 2,000,000 and it used over 4,000,000 hours. These early numbers do not well represent a production baseline. There are no other HPC compute or storage resources available to ALS.

11.2.3.2 Data and I/O

Scratch (temporary) space: ALS current usage of Scratch space is minimal, and solely due to individuals' usage.

Permanent (can be shared, NERSC Global Filesystem /project): ALS real-time data analysis writes to the NERSC NGF /project space. We routinely write ~35-40 TB per month, of which ~5 TB per month are considered permanent. We regularly run tools to reclaim disk space by two mechanisms:

1> Delete impermanent (i.e., reproducible data) data.

2> Archive on HPSS and purge older permanent (i.e., raw data) data.

Other uses of /project are modest in comparison. We estimate ~10-20 TB per year.

HPSS permanent archival storage: ALS real-time data analysis and our associated tools backup any irreproducible scientific data to HPSS before purging from disk. About 5 TB per month (60 TB/year) of permanent, *irreproducible* scientific data are archived to HPSS currently. We expect this to increase drastically in the 2017 time frame (see below).

We share data between Carver, DTNs, Science Gateway nodes, testbed nodes (such as Jessup), Hopper, and Dirac.

We use HDF5 libraries and do sequential I/O, one file per job (i.e., no parallel I/O). We anticipate utilizing parallel I/O in our codes during FY 2014.

Real-time analysis throughput is limited by a combination of startup & teardown times and data I/O. (i.e., CPU time is not our limiting factor).

11.2.3.3 Parallelism

The parallelism of the analysis varies significantly from beamline to beamline: tomographic reconstruction codes use ~100 cores concurrently (split across multiple jobs) while Monte-Carlo based GI codes (to be deployed in production in the next calendar year) can scale to 10's of thousands of cores. We do run multiple jobs concurrently when a new data set arrives at NERSC before the completion of the analysis on a previous set. This is compounded as datasets arrive from multiple beamlines. In general weak scaling (by increasing the amount of data to be analyzed) is more important to us than strong scaling (i.e. using more computational cores on a given data set). However, this may vary from beamline to beamline; in particular, the GISAXS codes are able to strong scale effectively.

11.2.4 HPC Requirements in 2017

11.2.4.1 Computational Hours Needed

As mentioned in 1.3.1, the current allocation is not reflective of the needs of ALS once the science/data portal is fully deployed. For 2014 ALS is requesting 15,000,000 hours because we are rolling out the first instance of the portal to one beamline. All of the users on that beamline will participate in the portal. We anticipate adding two more beamlines to the portal and to continue to develop analytical codes that take advantage of the HPC environment at NERSC. The 2014 request is 5,000,000 hours. By 2017, it is likely that ALS will bring on two more data intensive beamlines, COSMIC (Coherent Scattering and Diffraction Microscopy) and MAESTRO (Magnetic and Electronic Structure Observatory). In addition, there will be increased data flows from spectroscopy and photoemission work that that will be brought into the portal.

Based on these estimates, we anticipate requiring 45,000,000 computational hours.

11.2.4.2 Data and I/O

1. Scratch (temporary) space: We will be using high-performance scratch for I/O intensive processes and for temporary, intermediate data artifacts starting somewhere in FY2014-FY2015. We predict that we will need access to enough to handle 10-40 datasets in parallel. This translates into 5-20 TB of scratch.

2. Permanent (can be shared, NERSC Global Filesystem /project): We will be using ~100-400 TB of /project as cache for real-time raw data and derivatives. We would expect to achieve parallel I/O rates of 10-100 GB/s.

3. HPSS permanent archival storage: We will be adding approximately 1 PB of data to HPSS per year.

- 1 GB/sec I/O rates

11.2.4.3 Scientific Achievements with 32X Current Resources

Historically, the responsibility of a facility like ALS towards its users often ends when the users copies their data onto portable storage media. If ALS is to continue to enable cutting edge photon science, that relationship will need to be extended to include data management and analysis capabilities. Because the science is heterogeneous, predicting as specific achievement with more computing resources is difficult. That said, dramatically increased computing resources will enable more and richer data sets generated at ALS to be analyzed. As the number of users using the NERSC facilities grows, the simultaneous access to compute resources will be necessary.

11.2.4.4 Parallelism

By 2017, we expect to be simultaneously processing data from multiple ALS beam-lines in near real-time. Additionally, we expect concurrent analysis will be performed from users around the world via the web-portal. The parallelism of the analysis varies significantly from beamline to beamline from tomographic reconstruction codes that use ~100 cores concurrently to Monte-Carlo based GISAXS codes that can scale to 100's of thousands of cores (and have done so on Titan).

11.2.4.5 Memory

Simulation codes will likely have much higher needs that today. Extrapolation at this time is impossible.

Real-time analysis currently takes ~3GB/core and will likely increase by a factor of x4-x8 by 2017.

11.2.4.6 Many-Core and/or GPU Architectures

The GISAXS codes in particular have been optimized for GPUs and tomographic codes capable of efficiently utilizing a GPU are in development. We believe GPUs at NERSC, whether in the form of a rack, midsize cluster or large cluster would benefit the ALS workload. Porting to other many-core architectures is likely possible (particularly after the work in identifying areas for on-node parallelism has already been completed during the GPU porting phase).

11.2.4.7 Software Applications and Tools

The ALS beamline software landscape is largely ad-hoc and relies heavily on a limited number of experts to handle data and data analysis. The result is that beamline efficiency and ability to address important scientific questions are diminished. In order to use NERSC, codes will need to be made ready for the HPC environment. Section 1.2.2 lists some current codes and their requirements that will be made available to the users via the portal. Part of the ALS repo allocation includes capacity to modernize other codes to expand the utility of the portal.

11.2.4.8 HPC Services

ALS will require many of the components of the suite of services offered by NERSC. In particular, data analytics and visualization, collaboration tools, web interfaces, federated authentication services, and gateway support will be needed.

11.2.4.9 Time to Solution and Throughput

Real-time feedback to during ALS beam time is a capability critically needed by many ALS beamline users, yet unobtainable for very large data sets. Time dependent studies of crack formation in advanced composite materials under stress, or of dendrite formation leading to failure of Lithium ion batteries, or of supercritical CO_2 flow through rock for geologic carbon sequestration are all examples of time-resolved, in situ experiments at the ALS that require real-time feedback. Beam time is a precious commodity that may only be available to a particular user for one relatively short period during a year. Making the most efficient use of that time requires real-time feedback. Each of these science programs requires beam time feedback and are currently doing a fraction of the science possible with the ALS facility. Real-time feedback is also needed to evaluate experimental results in the context of data simulations before proceeding with a collection of a number of large data sets. While simulations might be run prior to beam time, the evaluation of data quality must be done in real-time. This requirement will change the way NERSC interacts with a class of its future users. Traditionally, execution of jobs is controlled by a queue. That process will not satisfy the requirements of some ALS users. The solution to this will require collaboration between ALS, NERSC, and computational staff. This may also drive a policy discussion between ASCR and BES.

11.2.4.10 Data Intensive Needs

Real-time & On-Demand Queues: To permit time-resolved, in-situ experiments and to provide beam-time feedback to scientists as they conduct experiments, we need access to CPU resources that can be reliably marshaled on demand. Techniques like message queues or dynamic resource allocation will need to be developed and integrated with our workflow systems.

High Throughput Queues: HTC resources for large data will be required to handle high Velocity data streams from beamlines like COSMIC at the ALS.

Science Servers & Services: Experiment/Facility-specific services running on either dedicated or shared servers (not interactive nodes, nor batch nodes with time or CPU limits), will be necessary for real-time analysis orchestration, data management processes, and user interfaces and/or data sharing. Science Data Gateways are anticipated to be a continued service provided.

GPUs & Special Architectures: Several simulation and analysis codes being currently developed run exceptionally well on GPUs. Understanding how these special-purpose and/or customized architectures could be integrated with NERSC CPU, disk, and tape resources will be an important component of our research over the next 5 years.

11.2.4.11 What Else?

Looming data policy requirements are likely to include access to facility generated data by a class of user not involved with generation of the data. Such a user may not even be a collaborator of the data-generating user. A user of this sort has been referred to as a Data User. This form of "public access" to data generated at a federally supported facility is best facilitated by a professionally managed facility like NERSC. In 2017 and beyond, it is possible that all data from ALS will need to be archived at NERSC. This will put added demand on NERSC services such as consulting or account support, data analytics and

visualization, training, collaboration tools, web interfaces, federated authentication services, gateways, etc. To date, there is no mechanism to support Data Users.

11.2.4.12 Requirements Summary

	Used at NERSC in 2013	Needed at NERSC in 2017
Computational Hours (Hopper core-hour equivalent)	4.3 M	45 M
Scratch storage and bandwidth	0 TB	10 TB
	0 GB/sec	10 GB/sec
Shared global storage and bandwidth (/project)	75 TB	200 TB
	1 GB/sec	10 GB/sec
Archival storage and bandwidth (HPSS)	536 TB	5,000 TB
	1 GB/sec	4 GB/sec
Number of conventional cores used for production runs	50 - 50,000	50 - 200,000
Memory per node	3 GB	12 GB
Aggregate memory	(NA) TB	(NA) TB

11.3 Advanced Modeling for Next-Generation BES Accelerators

Principal Investigator: Robert D. Ryne (Lawrence Berkeley National Laboratory)
NERSC Repository: m669

11.3.1 Project Description

11.3.1.1 Overview and Context

Particle accelerators are among the most versatile and important tools of scientific discovery. The Nation's accelerators are responsible for a wealth of advances in materials science, chemistry, the biosciences, particle physics, and nuclear physics. They also have important applications to the environment, energy, and national security, and are highly beneficial to the US economy by helping to maintain leadership in science and technology. Accelerators also have a direct impact on the quality of people's lives through medical applications including radioisotope production, pharmaceutical drug design and discovery, and through the thousands of accelerator-based irradiation therapy procedures that occur daily at U.S. hospitals.

The success of the Linac Coherent Light Source (LCLS) at SLAC marks the beginning of a new era in accelerator science, the era of "4th generation" light sources based on x-ray free electron lasers (XFELs). These facilities enable the exploration of systems with unprecedented temporal resolution, allowing ultra-fast biochemical processes to be observed and explored for the first time. The extraordinary opportunities presented by 4th generation light sources have led to their planned development worldwide. R&D is already underway for the design and development of new, novel approaches to "seed" the FEL to reduce size and cost, and to provide greater control of the radiation production.

Large-scale computational modeling is essential for the design of future light sources. These devices involve the interaction of particle beams, synchrotron radiation, and lasers, over a very wide range of spatial and temporal scales. The physical phenomena that are modeled involve 3D nonlinear dynamics in applied electromagnetic fields and nonlinear collective interactions. Researchers are now able to model some collective effects with a "real-world" number of simulation particles – 6.24 billion particles in a 1 nanoCoulomb bunch. It has been demonstrated in simulations that these single particle effects are important and need to be included in the design process to reliably predict light source performance.

While parallel simulation codes have proven very successful for existing light sources like LCLS, there remain important and challenging computational issues for the exploration, development, and optimization of future concepts. Examples include:

- Coherent synchrotron radiation (CSR): CSR is among the most important beam physics issues driving the design of future X-ray FELs, but it remains a major simulation challenge. Parallel beam dynamics codes, even those with 3D space-charge effects, often use simplified models of CSR, frequently a 1D model. The success of the LCLS has shown that collective effects such as CSR and space charge are adequately modeled in certain regimes. However, there remain some unexplained discrepancies between experiment and simulation, and large-scale simulation with more realistic models is important for understanding these discrepancies. Furthermore, some future light source concepts involve new regimes

of shorter bunches and lower emittances for which simplified models (like the 1D CSR model) have not been validated.

- Seeding: The study of innovative schemes to seed the FEL (which could dramatically reduce overall facility cost) requires high fidelity modeling at submicron wavelength and with 3D effects included, a capability currently not possible with existing codes. BES and other agencies are supporting advanced modeling R&D to address this situation.

- Parallel Design Optimization: Until recently, it was possible to perform parallel simulations to evaluate accelerator designs, but it was difficult to use large-scale simulation as a design tool. R&D efforts are now underway to develop and implement parallel design optimization capabilities. The tools include single- and multi-objective optimization, and allow a range of optimization approaches including, e.g., differential evolutionary algorithms.

Parallel 3D multi-physics modeling, including an accurate model of CSR, combined with parallel optimization capability, will provide an invaluable tool to meet the challenges of future light source concepts. Given the importance of light sources, and their potential billion-dollar cost, large-scale modeling is crucial for design optimization, cost and risk reduction, and the exploration of innovative ideas for which it would be too difficult or too expensive to perform physical experiments to test new design concepts.

11.3.1.2 Scientific Objectives for 2017

As will be described below, parallel beam dynamics simulation including 3D space-charge effects has come into widespread use throughout the accelerator community. In contrast, up to now there has been essentially no usable code for modeling 3D radiative effects. Codes such as IMPACT and Elegant use a 1D CSR model. But 3D effects are known to be important in certain situations, and they are likely to become more important through the increased reliance on complex beam manipulation systems that mix that transverse and longitudinal beam phase space distributions. The only code with 3D CSR capability, the CSRtrack code developed in Europe, runs very slowly in 3D mode and the capability is rarely used. Radiative phenomena like CSR are arguably among the most challenging phenomena to simulate in beam dynamics codes. Compared to the advances in modeling space-charge effects, modeling radiative phenomena lags far behind. But that is about to change. One of our main goals is to remedy this situation so that radiative phenomena can be modeled with the required realism and resolution needed to design the next generation of light sources.

Our objectives are: (1) To develop a parallel, scalable capability for including radiative effects such as CSR in accelerator design codes; (2) to develop and implement methods for efficient modeling of collective phenomena with wide-ranging and very high resolution (from 10's of microns down to sub-nanometer scale); (3) to develop methods for modeling shot-noise effects, (4) to embed our light source modeling capabilities in a parallel design optimization framework; (5) to apply these capabilities to design and optimize BES light sources and explore and develop advanced concepts, and (6) to distribute and deploy these advanced modeling capabilities to the BES light source design community.

11.3.2 Computational Strategies (now and in 2017)

11.3.2.1 Approach

Thanks to advances in mathematical models for treating radiative phenomena, parallel optimization algorithms, and the availability of the latest HPC resources, we are on the verge of new era in advanced simulation of light sources.

Parallel simulation of beam dynamics in particle accelerators began in earnest in the mid-1990s. Prior to that time, the accelerator community had made great progress in developing methods and serial codes for modeling single-particle nonlinear dynamics (e.g. the development of Lie algebraic mappings, symplectic integrators, etc.) and for modeling 2D space-charge effects. With the advent of parallel computing, space-charge effects could be usefully modeled in 3D for the first time, and these effects were combined with high-order beam optics effects using techniques such as split-operator methods. Throughout the 2000's and up to the present time, the trend in beam dynamics codes has been toward increasingly large scale, multi-physics modeling. Parallel beam dynamics codes now contain, and are routinely used to model, a variety of phenomena including high-order optics, 3D space-charge, beam-beam effects, structure wakefields, 1-D CSR, electron-cloud effects, and beam-material interactions. As an example of how far we have come, consider that in 2013 the IMPACT-T and IMPACT-Z parallel PIC codes were combined with the GENESIS FEL code to produce a single parallel executable, and were used for the first start-to-end simulation of a future light source simulated with a real-world number of electrons (2 billion in this example). The simulation took 14 hours on 2,048 cores of Hopper, and included 3D space-charge effects, structure wakefield effects, and 1D CSR effects.

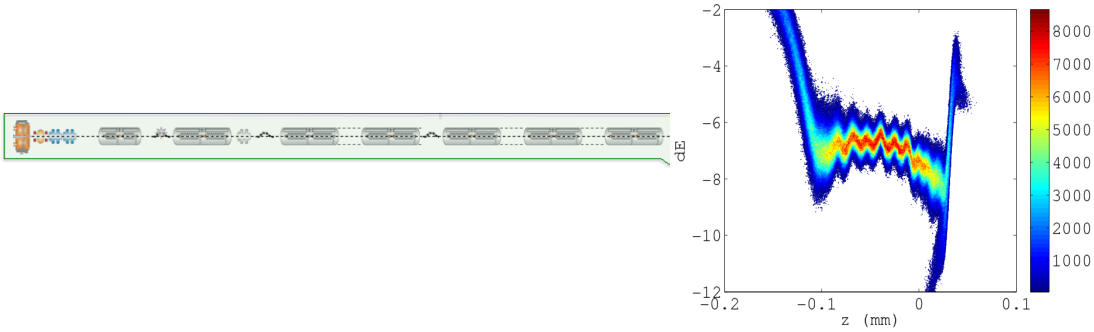


Figure 1: Schematic of a future light source (left), and results from a start-to-end simulation showing the final longitudinal phase space distribution. The simulation used a "real world" number of electrons (2 billion) and included 3D space-charge effects, structure wakefields, and 1D coherent synchrotron radiation (CSR) effects. This parallel simulation was performed on Hopper at NERSC, and required 10 hours on 2048 cores. By embedding this capability in a parallel optimizer and scaling to of order 100k cores, it is now possible to perform start-to-end global optimization using this model. The next major step in code development for future light sources involves developing techniques for high-resolution, 3D modeling of radiative effects such as CSR.

Despite the major advances in parallel multi-physics beam dynamics modeling, the simulation of radiative phenomena has remained a major challenge and has lagged far behind the treatment of 3D space-charge effects. Radiative phenomena are critical to future light sources. Examples include coherent synchrotron radiation (CSR), incoherent synchrotron radiation (ISR), and undulator radiation. A first-principles classical treatment usually involves the Lienard-Wiechert formalism. Since this involves quantities when the radiation was emitted (i.e. at retarded times and locations), it requires storing a history of each particle's trajectory. Also, CSR phenomena can exhibit large fluctuations that are physical, not numerical, hence it is often necessary to use as close to a real-world number of particles as possible. The calculation of retarded quantities is iterative and extremely time consuming. Consider that the calculation of an electric field component on a grid in an electrostatic code, e.g., $x/|r|^3$, requires only a small number of floating point operations; by contrast the calculation of the L-W field requires a small simulation code itself. In addition, to embed such a capability in a self-consistent beam dynamics code greatly compounds the effort, leading to massive requirements for FLOPs and memory.

Despite these challenges, recent progress indicates that we are on the verge of being able to model 3D radiative effects using a Lienard-Wiechert approach. This is made possible by new, convolution-based Lienard-Wiechert solvers. In this approach, the Green function, which normally depends on both the observation point and the retarded quantities, is reduced to a function of just one quantity. This in turn makes it possible to solve for the fields in O(N log N) operations, where N is the number of grid points. An example is shown in Figure 2, which compares Lienard-Wiechert summation with convolution-based simulation for two test problems. The results are nearly identical. It is worth emphasizing that this is not an electrostatic calculation (for which the technique is well established), this is an electromagnetic calculation involving the retarded Green function. The brute force summation scales as the square of the number of particles, making it totally unsuitable for practical use.

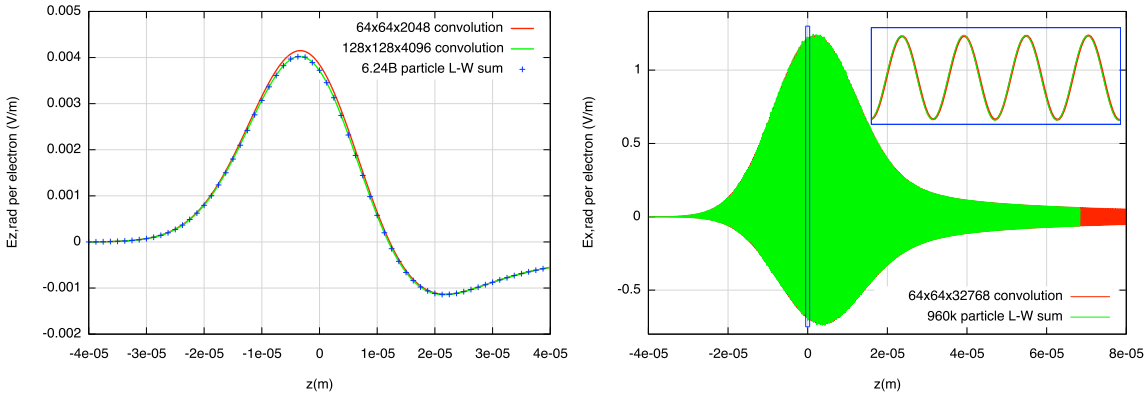


Figure 2: Comparison of two methods -- brute-force Lienard-Wiechert (L-W) summation and L-W convolution with a retarded Green function -- for two test problems. Left: z-component of the radiation electric field of a 1 GeV Gaussian bunch inside a dipole magnet. Right: x-component of the radiation electric field of a 125 MeV modulated Gaussian bunch inside an undulator of a free electron laser. The two methods are in excellent agreement for both test problems. The narrow blue rectangle on the right figure shows the domain of the inset, and demonstrates excellent agreement even at the level of the radiation wavelength (250 nm in this example).

11.3.2.2 Codes and Algorithms

IMPACT, IMPACTgs: The IMPACT code suite contains 2 parallel particle-in-cell (PIC) codes, along with auxiliary programs, to simulate high intensity, high brightness beam transport in particle accelerators, including 3D space-charge effects. It has demonstrated more than 90% efficiency in weak scaling on up to 100,000 processors. IMPACTgs is parallel code combining IMPACT with the GENESIS FEL code.

CSR3D: This is a new parallel PIC code that has been developed to model both space-charge effects and radiative effects in charged particle beams in accelerators. It makes use of the LW3D convolution-based solver, which is a parallel code for computing the Lienard-Wiechert fields associated with a beam bunch in an accelerator.

11.3.3 HPC Resources Used Today

11.3.3.1 Computational Hours

We used 5.2 M hours in 2013.

11.3.3.2 Parallelism

We typically use 2,000 to 10,000 cores. We have done simulations on up to 100,000 cores.

We expect we could use up to a million cores at this time for problems involving parallel design optimization. We already have parallel optimizers that use 2-level parallelism, one level for each "point" simulation in parameter space and a second level for the optimizer. Scalability of these simulations is dominated by the scalability of the "point" simulations, which have already been demonstrated up to 10K, but are typically run with 2K cores. Multiplying this by of order 100 "population members" in a differential evolutionary optimizer leads to simulations of 200k to 1M cores.

Accelerator design is an activity that normally requires timely interaction with the accelerator designer. While long wait times are acceptable for occasional "heroic" simulations, generally speaking wait times of more than a day are unacceptable, and wait times of hours are highly preferable. We would gladly use 100K processors for our typical design optimization runs if they would start within a few hours from the time we submit them.

We do not use high throughput computing.

Whether weak scaling or strong scaling is important depends on what is being simulated. Strong scaling is more important when we are modeling a single accelerator design. On the other hand, when we are doing parallel design optimization, weak scaling is important since it allows us to use more population members in a differential evolutionary optimizer, thereby leading to a faster time-to-solution. However, since the communication associated with the optimizers is small, the dominating factor is the strong scalability of the point simulations.

11.3.3.3 Scratch Data

Scratch is not an issue for us at this time.

11.3.3.4 Shared Data

We have a project directory called m669. We use it to share data, and we use it like an extended home directory.

11.3.3.5 Archival Data Storage

We usually store reduced data, so archival storage is not normally an issue for us. If we stored raw data from all the steps of a simulation, the storage requirement would be as follows for the following example: 2 billion particles x 9 variables/particle = 144 GB per step in double precision. Multiplied by 10,000 steps, this corresponds to 14 PB.

11.3.4 HPC Requirements in 2017

11.3.4.1 Computational Hours Needed

We expect to need at least 100M hours in CY2017.

The primary factor is the transition to performing 3D simulations of radiative phenomena like CSR instead of 1D simulations.

11.3.4.2 Parallelism

The incorporation of convolution-based Lienard-Wiechert solvers in our codes will greatly increase the FLOPs count in our simulations. This will improve the scalability of our "point" simulations, which we expect to increase from 2K-10K cores (presently) to 100K cores in 2017. On the other hand, as mentioned above, wait time in queues is an important factor in accelerator design. Our "typical" use will be to use as many cores as possible while still having wait times of less than a day.

For parallel design optimization, multiplying 100K cores by 100 population members of a differential evolutionary algorithm leads to 10M cores.

We will not typically need more than one job running at once. Although our parallel optimizing runs, which are actually performed in a single job, and can be thought of as concurrent, loosely coupled jobs.

11.3.4.3 I/O

Our I/O requirements will vary depending on the type of analysis being performed. In some cases very little I/O is required, while in others (such as exploring beam phenomena at high resolution) I/O requirements can be a several tens of TB, to as much as a PB if we stored all particle information from every step. Our parallel I/O activities and data analysis activities will build on our already successful collaboration with the ExaHDF5 team.

Storing O(100TB) datasets for future runs will need scaling of I/O hardware resources. We believe that an aggregate bandwidth of 500GB/s-1TB/s will be required for storing our datasets on disk. It is worth noting that in-situ visualization and analysis is applicable to a small fraction of our small use cases, and we will need to store datasets at a high temporal resolution going forward.

We would like to keep the percentage of total runtime devoted to I/O as low as possible. Between 5-10% would be considered reasonable for our work.

11.3.4.4 Scratch Data

We will require O(100TB) for large runs in the future. To the extent that /scratch systems will be faster than project, we want to avail of high performance I/O going forward.

What is causing the growth in scratch space is the use of higher resolution field data, and the desire to store a larger number of time steps.

11.3.4.5 Shared Data

Each full particle dump in a 2-billion particle simulation requires 144 GB. We are not likely to share more than several such dumps. As a result, our present project quota of 4 TB should be sufficient.

11.3.4.6 Archival Data Storage

We expect to store around 300 TB.

I/O is parallel and collective (writing to single shared HDF5 file).

11.3.4.7 Memory Required

We can control how much memory per node is used by controlling how many particles each node handles. But since Lienard-Wiechert codes use stored particle history data, we prefer to have as much memory as possible. Otherwise our codes become communication-dominated (i.e. if there are too few particles/node, then there would be too few FLOPs per/node, and we would be dominated by global operations like parallel FFTs).

11.3.4.8 Emerging Technologies and Programming Models

We are not ready for this now, but it would be advantageous in a Lienard-Wiechert code since there are many FLOPs associated with the iterative search to find so-called "retarded" quantities, and very little data movement required.

If it appears that the NERSC system in 2017 will have GPUs, then we would request help (either directly or through NERSC training classes) to GPU-ize the Lienard-Wiechert kernel.

11.3.4.9 Software Applications and Tools

We need Fortran compilers, serial FFTW, parallel FFTs, H5PART, H5BLOCK, and ExaHDF5.

11.3.4.10 HPC Services

We use consulting and analytics support and collaborate with the ExaHDF5 team.

11.3.4.11 Time to Solution and Throughput

As mentioned previously, accelerator design usually requires feedback to the designer in a matter of hours to at most a day. For that reason, we often run short premium jobs hoping that they will backfill and start quickly. Ideally, we would like to find a way to have several-hour jobs start running quickly.

11.3.5 Data Intensive Needs

We greatly value and appreciate NERSC's systems and resources. The one thing that would help us a lot is to be able to have jobs a few thousand processors, lasting up to, say, 12 hours, start running more quickly.

We do not have a data management plan in place now.

11.3.5.1 Requirements Summary

	Used at NERSC in 2013	**Needed at NERSC in 2017**
Computational Hours	5.2 M	100 M
Typical number of cores* used for production runs	few thousand	20-100K
Maximum number of cores* that can be used for production runs	> 100K	> 1M
Data read and written per run		O(100 TB) written for large runs
Maximum I/O bandwidth		500 - 1000 GB/sec
Percent of runtime for I/O		5-10%
Scratch File System space		100 TB
Shared filesystem space	0.25 TB	4 TB
Archival data	39 TB	300 TB
Memory per node	64GB	64 GB or greater

* "Conventional" cores

Appendix A. Attendee Biographies

Domain Scientists

Michael Banda is the Deputy Division Director for Operations of the Advanced Light Source. "Banda," as he prefers to be called, manages the overall operation of the ALS, including accelerator and beamline operations, user activities, safety, and environmental protection activities. He has been with Berkeley Lab since 1999. He began as the Deputy Division Director for Life Sciences, and then became the founding Deputy Division Director for Genomics at the time when the Joint Genome Institute was formed. In 2001, Banda moved on to become Deputy Division Director for Computing Sciences, where he held responsibilities with the Computational Research Division and the National Energy Research Scientific Computing Center (NERSC). His previous work with x-ray science includes an appointment as Professor of Radiology and Director of the Laboratory of Radiological Biology at the University of California, San Francisco, a unit that studied the effects of radiation in applications of biochemistry and cell biology.

Jacqueline Chen is a Distinguished Member of Technical Staff at the Combustion Research Facility of Sandia National Laboratories and Adjunct Professor of Chemical Engineering University of Utah, Utah. She received her Ph.D. in Mechanical Engineering from Stanford (1989) under the direction of Brian Cantwell. She is a world-renowned expert in the use of petascale direct numerical simulations (DNS) for turbulent combustion, with a focus on turbulence-chemistry interactions in canonical laboratory-scale flames. Se served as the co-editor of Proceedings of the Combustion Institute and is a member of the Editorial Advisory Boards of Combustion and Flame and Computational Science and Discovery.

Jack Deslippe is an HPC Consultant in the NERSC User Services Group where he specializes in the support of material science applications. He is engaged in evaluating and improving the suitability of these applications for potential N8 architectures and also works on bringing dynamic web-content to users through MyNERSC, the MOTD system, Completed Jobs Pages, ALS Science Gateway Projects and the NERSC mobile site, m.nersc.gov. Jack is a PI on a SCIDAC project (http://excited-state-scidac.org/) and is one of the lead developers of the BerkeleyGW package for computing the excited state properties of materials. Jack is the NERSC PI on the Berkeley Lab Directed Research project that is delivering real-time data analysis to ALS scientists through ESNET and NERSC resources. He is the developer the ALS analysis and simulation web-portal at NERSC. He received a Ph.D. from UC Berkeley in physics in 2011. His research centered on materials physics and nano-science: scaling many-body Green's function computational methods for the study of the optical properties of materials with large and complex structures.

Sanket Deshmukh is a postdoctoral researcher with Subramanian Sankaranarayanan in the Theory and Modeling Group, Center for Nanoscale Materials, at Argonne National Laboratory.

Andrew R. Felmy is a Laboratory Fellow at Pacific Northwest National Laboratory.

Scott French received his Ph.D. in Earth and Planetary Science at U.C. Berkeley Prof. Barbara Romanowicz in the Global Seismology Research Group at the Berkeley Seismological Laboratory. He recently joined NERSC as an HPC consultant.

Andreas Heyden is Assistant Professor of Chemistry at University of South Carolina. His research interests are in the areas of nanomaterial science and heterogeneous catalysis. His goal is to use computer simulations to obtain a deeper - molecular - understanding of key issues in these areas, such as the self-assembly process in catalyst synthesis, the structure of small metal clusters on high-surface-area supports, and the structure-performance relationship of single-site heterogeneous catalysts. He was granted a Ph. D. from Hamburg University of Technology in 2005.

Paul Kent is a member of the Nanotheory Institute at the Center for Nanophase Materials Sciences (CNMS) and the Computational Materials Science group in the Computer Science and Mathematics Division. He spent three years at NREL with Alex Zunger after completing a Ph.D. with Richard Needs at the University of Cambridge. For several years he worked with Mark Jarrell at the University of Cincinnati on high-temperature cuprate superconductors. He has been at Oak Ridge National Laboratory since 2009.

Yun Liu is a postdoctoral researcher in the MIT laboratory of Professor Jeffrey Grossman.

Thomas Miller graduated from Texas A&M University with honors in 2000 as a major in chemistry and mathematics. He received a British Marshall Scholarship to pursue graduate study in the U.K., which he used to obtain an M. Phil. from University College London in 2002. He then attended the University of Oxford on an NSF graduate research fellowship, earning a D. Phil. from Balliol College in 2005. Tom joined the Caltech faculty as an assistant professor in 2008, and he was promoted to full professor in 2013. While at Caltech, he has received the Dreyfus New Faculty Award, Sloan Research Fellowship, NSF CAREER Award, American Chemical Society Hewlett-Packard Outstanding Junior Faculty Award, Associated Students of Caltech Teaching Award, and the Dreyfus Teacher-Scholar Award.

Jeffrey Neaton was appointed director of the Molecular Foundry at LBNL in 2013. He also leads the Theory group at the Molecular Foundry. Jeff received his Ph.D. in Physics from Cornell University in 2000, under the guidance of Neil W. Ashcroft. After a departmental postdoc in the Department of Physics and Astronomy at Rutgers University, he joined the Molecular Foundry at Lawrence Berkeley National Laboratory in 2003. His current research interests center on computational nanoscience, in particular the development and application of methods for calculating the structural, spectroscopic, and transport properties of inorganic and molecular nanostructures, particularly at interfaces and contacts. Present areas of interest include the electronic properties of the metal-organic interface, hybrid silicon-organic interfaces, and single-molecule junctions; self-assembly; nanoparticle assemblies; photovoltaics; hydrogen storage; ultrathin epitaxial films of transition metal oxides, such as ferroelectrics and multiferroics; and structural and electronic phases of light elements under pressure.

Gregory Newman is a Senior Scientist at Lawrence Berkeley National Laboratory, Earth Science Division and Head of the Geophysics Department in the Earth Sciences Division. Prior to his appointment in January 2004, Dr. Newman worked nearly fourteen years at Sandia National Laboratories, Geophysical Technology Department. His interest, include large-scale, multi-dimensional, inverse and forward modeling problems arising in exploration geophysics, parallel computation and electromagnetic geophysics. He has over 20 years of experience in large-scale

geophysical field simulation and computation. In 2000, Dr. Newman was a Mercator Fellow at the Institute for Geophysics and Meteorology, University of Cologne, Federal Republic of Germany. The fellowship was awarded from the German National Science Foundation for a year of study in the Federal Republic of Germany. Studies at the Institute were directed on the formulation and implementation of 3D transient electromagnetic modeling and inversion algorithms for geophysical applications and lectures on the electromagnetic modeling and inversion. Dr. Newman was also affiliated with this institution from 1987-1989 as a Post Doctorate Appointee and an Alexander von Humboldt Fellow.

David Skinner is the Strategic Partnerships Lead at NERSC. Skinner holds a Ph.D. in theoretical chemistry from the University of California, Berkeley. His research focused on quantum and semi-classical approaches to chemical reaction dynamics and kinetics. He began working at NERSC/Berkeley Lab in 1999 as an HPC engineer and spent the last eight years leading the OSP group. During his 15-year career at NERSC, Skinner was the lead technical advisor to first two rounds of INCITE projects, led the SciDAC Outreach Center, and is an author of the Integrated Performance Monitoring (IPM) framework. He also published several papers on the performance analysis of HPC science applications and broadening the impact of HPC through Science Gateways.

Carl Steefel has over 21 years of experience in developing models for multicomponent reactive transport in porous media and applying them to topics in reactive contaminant transport and water-rock interaction. The reactive transport software CrunchFlow, for which he is the principal developer, is the culmination of this work. He investigated geochemical self-organization and complexity theory in water-rock interaction, while also developing the first routine for multicomponent nucleation and crystal size distributions in the Earth Sciences. Soon after, he presented the first multicomponent, multi-dimensional code for simulating water-rock interaction in non-isothermal environments. Steefel applies reactive transport modeling to such diverse settings as hydrothermal, contaminant, chemical weathering, and marine environments. He holds a Ph.D. in Geochemistry from Yale University and has been at LBNL since 1998.

Sotiris Xantheas Dr. Sotiris Xantheas is known in the chemical physics scientific community for his research in intermolecular interactions in aqueous ionic clusters and the use of ab-initio electronic structure calculations to elucidate their structural and spectral features. His research has ranged from the computation of potential energy surfaces for various chemical reactions using correlated wavefunctions to the elucidation of reaction paths governing carbene ring opening processes and the location and characterization of intersections of potential energy surfaces of the same symmetry in polyatomic systems. He has recently utilized the results of high-level electronic structure calculations to parameterize a family of ab-initio based interaction potentials for water and used those potentials to simulate the macroscopic properties of liquid water and ice.

NERSC Editors

Richard Gerber is NERSC Senior Science Advisor and User Services Group Lead. Together, with Harvey Wasserman he organizes the NERSC High Performance Computing and Storage Requirements Reviews for Science and edits the reports. He holds a Ph.D. in physics from the University of Illinois at Urbana-Champaign, specializing in computational astrophysics; held a National Research Council postdoctoral fellowship at NASA-Ames Research Center 1993-1996; and has been on staff at NERSC since.

Harvey Wasserman is a member of the NERSC User Services Group and helps to organize the NERSC High Performance Computing and Storage Requirements Reviews.

Appendix B. Meeting Agenda

Tuesday, October 8		
8:00 AM	Informal discussions	
8:30 AM	Welcome, Overview of Requirements Reviews	Dave Goodwin, ASCR (NERSC Program Manager); Richard Gerber (NERSC)
8:45 AM	Computing in Basic Energy Sciences	James Davenport, BES
9:15 AM	NERSC's 10-Year Plan	Sudip Dosanjh, NERSC Director
9:45 AM	AM Break	
	Geosciences Case Studies	
10:00 AM	Large Scale 3D Geophysical Inversion & Imaging	Gregory Newman, LBNL
10:20 AM	Computational Studies in Molecular Geochemistry	Andy Felmy, PNNL
10:50 AM	Direct Numerical Simulation of the Poisson-Nernst-Planck Equation in Clay	Carl Steefel, LBNL
11:20 AM	Global-scale full-waveform seismic imaging of Earth's mantle	Scott French, UC Berkeley
	Materials Science Case Studies	
11:40 AM	Computational Resources for the Nanomaterials Theory Institute at the Center for Nanophase Materials Sciences	Paul Kent, ORNL
12:10 PM	Group Photo	
12:30 PM	Working Lunch Presentation. "Transitioning to NERSC-8 and Beyond: The NERSC Application Readiness Effort"	Jack Deslippe, NERSC
	Materials Science Case Studies (continued)	
1:00 PM	The Materials Project	David Skinner, NERSC
1:30 PM	Large-Scale Computation for Excited State Phenomena	Jeff Neaton, LBNL
2:00 PM	Computational Design of Novel Energy Materials	Yun Liu, MIT
2:30 PM	PM Break	
	Scientific User Facility Case Studies	
2:45 PM	Advanced Modeling for Next-Generation BES Accelerators	Robert Ryne, LBNL
3:00 PM	Advanced Light Source	Jack Deslippe, NERSC
	Combustion Case Studies	
3:30 PM	Direct Numerical Simulations of Clean and Efficient Combustion with Alternative Fuels	Jackie Chen, Sandia National Laboratories
	Chemical Sciences Case Studies	
4:00 PM	Rational Catalyst Design for Energy Production	Andreas Heyden, Univ. South Carolina
4:30 PM	Group Discussions	All participants
5:30 PM	Cross-Cutting Issues in Data Storage, Transfer, and Analysis	David Skinner and all Participants
6:00 PM	Adjourn for the day	

Wednesday, October 9		
8:00 AM	Informal Discussion	
	Chemical Sciences Case Studies (Cont'd)	
8:30 AM	Chemical reactivity, solvation and multicomponent heterogeneous processes in aqueous environments	Sotiris Xantheas, PNNL
9:00 AM	Molecular Dynamics of PNIPAM Agglomerates and Composite Architectures	Sanket Deshmukh, ANL
9:30 AM	Sampling Diffusive Dynamics on Long Timescales, and Simulating the Coupled Dynamics of Electrons and Nuclei	Tom Miller, Caltech
10:00 AM	AM Break	
10:15 AM	High-Level Findings Report	All Participants
11:00 AM	Schedule for Report	Richard Gerber & Harvey Wasserman
11:15 AM	Case Study Report Refinement and Discussions	All Participants
12:00 PM	Working Lunch: Case Study Breakout Sessions	All Participants
1:00 PM	Adjourn	

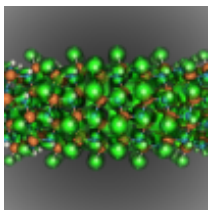
Appendix C. Abbreviations and Acronyms

ALCC	ASCR Leadership Computing Challenge
ALCF	Argonne Leadership Computing Facility
ALS	Advanced Light Source
AMR	Adaptive Mesh Refinement
API	Application Programming Interface
ASCR	Advanced Scientific Computing Research, DOE Office of
AY	Allocation Year
BER	Biological and Environmental Research, DOE Office of
CG	Conjugate Gradient
CSR	Coherent synchrotron radiation
CUDA	Compute Unified Device Architecture
DFT	Density Functional Theory
DNS	Direct Numerical Simulation
DTN	(NERSC) Data Transfer Node
EMSL	Environmental Molecular Sciences Laboratory at PNNL
ESG	Earth System Grid
ESnet	DOE's Energy Sciences Network
FEL	Free Electron Laser
FEM	Finite Element Modeling
FFT	Fast Fourier Transform
GA	Global Arrays
GI	Grazing-Incidence
GPGPU	General Purpose Graphical Processing Unit
GPU	Graphical Processing Unit
HDF	Hierarchical Data Format
HPC	High-Performance Computing
HPSS	High Performance Storage System
I/O	input output
IDL	Interactive Data Language visualization software
INCITE	Innovative and Novel Computational Impact on Theory and Experiment
LBNL	Lawrence Berkeley National Laboratory
LCLS	Linac Coherent Light Source
MD	Molecular Dynamics
MKL	(Intel) Math Kernel Library
MP	Massively Parallel
MPI	Message Passing Interface
NERSC	National Energy Research Scientific Computing Center
NetCDF	Network Common Data Format
NGF	NERSC Global Filesystem
NISE	NERSC Initiative for Science Exploration
NREL	DOE National Renewable Energy Laboratory
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
OS	operating system
PDE	Partial Differential Equation
PDSF	NERSC's Parallel Distributed Systems Facility
PES	Potential Energy Surface

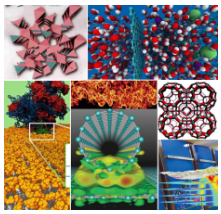
PNNL	Pacific Northwest National Laboratory
QM	Quantum Mechanics
QMC	Quantum Monte Carlo
SC	DOE's Office of Science
SciDAC	Scientific Discovery through Advanced Computing
SLAC	SLAC National Accelerator Laboratory
SNL	Sandia National Laboratories
SPH	Smoothed Particle Hydrodynamics
STF	Solar Thermal Fuels
TDDFT	Time-dependent density functional theory
VASP	Vienna Ab initio Simulation Package
XFEL	X-ray Free Electron Laser
XSEDE	Extreme Science and Engineering Discovery Environment

Appendix D. About the Cover

Image showing a portion of NERSC's "Hopper" system, a Cray XE6 installed during 2010. Hopper is NERSC's first peta-FLOP resource, with a peak performance of 1.28 PetaFLOPs/sec, 153,216 compute cores, 212 Terabytes of memory, and 2 Petabytes of disk. Hopper placed number five on the November 2010 Top500 Supercomputer list.



Results from a simulation of a 1 nanometer-wide indium nitride wire showing electron density distribution around a positively charged "hole." Strong quantum confinement in these small nanostructures enables efficient light emission at visible wavelengths, researchers found. (Simulation and analysis by Dylan Bayerl and Emmanouil Kioupakis, University of Michigan. Visualization created by Burlen Loring, Lawrence Berkeley National Laboratory, using ParaView. See "Visible-Wavelength Polarized-Light Emission with Small-Diameter InN Nanowires," American Chemical Society *Nano Lett*., 2014, 14 (7), pp 3709–3714)



Montage depicting research activities within the DOE Office of Basic Energy Sciences at NERSC. Image credits, from top, left: Structure of Mullite, from the cover of the Journal of the American Ceramic Society, image created by Prof. Wai-Yim Ching of U. Missouri KC; 3D visualization of water molecules (red and white) and sodium and chlorine ions (green and purple) in saltwater, on the right, encountering a sheet of graphene (pale blue, center) from a simulation related to water desalination. Graphic: David Cohen-Tanugi, MIT; visualization showing a ribosome (red-blue) in complex with a translocon channel (green) that is embedded in a cell membrane (yellow, white), Image credit: Bin Zhang and Thomas Miller, Caltech 2012; turbulent mixing and reaction chemistry in the DNS simulation of planar jet flame, image courtesy of Evatt R. Hawkes, Ramanan Sankaran, James C. Sutherland, Jacqueline H. Chen; Image showing results of a first-principles electronic structure and transport study of the junction between a carbon nanotube and graphene, a type of junction that may turn out to be useful for transistors, from the cover of the Applied Physics Letters, work done by Brandon G. Cook, William R. French, and Kalman Varga, Vanderbuilt U.; model of a zeolite molecule investigated using a waste recycling Monte Carlo simulation to evaluate thermodynamics and kinetics associated with molecular absorption and motion through the molecule, image courtesy of Jihan Kim, Jocelyn M. Rodgers, Manuel Athènes, and Berend Smit, Journal Chemical Theory and Computation; Image representing the result of a 3D numerical seismic and electromagnetic wave propagation and diffusion simulation superimposed on a portion of NERSC's Cray XT4 supercomputer; image courtesy of LBNL Earth Sciences Division, http://esd.lbl.gov/departments/geophysics/core_capabilities/computational_geophysics.html

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.