Title

High Performance Computing and Storage Requirements for Biological and Environmental Research
Target 2017

Author

Gerber, Richard

Publication Date

2013-05-23

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 Dorothy Koch, Renu Joseph, and Susan Gregurick serve as the Biological and Environmental Research (BER) 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 Biological and Environmental Research, and the Office of Advanced Scientific Computing Research, Facilities Division.

High Performance Computing and Storage Requirements for Biological and Environmental Research
Target 2017

Report of the HPC Requirements Review

Conducted September 11-12, 2012

Rockville, MD

DOE Office of Science

Office of Biological and Environmental Research (BER)
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

The National Energy Research Scientific Computing Center (NERSC) is the primary computing center for the DOE Office of Science, serving approximately 4,500 users working on some 650 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 September 2012 NERSC, DOE's Office of Advanced Scientific Computing Research (ASCR) and DOE's Office of Biological and Environmental Research (BER) held a review to characterize High Performance Computing (HPC) and storage requirements for BER research through 2017. This review is the seventh in a series that began in 2009 and it is the second for BER. The report from the 2009 BER 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 BER computing and storage needs. High-level findings are:

1. Scientists need access to significantly more computational and storage resources to achieve their goals and reach BER research objectives. BER anticipates a need for six billion computational hours (25 times 2012 usage) and 107 PB of archival data storage (10 times 2012 usage) at NERSC in 2017.
2. Simulation and analysis codes will need to access, read, and write data at a rate far beyond that available today.
3. Support for high-throughput job workflows is needed.
4. State-of-the-art computational and storage systems are needed, but their acquisition must not interrupt ongoing research efforts.
5. NERSC needs to support data analytics and sharing, with increased emphasis on combining experimental and simulated data.

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 BER. The case study topics were selected by the NERSC meeting coordinators and BER program managers to represent the BER production computing workload. Prepared by BER 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 BER Mission

The U.S. Department of Energy's Office of Biological and Environmental Research (BER) conducts research in the areas of Climate and Environmental Sciences and Biological Systems Science. BER's scientific impact has been transformative. In 1986, the Human Genome Project gave birth to modern biotechnology and genomics-based systems biology. Today, with its Genomic Sciences Program and the DOE Joint Genome Institute (JGI), BER researchers are using powerful tools of plant and microbial systems biology to pursue breakthroughs needed to develop cost-effective cellulosic biofuels. Our three DOE Bioenergy Research Centers lead the world in fundamental biofuels research.

Since the 1950s, BER has been a critical contributor to climate science research in the U.S., beginning with studies of atmospheric circulation—the forerunners of climate models. Today, BER supports the Community Earth System Model, a leading U.S. climate model, and addresses two of the most critical areas of uncertainty in contemporary climate science—the impact of clouds and aerosols—through support of the Atmospheric Radiation Measurement Climate Research Facility, which is used by hundreds of scientists worldwide.

BER plays a unique and vital role in supporting research on atmospheric processes; terrestrial ecosystem processes; subsurface biogeochemical processes involved in nutrient cycling, radionuclide fate and transport, and water cycling; climate change and environmental modeling; and analysis of impacts and interdependencies of climatic change with energy production and use. These investments are coordinated to advance an earth system predictive capability, involving community models open to active participation of the research community. For more than two decades, BER has taken a leadership role to advance an understanding of the physics and dynamics governing clouds, aerosols, and atmospheric greenhouse gases, as these represent the more significant weaknesses of climate prediction systems. BER also supports multidisciplinary climate and environmental change research to advance experimental and modeling capabilities necessary to describe the role of the individual (terrestrial, cryospheric, oceanic, and atmospheric) component and system tipping points that may drive sudden change. In tight coordination with its research agenda, BER supports two major national user facilities, i.e., the ARM Climate Research Facility and Environmental Molecular Sciences Laboratory, and significant investments are provided to community data base and model diagnostic systems to support research efforts.

† 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 4,500 scientists working on over 650 projects of national importance. Operated by Lawrence Berkeley National Laboratory (LBNL), NERSC is the primary high-performance computing facility for scientists in all 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 Biological and Environmental Research. 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.

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

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

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

[3] January through July. Usage for all of 2012 at NERSC was 34.4 million hours.

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 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 September 2012, ASCR (which manages NERSC), BER, and NERSC held a review to gather HPC requirements for current and future science programs supported by BER. This report is the result.

This document presents a number of findings, based upon a representative sample of projects conducting research supported by BER. The case studies were chosen by the DOE Program Office Managers and NERSC staff to provide broad coverage in both established and incipient BER 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 current and future science, a brief description of computational methods used, and a description of current and future computing needs. Since supercomputer architectures are trending toward systems with chip multiprocessors containing hundreds or thousands of cores per socket and millions of cores per system, participants were asked to describe their strategy for computing in such a highly parallel, "many-core" 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 Biological and Environmental Research.

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. A second round of meetings, of which this is the first, will target needs for 2017.

Specific findings from the review follow**.**

5 Workshop Demographics

5.1 Participants

5.1.1 DOE / NERSC Participants and Organizers

Name	Institution	Area of Interest
Todd Anderson	DOE / BER	BER Program Manager
Shane Canon	NERSC	Technology Integration Group Lead
Richard Gerber	NERSC	Meeting Organizer
Dave Goodwin	DOE / ASCR	NERSC Program Manager
Susan Gregurick	DOE / BER	BER Program Manager
Renu Joseph	DOE / BER	BER Program Manager
Dorothy Koch	DOE / BER	BER Program Manager
Yukiko Sekine	DOE / ASCR	NERSC Program Manager
Harvey Wasserman	NERSC	Meeting Organizer
Katherine Yelick	NERSC / Berkeley Lab	Associate Laboratory Director for Computing Sciences
Sudip Dosanjh	NERSC	NERSC Director

5.1.2 Domain Scientists

Name	Institution	Area of Interest	NERSC Repo(s)
Mohammed AlQuraishi	Stanford University	Bioscience	m926
David Bader	Lawrence Livermore National Laboratory	Climate	
Thomas Bettge	National Center for Atmospheric Research	Climate	mp9
Tom Brettin	DOE / BER	BER Program Manager	kbase
William Collins	DOE / BER	Climate	m1024, m1040, m1343, m1196
Gilbert Compo	DOE / BER	Climate	m958
Robert Egan	Joint Genome Institute	Genomics	m342
David Goodstein	Joint Genome Institute	Genomics	m342
Ruby Leung	Pacific Northwest National Laboratory	Climate	m1040, m1209, m1178
Stephen Price	Los Alamos National Laboratory	Climate	m1041, m1343
Victor Markowitz	Berkeley Lab	Genomics	m1045
Loukas Petridis	Oak Ridge National Laboratory	Bioscience	m906
Cristiana Stan	Institute of Global Environment and Society (IGES)	Climate	m1441
Jin-Ho Yoon	Pacific Northwest National Laboratory	Climate	mp9, m1199, m1178
Timothy Scheibe	Pacific Northwest National Laboratory	Environmental Science	m749

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 each used in 2012. These projects accounted for about three-fifths of computer time and archival storage used by BER at NERSC that year.

NERSC Project ID (Repo)	NERSC Project Title	Principal Investigator	Workshop Speaker(s)	Hours Used at NERSC in 2012 (M)	Archival Data at NERSC 2012 (TB)
Climate / Environmental Science					
mp9	*Climate Change Simulations with CESM: Moderate and High Resolution Studies*	Warren Washington	Tom Bettge	34.4	1,542
m958	*Sparse Input Reanalysis for Climate Applications (SIRCA) 1850-2012*	Gil Compo	Gil Compo	11.7	1,005
m1199	*Improving the Characterization of Clouds, Aerosols and the Cryosphere in Climate Models*	Philip Rasch	Jin-Ho Yoon	12	158
m1204	*Center at LBNL for Integrative Modeling of the Earth System (CLIMES)*	William Collins	William Collins	4.1	268
m1040	*Investigation of the Magnitudes and Probabilities of Abrupt Climate TransitionS (IMPACTS)*	William Collins	William Collins	9.7	608
mp193	*Program for Climate Model Diagnosis and Intercomparison*	James Boyle	David Bader	7.8	960
m1343	*Projections of Ice Sheet Evolution Using Advanced Ice and Ocean Models*	William Collins	Stephen Price	2.3	58
m1441	*Simulations of Anthropogenic Climate Change Using a Multi-scale Modeling Framework*	Cristiana Stan	Cristiana Stan	5.2	10.6
m1178	*Development of Frameworks for Robust Regional Climate Modeling*	Ruby Leung	Ruby Leung	6.2	200
m749	*Hybrid Numerical Methods for Multiscale Simulations of Subsurface Biogeochemical Processes*	Tim Scheibe	Tim Scheibe	3.7	6
Total of projects represented by case studies				**97.1**	**4,816 TB**
NERSC 2012 BER Climate / Environmental Total				**165**	**8,409 TB**
Percent of NERSC Climate represented by case studies				**59%**	**57%**

Bioscience					
m906	*Molecular Dynamics Simulations of Protein Dynamics and Lignocellulosic Biomass*	Jeremy Smith	Loukas Petridas	7.7	47
m926	*Computational Prediction of Transcription Factor Binding Sites*	Harley McAdams	Mohammed AlQuraishi	0.5	0
m342	*Joint Genome Institute - Production Sequencing and Genomics*	Edward Rubin	D. Goodstein / R. Egan / S. Canon	32	1,300
m1045	*Microbial Genome and Metagenome Data Processing and Analysis*	Victor Markowitz	Victor Markowitz		
kbase	*Systems Biology Knowledge Base*	Shane Canon	S. Canon / Brettin	0.017	1
Total of projects represented by case studies				**40**	**1,348 TB**
NERSC AY2012 BER Bioscience Total				**70**	**1,760 TB**
Percent of NERSC BER Bioscience represented by case studies				**57%**	**77%**
Totals					
Total Represented by All Case Studies				**137 M**	**6,164 TB**
All BER at NERSC in 2012				**235 M**	**10,170 TB**
Percent of NERSC BER 2012 Allocation Represented by Case Studies				**58.3%**	**60.6%**

6 Findings

6.1 Summary of Requirements

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 need access to significantly more computational and storage resources to achieve their goals and reach BER research objectives.

a) Researchers attending the review estimate needing 3.5 billion production computing hours in 2017. Normalized to the entire BER production workload this is 6 billion hours, or about 25 times 2012 BER usage at NERSC.
b) The need for permanent archival data storage will continue to increase, reaching more than 100 PB for BER at NERSC by 2017. This is more than 10 times what was stored in 2012.
a) Key BER projects estimate needing a 30-fold increase in their online data storage capacity at NERSC. This translates into a need for 30 PB of permanent disk storage.

6.1.2 Simulation and analysis codes will need to access, read, and write data at a rate far beyond that available today.

a) Data transfer rates from computational systems to disk and long-term storage must increase by a factor of approximately 10 to support the anticipated workloads of 2017. Extrapolating from Hopper's global scratch I/O bandwidth, this translates to a system bandwidth of 700 GB/sec to scratch disk.
b) As simulation sizes grow the need to checkpoint individual computational jobs and output data to disk will increase. I/O system capability must keep pace such that time spent performing I/O does not overwhelm the time spent performing computations. The climate community has a target I/O time of one percent of the total runtime.

6.1.3 Support for high-throughput job workflows is needed.

a) Adequate job throughput is required to support future international climate assessments such as IPCC AR6.
b) For some BER users, effective throughput of many (possibly interdependent) runs is the most important factor for scientific productivity. Long wait times in a batch queue severely limit productivity.
c) Ensembles of runs are required for uncertainty quantification (UQ) and to study models' sensitivity to different choices of parameters.

d) The bioinformatics community needs support for high-throughput computing and complex workflows.

6.1.4 State-of-the-art computational and storage systems are needed, but their acquisition must not interrupt ongoing research efforts.

a) Emerging architectures provide an opportunity for increased scientific productivity, but ongoing computational research must be supported without a lengthy interruption during the transition to new platforms.
b) Access to early prototype and testbed systems are needed to prepare for new architectures while still running production jobs on existing machines.
c) Frequent and effective user training is needed for the transition to new architectures.

6.1.5 NERSC needs to support data analytics and sharing, with increased emphasis on combining experimental and simulated data.

a) Data portals, like the Earth Systems Grid (ESG), for sharing data and simulation results are important to serve large data sets among many users.
b) High-speed external networks are required. Several projects are multi-site collaborations and need increased bandwidth for HPSS to external networks.
c) This community is heavily using NERSC data services and needs dedicated resources such as the NERSC Scientific Gateway Nodes and Data Transfer Nodes to facilitate data movement.
d) There is a need for a standard way to provide data provenance (where it came from, where it was generated, who worked on it, where and by what it was compiled).

6.2 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.

6.2.1 Readiness for next-generation architectures (many-core) varies

Some groups, especially with newer codes, are committed to porting immediately to new programming models to try to take advantage of existing platforms (e.g., using CUDA, OpenACC). Others are waiting for community codes to be ready before moving to new architectures. Still others are “waiting it out” to see which programming paradigm and/or language gains acceptance.

6.2.2 Scientific productivity is the key objective

Enabling and maintaining scientific productivity, while still advancing the state of the art, is required when acquiring new systems and offering new services: “Leading without bleeding.” It’s not just computational power that maters most, it’s support for doing science.

6.2.3 The bioinformatics workload is rapidly increasing

A quickly growing need to support the bioinformatics workload presents a number of challenges to traditional HPC centers. There is a need to support massive numbers of high-throughput and low concurrency jobs, some of which need very long runtimes and/or large-memory architectures.

6.2.4 Uncertainty Quantification (UQ) will play a more prominent role

Quantifying uncertainty in simulations and validating those simulations to much higher precision is becoming required. This will increase the demand for additional compute cycles and better workflow management systems. How this will play out is uncertain, but it has the potential to drastically increase the BER community's aggregate resource requirements.

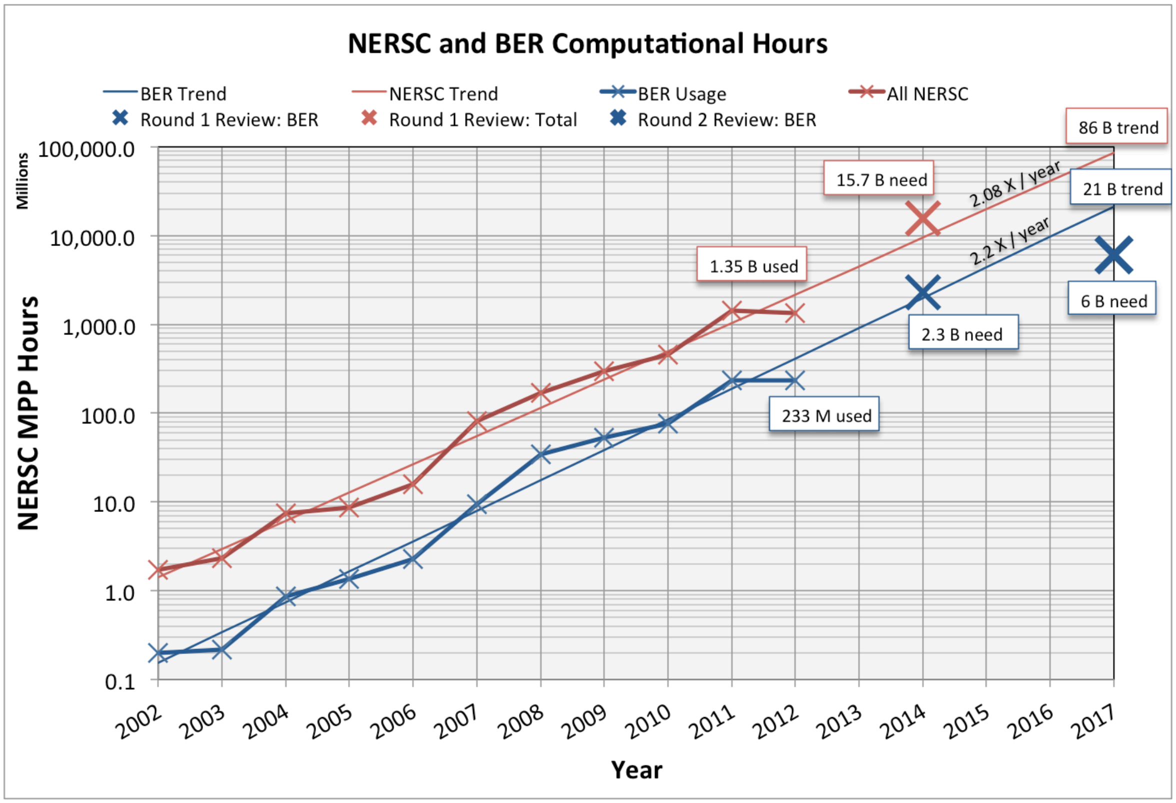
6.2.5 Data and analytics present new challenges

Larger simulations and data sets may require new approaches to data managements, data analytics and scientific visualization. Improvements in hardware may not be adequate to accommodate current I/O methods and in-situ analysis and/or data reduction may be required. The BER community will need assistance to implement these types of analysis methods.

7 Computing and Storage Requirements

The following table lists the 2017 computational hours and archival storage needed at NERSC for research represented by the case studies in this report. "Total Scaled Requirement" at the end of the table represents the hours needed by all 2012 BER NERSC projects if increased by the same factor as that needed by the projects represented by the case studies. The 25-fold increase over 2012 NERSC use does not include the KBase value.

Case Study Title	**PI**	**Hours Needed in 2017**		**Archival Data Storage Needed in 2017**	
		Million Hours	**Factor Increase**	**TB**	**Factor Increase**
Climate Change Simulations with CESM	Washington	1,000	29	30,000	19
Sparse Input Reanalysis for Climate Applications (SIRCA) 1850-2012	Compo	670	57	8,000	8
Improving the Characterization of Clouds, Aerosols and the Cryosphere in Climate Models	Rasch	200	17	2,000	13
CLIMES and IMPACTS	Collins	150	11	4,000	4.6
Climate Science for a Sustainable Energy Future (CSSEF)	Bader	150	19	7,500	7.8
Projections of Ice Sheet Evolution Using Advanced Ice and Ocean Models	Price	156	68	300	5.2
Simulations of Anthropogenic Climate Change Using a Multi-scale Modeling Framework	Stan	55	11	150	14
Development of Frameworks for Robust Regional Climate Modeling	Leung	100	16	3,200	16
Hybrid Numerical Methods for Multiscale Simulations of Subsurface Biogeochemical Processes	Scheibe	120	32	200	33
Molecular Dynamics Simulations of Protein Dynamics and Lignocellulosic Biomass	Smith	360	47	100	2
Computational Prediction of Transcription Factor Binding Sites	McAdams	30	60	0	N/A
Joint Genome Institute - Production Sequencing and Genomics	Rubin / Markowitz	400	12.5	7,500	5.8
Kbase Systems Biology Knowledge Base	Canon	100	6,000	2,000	2,000
Total Represented by Case Studies		**3,491**		**64,950**	
Percent of NERSC BER Represented by Case Studies		**58.3%**		**60.6%**	
All BER at NERSC Total Scaled Requirement		**6,000 M**	**25.5**	**107,000**	**10.5**



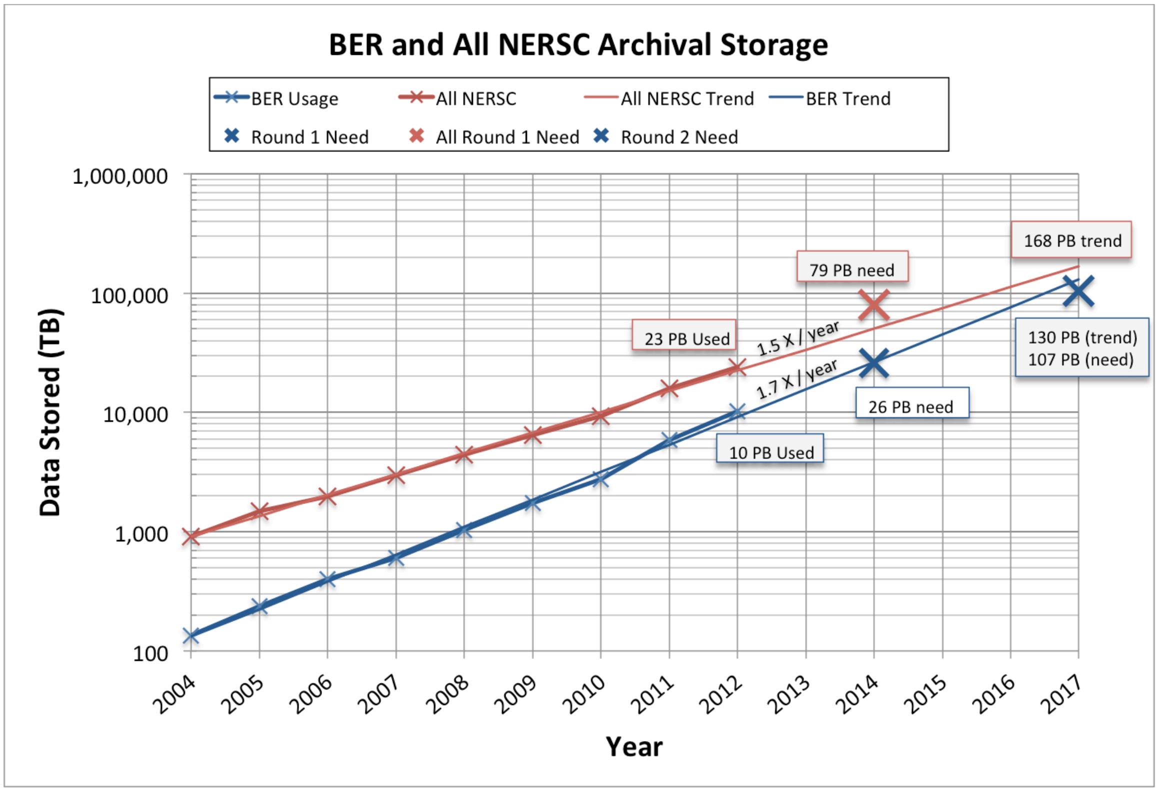


Figure 1 Computational and archival storage usage and needs for the Office of Biological and Environmental Research as well as the sum total of all six DOE Office of Science program offices (All NERSC).

8 Climate and Environmental Science Case Studies

8.1 Overview

Drs. Renu Joseph and Dorothy Koch, Program Managers, Climate and Environmental Sciences Division, DOE

The Climate and Environmental Sciences Division (CESD) focuses on fundamental research that advances a robust predictive understanding of Earth's climate and environmental systems and informs the development of sustainable solutions to the Nation's energy and environmental challenges. As provided by the 2012 CESD Strategic Plan (http://science.energy.gov/~/media/ber/pdf/CESD-StratPlan-2012.pdf), there are five goals which frame the Division's programs and investments: (a) synthesize new process knowledge and innovative computational methods that advance next generation, integrated models of the human-earth system; (b) develop, test and simulate process-level understanding of atmospheric systems and terrestrial ecosystems, extending from bedrock to the top of the vegetative canopy; (c) advance fundamental understanding of coupled biogeochemical processes in complex subsurface environments to enable systems-level prediction and control; (d) enhance the unique capabilities and impacts of the ARM and EMSL scientific user facilities and other BER community resources to advance the frontiers of climate and environmental science; and (e) identify and address science gaps that limit translation of CESD fundamental science into solutions for DOE's most pressing energy and environmental challenges. Leadership-class computing facilities and DOE NERSC are critical for the computationally intensive simulations of high-resolution models needed to address these priorities.

CESD focuses on three research activities, each containing one or more programs and/or linkages to national user facilities. These activities are: (1) The Atmospheric System Research activity which seeks to understand the physics, chemistry, and dynamics governing clouds, aerosols, and precipitation interactions, with a goal to advance the predictive understanding of the climate system; (2) The Environmental System Science activity that seeks to advance a robust predictive understanding of terrestrial surface and subsurface ecosystems, within a domain that extends from the bedrock to the top of the vegetated canopy and from molecular to global scales. 3) The Climate and Earth System Modeling activity which seeks to develop high fidelity community models representing earth and climate system variabilities and change, with a significant focus on the response of systems to natural and anthropogenic forcing.

The primary programs that actively use NERSC facilities are: 1) The Earth System Modeling (ESM) program that develops advanced numerical algorithms to represent the dynamical and biogeophysical elements of the earth system and its components; 2) The Regional and Global Climate Modeling Program which focuses on understanding the natural and anthropogenic components of regional variability and change, using simulations, and diagnostic measures; 3) The subsurface research program whose focus is to develop robust predictive models of subsurface biogeochemical processes to understand the structure and function of complex subsurface systems.

NERSC and other DOE leadership class computational facilities are essential to advance the robust predictive understanding of the earth's climate and environmental systems. For example, NESRC supports computationally intense and long-term simulations from state-of-the-art global climate models. These simulations from global climate and earth system models have contributed model output to all the Intergovernmental Panel on Climate Change (IPCC) reports (reports from 1-5). NESRC resources also contribute extensively to model development efforts of the DOE-NSF jointly funded Community and Earth System Model (CESM), by allowing for development and testing of the various model components. Development of the Atmospheric, Oceanic, Biogechemistry, and Land-and-Sea Ice model CESM components would not be possible without NERSC. In addition, the computational resources needed to understand and quantify the uncertainties in global models (and their individual components) are significant and NERSC resources have contributed extensively to development of uncertainty quantification methods in the Earth system. Modeling and understanding the implications of sea-level rise is another area that requires enormous computer resources because of the length of the simulations needed for capturing ice sheet evolution. An example of a subsurface project utilizing NERSC is Advanced Simulation Capabilities for Environmental Management (ASCEM), in which an integrated multi-scale modeling framework is being developed to link different subsurface flow, transport, and reaction process models at continuum, pore, and sub-pore scales. The Next Generation Ecosystem Experiment (NGEE) Arctic Project will use a similar approach to understand and model the evolution of Arctic permafrost systems. NERSC has played and will continue to play a vital role in enabling modeling and simulation for DOE climate and environmental research.

8.2 Climate Change Simulations with the Community Earth System Model

Principal Investigators: Warren Washington (NCAR)
Case Study Author: Thomas Bettge (NCAR)
NERSC Repository: mp9

8.2.1 Summary and Scientific Objectives

The goals of the Climate Change Prediction (CCP) group at NCAR are to understand and quantify contributions of natural and anthropogenic-induced patterns of climate variability and change in the 20th and 21st centuries by means of simulations with the Community Earth System Model (CESM). With these model simulations, researchers are able to investigate mechanisms of climate variability and change, as well as to detect and attribute past climate changes, and to project and predict future changes. The simulations are motivated by broad community interest and are widely used by the national and international research communities.

The types of fully-coupled CESM simulations conducted by the CCP include simplified forcing experiments, long pre-industrial control runs, large ensembles of 20th-century simulations with various combinations of natural and anthropogenic forcing, and large ensembles of future climate simulations using different emission scenarios. Single-forcing runs, isolating the contributions to climate change of individual natural (e.g., solar and volcano) and anthropogenic (e.g. GHG, ozone, aerosol) forcing, complement the runs with all-inclusive forcing by contributing to studies of climate change detection and attribution. Analyses typically target changes in mean climate and associated uncertainties due to natural variability obtained from the large number of ensemble members, changes in variability and extremes, and changes across collections of ensemble members with different scenarios to assess forcing-related uncertainties. Advancements in high-end computing technology has allowed, and will continue to allow, the use of increased horizontal and vertical grid resolution in both the atmosphere and the ocean to facilitate analysis of regional climate regimes within projected forcing scenarios. With increasing model resolution, we also produce and analyze decadal hindcast and initialized prediction experiments to better quantify time-evolving regional climate change over the next few decades.

The CESM project can be divided into three categories, each with a direct high-end computing requirement:

- **CESM Development and Validation:** Research and Development of scientific processes/methods and computational algorithms requires **easy access** to high-end computing platforms to test, iterate, and validate procedures.

- **Climate Change Prediction using CESM (the focus of CCP)**

To contribute to national and international missions and goals with scientifically and computationally validated CESM versions, the community needs **consistent access** to stable high-end computing platforms for extended production.

• **High Resolution Fully Coupled CESM Simulation:** Testing cutting-edge high-resolution CESM configurations with extended integrations requires **priority access** to high–end computing platforms (sometimes in a pre-release state from general community use) which will allow potential transformations in climate change science in reasonable timeframes.

Historically, the high-end computing resources of the entire CESM project - including research, development, validation, and production - are provided by allocation requests at several HPC sites across several government agencies and/or government-funded organizations. The resources are organized and targeted according to the general mission of the granting agency and the goals outlined in the CESM strategic plan (http://www.cesm.ucar.edu/management). While the core CESM project is managed at NCAR, a large community representing scientists from universities and national laboratories participate in numerous CESM research themes and working groups.

CCP has contributed input to the Atmospheric Model Intercomparison Project (AMIP), Coupled Model Intercomparison Project (CMIP), and Intergovernmental Panel on Climate Change (IPCC), and other projects. Data storage – both near-term and long-term – is an important aspect of analysis for both primary scientists within CCP as well as a broad community to whom access is granted. Subsets of the primary data are distributed via the DOE Earth System Grid (ESG, http://www.earthsystemgrid.org). The CESM Data Management and Data Distribution Plan (2011) (http://www.cesm.ucar.edu/management/docs.html) guides the production, storage, and distribution of data produced by CESM. The overall goal of this plan is to provide the best possible access to, and easiest use of, high-quality CESM data to and by diverse users within the constraints of available resources. *An overarching goal for storage of data produced by CESM is to archive the data, whenever possible, at either the site of generation or its associated data archive center, and thus avoid moving massive amounts of data over a wide area network.*

8.2.2 Scientific Objectives for 2017

Over the next five years we plan to

1) Improve our understanding of many of the component processes represented in the CESM, including cloud physics; radiative transfer; atmospheric chemistry, including aerosol chemistry, boundary-layer processes, polar processes, and biogeochemical processes; and the interactions of gravity waves with the large-scale circulation of the atmosphere;
2) Better understand how these component processes interact;
3) Develop more sophisticated codes to better represent dynamical geophysical fluid processes;
4) Further calibrate our models against improved observations of the atmosphere, including those enabled by major advances in satellite observations.

Models with increased spatial resolution covering longer intervals of simulated time are required to meet these objectives. It is crucial that increasing computer power, both in the U.S. and abroad, be available to support these more elaborate and sophisticated models and studies.

The full suite of CESM development and production plans involves the activities of many working group scientists, especially those of CESM collaborators and working group co-chairs at the DOE National Laboratories. CESM development activities also result from working group participation in activities such as the CLIVAR Climate Process Teams (CPTs) and from involvement of CESM scientists with university collaborators in agency proposals, including the NSF/DOE/USDA call for decadal and regional climate prediction using Earth System Models (EaSM). Development activities are also facilitated by the recent DOE proposal "Climate Science for a Sustainable Energy Future" (CSSEF). This involves the direct participation of several CESM working group co-chairs and focuses on reducing uncertainty and confronting models with observations with the aim of developing the "next generation plus one" version of CESM (i.e., CESM3.0).

Over the next five years general CESM development plans include the following.

- **Coupling across components and understanding interactions:** A key attribute of CESM is the ability to simulate coupled interactions across different components of the climate system, including physical, chemical and biological elements. Proposed development work in this regard focuses on three main aspects: evaluating model performance against observations; understanding the behavior of and refining the representation of physical processes; and expanding capabilities for coupling across components.
- **New parameterizations and processes:** To address the evolving scientific needs of the CESM community, progress demands that new processes be introduced and new parameterizations of existing processes be developed and tested. The incorporation of more earth system components and efforts to run the CESM across a wider range of resolutions, incur unique challenges for parameterization development.
- **High resolution and new dynamical cores:** With increases in computer resources, a societal need for climate information at more regional scales, and scientific questions associated with scale-interactions and high-resolution phenomenon, important development efforts are focused on high-resolution simulations and new dynamical cores that enable these resolutions. These developments are occurring throughout the component models. For the atmosphere, development work will consider global resolutions up to 1/8 degree and regionally refined grids.

Software development

Software development covers three traditional and well-defined tasks: model testing, performance tuning, and debugging. Every combination of model configuration and production machine undergoes on the order of 100 short tests to ensure reliability before being made available for community use. Allocations will be needed for debugging problems as they arise inevitably from systems issues, or from new dynamic capabilities and parameterizations, processor layouts and resolutions. Performance tuning optimizes the number of MPI tasks and OpenMP threads for each CESM component, resolution and targeted processor count. Additionally, work is now beginning on the use of GPU and/or many-core chip architectures in preparation for the next generation supercomputers. The computing resources required for software development should be regarded as a wise investment with a high return in the form of reduced probability of encountering problematic code, centralized debugging by experts, and efficient use of allocated processors.

We anticipate that in 2017 the CESM configuration employed for CCP activities will contribute to continued advancement of climate science knowledge using significant contributions from the above four model development themes. In particular, we anticipate in 2017 the use of an atmospheric horizontal resolution of 0.25^o (versus 1.0^o in 2012). Even by taking into account a more computationally efficient and scalable dynamical core, the compute time needed for a single set of control-historical-future simulations will require on the order of 12 times our present allocations. In addition, we anticipate that more accurate physical processes will be required to simulate the climate system, including explicit cloud resolving models, ice sheet (ocean and land) models, atmospheric chemistry models, and models to simulate land biogeochemical processes. These additional components could increase the cost by another factor of two, or 24 times our present allocation.

Fortunately, we are optimistic that increasing computer power, both in the U.S. and abroad, will be available to support more elaborate and more sophisticated models and modeling studies, using increased spatial resolution and covering longer intervals of simulated time. We anticipate that standard CESM climate change production simulations will use O(20,000-50,000) processors, versus O(2,000) in 2012, and that GPU cores will be used for the most intensive parts of the computations.

8.2.3 Computational Strategies

8.2.3.1 Approach

The Community Earth System Model (CESM) is a coupled climate model for simulating Earth's climate system. Composed of five separate models simultaneously simulating the Earth's atmosphere, ocean, land, land-ice, and sea-ice, plus one central coupler component, CESM allows researchers to conduct fundamental research into the Earth's past, present, and future climate states.

The CESM system can be configured several different ways from both a science and technical perspective. CESM supports several different resolutions and component configurations. In addition, each model component has input options to configure specific model physics and parameterizations. CESM can be run on many different hardware platforms and has a relatively flexible design with respect to processor layout of components. CESM also supports both an internally developed set of component interfaces and Earth System Modeling Framework (ESMF) compliant component interfaces.

The CESM coupled model software is based on a framework that divides the complete climate system into component models that are connected by a coupler component. The coupler controls the execution and time evolution of the complete system by synchronizing and controlling the flow of data between the various components. It also communicates interfacial states and fluxes between the various component models while ensuring the conservation of fluxed quantities. While the primary models can be treated as standalone software components when removed from the CESM software stack, the coupler is implemented as a single executable and is the main program for the entire coupled model. It arranges the component models to run sequentially, concurrently, or in some mixed sequential/concurrent layout, and performs flux calculations, mapping (regridding), diagnostics, and other calculations. Its functions can be run on a subset of the total processors. Each primary model can be configured as *active* (provides prognostic boundary information to the coupler) or *data driven* (provides climatological or steady-state boundary information to the coupler).

8.2.3.2 Codes and Algorithms

The CESM consists of a system of five geophysical component models: atmosphere, land, ocean, sea ice, and land ice. A land ice model has recently been introduced into CESM, but it is currently used at low resolution and has little effect on model performance, so is not discussed here.

CAM

The Community Atmosphere Model (CAM) is characterized by two computational phases: the dynamics, which advances the evolutionary equations for the atmospheric flow, and the physics, which approximates subgrid phenomena such as precipitation processes, clouds, long- and short-wave radiation, and turbulent mixing. Separate data structures and parallelization strategies are used for the dynamics and physics. The dynamics and physics are executed in turn during each model simulation time step, requiring some data motion between the two data structures each time step.

CAM includes multiple compile-time options for the dynamics, referred to as dynamical cores or "dycores." The default dycore is a finite-volume method (FV) that uses a tensor-product latitude × longitude × vertical-level computational grid over the sphere. The parallel implementation of the FV dycore is based on two-dimensional tensor-product "block" decompositions of the computational grid into a set of geographically contiguous subdomains. A latitude-vertical decomposition is used for the main dynamical algorithms and a latitude-longitude decomposition is used for a Lagrangian surface remapping of the vertical coordinates and (optionally) geopotential calculation. Halo updates are the primary MPI communications required by computation for a given decomposition. OpenMP is used for additional loop-level parallelism.

CAM physics is based on vertical columns and dependencies occur only in the vertical direction. Thus, computations are independent between columns. The parallel implementation of the physics is based on a fine-grain latitude-longitude decomposition. The computational cost in the physics is not uniform over the vertical columns, with the cost for an individual column depending on both geographic location and on simulation time. A number of predefined physics decompositions are provided (selected by the user at compile time) that attempt to minimize the combined effect of load imbalance and the communication cost of mapping to/from the dynamics decompositions.

Transitioning from one grid decomposition to another, for example, latitude-vertical to latitude-longitude or dynamics to physics, may require that information be exchanged between processes. If the decompositions are very different, then every process may need to exchange data with every other process. If they are similar, each process may need to communicate with only a small number of other processes (or possibly none at all).

Common performance optimization options include:

- Number of OpenMP threads per process;

- Number of processes to use in the dynamics latitude-vertical decomposition, in the dynamics latitude-longitude decomposition, and in the physics latitude-longitude decomposition (none of which need to be the same);

- For a given process count, the two-dimensional virtual processor grid used to define a dynamics decomposition;

- Physics load balancing option (and decomposition); and
- MPI communication algorithms and protocols used for each communication phase, e.g., halo update or potentially nonlocal communication operators for mapping between decompositions.

These represent a large number of options, and other options are available for special cases; for example, parallelizing over tracer index when advecting large numbers of tracers. Fortunately, reasonable defaults have been identified for most of these, and optimization begins from a reasonable initial set of settings.

POP

The Parallel Ocean Program (POP) approximates the three-dimensional primitive equations for fluid motions on a generalized orthogonal computational grid on the sphere. Each timestep of the model is split into two phases. A three-dimensional “baroclinic” phase uses an explicit time integration method. A “barotropic” phase includes an implicit solution of the two-dimensional surface pressure using a preconditioned conjugate gradient (CG) solver.

The parallel implementation is based on a two-dimensional tensor-product “block” decomposition of the horizontal dimensions of the three-dimensional computational grid. The vertical dimension is not decomposed. The amount of work associated with a block is proportional to the number of grid cells located in the ocean. Grid cells located over land are “masked” and eliminated from the computational loops. OpenMP parallelism is applied to loops over blocks assigned to an MPI process. If specified at compile time, the number of MPI processes and OpenMP threads will be used to choose block sizes such that enough blocks are generated for all computational threads to be assigned work. The block sizes can also be specified manually.

The parallel implementation of the baroclinic phase requires only limited nearest-neighbor MPI communication (for halo updates) and performance is dominated primarily by computation. The barotropic phase requires both halo updates and global sums (implemented with local sums plus MPI_Allreduce with a small data payload) for each CG iteration. The parallel performance of the barotropic phase is dominated by the communication cost of the halo updates and global sum operations.

Two different approaches to domain decomposition are considered here: “Cartesian” and “spacecurve.” The Cartesian option decomposes the grid onto a two-dimensional virtual processor grid, and then further subdivides the local subgrids into blocks to provide work for OpenMP threads. The spacecurve option begins by eliminating blocks having only “land” grid cells. A space-filling curve ordering of the remaining blocks is then calculated, and an equipartition of this one-dimensional ordering of the blocks is used to assign blocks to processes.

The common performance optimization options are:

- Number of OpenMP threads per process;
- Choice of Cartesian or spacecurve decomposition strategy; and
- Block size

CLM

The Community Land Model (CLM) is a single column (snow-soil-vegetation) model of the land surface, and in this aspect it is embarrassingly parallel. When using the FV dycore in the atmosphere model, CLM typically uses the same horizontal computational grid as the atmosphere. However, CESM supports the option of CLM using a totally different grid.

Spatial land surface heterogeneity in CLM is represented as a nested subgrid hierarchy in which grid cells are composed of multiple landunits. Landunits are composed of multiple snow/soil columns and snow/soil columns are composed of multiple plant functional types (PFTs). Grid cells are grouped into blocks of nearly equal computational cost, and these blocks are subsequently assigned to MPI processes.

When run with MPI-only parallelism, each process has only one block. When OpenMP is enabled, the number of blocks per process is by default set to the maximum number of OpenMP threads available. This number can be overridden at runtime.

A single load balancing algorithm is implemented that has proven to work well across a variety of computer architectures and problem specifications. At the current time, the common performance optimization options are:

- Number of OpenMP threads per process; and
- Number of grid cell blocks assigned to each process.

The default of one block per computational thread is typically optimal. Moreover, for a fixed core count, MPI-only often outperforms hybrid MPI/OpenMP runs.

CICE

The Community Ice Code (CICE) sea ice model is formulated on a two-dimensional horizontal grid representing the earth's surface, typically using the same horizontal grid as POP. An orthogonal vertical dimension exists to represent the sea ice thickness. Similar to POP, the parallel implementation decomposes the horizontal dimensions into two-dimensional blocks. The vertical dimension is not decomposed. CICE exploits MPI and OpenMP parallelism over the same dimension, namely grid blocks. Currently, the CICE decomposition is static and set at initialization. Like POP, a block size will be chosen based on the total number of computational threads, or a block size can be specified manually.

The relative cost of computing on the sea ice grid varies significantly both spatially and temporally over a climate simulation because the sea ice distribution is changing constantly. This has a huge impact on the load balance of the sea ice model in a statically decomposed model. The load balance will be generally optimized if grid cells from varied geographical locations are assigned to each process. CICE also performs regular and frequent halo updates with a resultant performance cost that also depends on the assignment of grid cells to processes.

Optimal static load balance is achieved by balancing the computational load imbalance and the communication cost of halo updates. The common performance optimization options are:

- number of OpenMP threads per process;
- choice of Cartesian or spacecurve decomposition strategy; and
- block size.

Coupler

The CCSM coupler is responsible for several actions including rearranging data between different process sets, coordinating the interaction and time evolution of the component models, interpolating (mapping) data between different grids, merging data from different components, flux calculations, and diagnostics. Many of the algorithms are trivially parallel and require no communication between grid cells. Two-dimensional boundary data (flux and state information) are exchanged periodically through the coupler component.

The coupler receives grid information in parallel at runtime from all of the model components. Domain decompositions are determined on the fly based upon the model resolutions, the component model decompositions, and the processors used by the coupler. Both rearrangement and mapping require interprocess communication, and the choice of MPI communication algorithm and protocol used to implement these affect performance. The number of options is small compared to those in CAM currently, but this may increase in the near future. Performance depends primarily on the number of processes assigned to each of the components, including the coupler, and the placement of these components relative to the coupler processes. The coupler is the one component that cannot be optimized separately from the full CESM.

To summarize, the performance optimization options are:

- MPI communication algorithms and protocols used in transferring data to and from the geophysical components; and

- number and layout of processes used for each component.

OpenMP parallelism has been introduced in a development version of the coupler, but is not used because coupler performance is not a limiting factor in CESM performance.

8.2.4 HPC Resources Used Today

8.2.4.1 Computational Hours

Table 1 shows the number of hours used for climate related activities of our project. We have attempted to normalize the units in the table (millions of hours) to a Hopper processor, as shown in the caption. Total CESM usage is the aggregate of research, development, and testing of all CESM components, including, for example, validation of new parameterizations and processes, and the highest resolution, cutting-edge simulations. In addition, we have provided an estimation of the CPP-only usage attributed to climate change simulations that we have used to contribute to national and international programs for climate model intercomparisons, climate change detection/attribution, and future climate change projections. The CCP-only, which dominates the NERSC usage, was performed with the current validated version of CESM that allows long simulations with reasonable turnaround to meet report and project timelines.

Site	2011 Usage All CESM	2011 Usage CCP-only	2012[3] Usage All CESM	2012[3] Usage CCP-only
NERSC (mp9)	32.5	22.1	23.8	15.3
OLCF	48.0	8.8	37.5	8.7
ALCF	10.8	0	4.0	0
NCAR CSL	30.2	6.3	17.6	3.7

Table 1. Supercomputer Usage for Climate Change Simulations with the Community Earth System Model (CESM) Today. Usage is provided for 2011 and Jan-Jul 2012. Units are Millions of hours, based upon the following conversions to equivalent NERSC processor hours (Cray XE6:other machine): NERSC (1:1), OLCF (1:1), ALCF (1:4), NCAR (4:1). The NCAR CSL PE-hours are for the NCAR/CESM project only. CCP uses approximately 20% of this allocation. The Climate System Laboratory (CSL) at NCAR provided over 60M pe-hours in 2011 and 2012 for a broad range of CSL (CESM and non-CESM) projects.

8.2.4.2 Compute Cores

The goal of the climate change and variability simulations performed by our group at NERSC is to use scientifically validated and publically released versions of CESM at resolutions that allow long earth system simulations to advance climate change science in a reasonable timeframe. For example, deadlines exist for reports at annual meetings, targeted journal publications, and ultimately the upcoming IPCC report. In simple terms, this means that the version of CESM must use earth system dynamical and physical components that have been scientifically vetted, thoroughly tested, and are well behaved, and it must use initial datasets that have been flux balanced via very long control integrations in a comprehensive configuration that balances high performance and a sufficient production rate. For example, a suitable production time for a century-long simulation is less than one wallclock month. The actual time depends upon the number of ensemble members needed for statistical significance, the desired length of each experiment, resources available, number of cores used, queue wait time, and other factors.

A standard (typical) CESM production simulation at NERSC today consists of five components – atmosphere, ocean, land, ice, and the coupler – at a horizontal resolution of 1°. Each job of this type uses a maximum of 2,064 processors within a single executable. The component models are distributed on the processors in a fashion that achieves both optimal performance and optimal load balance. Some of the models run in parallel on separate subsets of the requested processors, while others run sequentially. If the number of processors is changed, the performance and load balance is changed accordingly to achieve the most effective use of the machine. On 2,064 processors, the production rate for this model is 10 model years per wallclock day. Coarse parallelism on Hopper is achieved by running multiple simulations within a single job submission. With the above model

[3] January through July. Usage for all of 2012 at NERSC was 34.4 million hours.

configuration we normally gang together several simulations in order to receive the discount applied to jobs that run on a large number of processors.

Because of our project's goals outlined above, we rarely use more than 2,064 processors for a single simulation. The two restrictions that limit the number of processors we typically employ are (1) use of a certified version of CESM with an acceptable initial control dataset, and (2) lack of code scalability. The culprit of both these restrictions is the 1° horizontal resolution. In a nutshell, CESM demonstrates scalability at higher resolutions, but at higher expense (due to increased resolution), which in turn means that either the model has not been validated or no long control simulation exists. Indeed, 2,064 is the current sweet spot.

8.2.4.3 Checkpointing

The CESM model is designed to create and write a checkpoint file at regular intervals during a job submission. Generally, these files are created at subintervals of the expected total job runtime as well as upon normal termination of job. For the configuration described in section 3.2 above, the restart dataset size is 4.1 GB. Our model timing statistics show that the creation of a checkpoint file is 1% of the total runtime.

8.2.4.4 Data and I/O

If we define a "run" as an eight-member ensemble within a single job submission (as described in section 3.2), then a single 24-hour wallclock job produces 80 simulation years. Each simulation year produces 30 GB of data, so a single submission writes 2.4 TB of data. According to the output from the pyLMT web portal, the maximum (burst) bandwidth to write data within the CESM is 4.5 GB/sec. Given that the data are written once per simulated month, this means that 2.5 GB of data are written every 72 seconds. At a burst rate of 4.5 GB/sec, the I/O required by this model is less than 1% of the total compute time, which is consistent with the CESM software's internal timers.

8.2.4.5 Project Data

We currently make frequent use of a project directory, "ccsm1," that has an extended quota of 10TB and about 3 TB stored in it.

8.2.5 HPC Requirements in 2017

8.2.5.1 Computational Hours Needed

In 2017 we anticipate that the standard validated CESM for performing climate change prediction simulations will use a 0.25° CAM and 1.0° POP. Other components will be of similar, consistent resolution. A single standard (minimum) climate change simulation experiment consists of a 156-year historical simulation (1850-2005), and four future 100-year scenarios. The 556 years are repeated to create an ensemble for statistical variability analysis using a minimum of five ensemble members. The total number of years for an experiment is thus (156 + 4*(100))*5 = 2780 years. Table 2 summarizes the compute hours needed for a complete experiment using the current production CESM and the version expected in 2017.

Resolution	Simulated Years Required	Cost per Simulated Year (hours)	Data Storage per Simulated Year (GB)	Total Cost (hours)	Total Storage Required (TB)
1º (2012)	2780	5,000	30	14 M	84
0.25º (2017)	2780	60,000	500	167 M	1,400
0.25º (2017)*	2780	120,000	700	334 M	1,946

Table 2. Cost of a typical end-to-end climate change experiment for 2012, and anticipated by 2017. Note that total cost for a single experiment represents about half the total historical need and award given to our project.

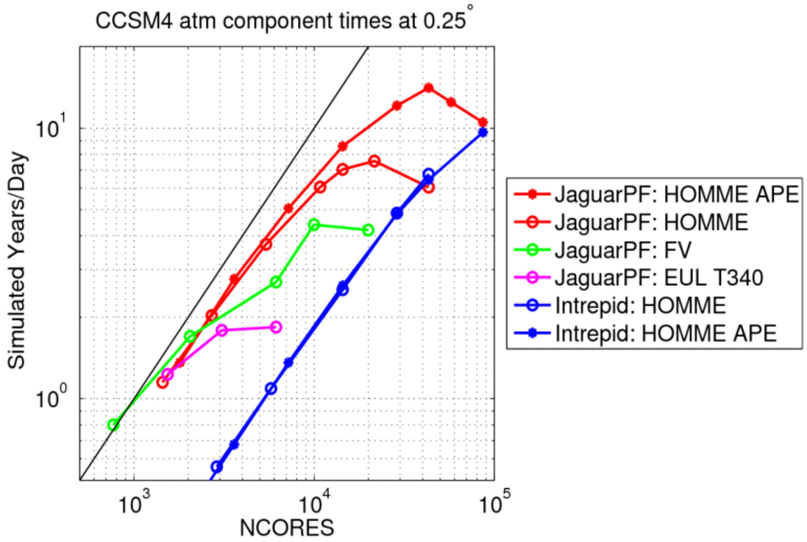
*The bottom row represents a model experiment with improved representation of several climate system components, including cloud resolving models, complex ice sheets, and significantly upgraded chemistry and biogeochemistry processes.

For simplicity the numbers provided above are qualified estimates of a typical climate change experiment. In practice, we rarely produce an end-to-end experiment at a single site in a single year because of (a) expense involved; (b) the total wallclock required (any single experiment is normally performed over a 1-2 year period); (c) the ability to usher the jobs through the runtime queues; and/or (d) the desire to diversify the science being accomplished. Our historical request and usage at NERSC reflects a range of climate science research beyond a single end-to-end climate change experiment. In fact, the usage history shows that our allocation is nearly evenly split between an end-to-end experiment and other climate change (detection/attribution, decadal predictability, etc.) experiments. Thus, the cost-per-simulated-year for a state-of-the-art climate model (in 2012 or 2017) in Table 2 provides a suitable cost-estimate model for anticipated climate change research requirements. In 2008, our project was awarded 1.3 M hours, and in 2013 we anticipate an award of ~30 M hours. Thus, by extrapolation the 20 M hours we received in 2012 would project ***a request/award of 400-500 M hours in 2017***. This estimate is consistent with both past usage and the scientific goals for the 0.25º production CESM in 2017. ***If anticipated improvements in earth system modeling are realized, the estimated need in 2017 could increase by a factor of two, to 800-1000 M hours.***

We anticipate that significant resources will be available elsewhere, most likely in proportion to Table 1. If accurate, the total time available at the sites listed in Table 1 is in the vicinity of one billion hours. It is difficult to anticipate the availability and accessibility of additional resources for climate science research.

8.2.5.2 Number of Compute Cores

With the standard CESM climate change simulation model configuration in 2017, we anticipate using upwards of 20,000-30,000 processor cores. Figure 1 shows the current scaling performance of the 0.25º CAM model on the Cray XT5 (Jaguar) at the OLCF. The fully coupled model will be rate limited by the CAM model, and thus the number of cores required will likely not exceed 10% greater than that required by CAM. Use of available GPU technology by CAM is currently under development, and we anticipate that the overall performance of portions of CAM will improve dramatically, although it is unknown how enhanced GPU performance will affect scaling.



When performing an ensemble of simulations to meet the goals of the science, jobs can be run concurrently including several experiments in a single job. Presently, we gang together enough experiments in a single submission on Hopper to fulfill the threshold of a discount queue charge factor for use of greater than 16K processors.

8.2.5.3 Checkpointing

The checkpoint file sizes will be proportional to horizontal resolution, and therefore will increase in size from 4 GB to 64 GB. Our goal is to maintain keeping the checkpoint file creation to less than 1% of the total runtime for an extended production job.

8.2.5.4 Data I/O

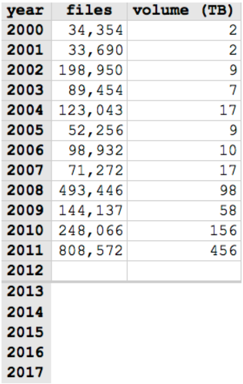
The amount of data read for a CESM climate simulation is minimal, usually only two-dimensional boundary fields representing the annual cycle of a physical quantity. The amount of data written will be proportional to the horizontal resolution increase – 500 GB per simulated year. As with checkpoint file writing, we also have a goal of keeping the data I/O volume to a level where it is 1% of the runtime. Using this goal we estimate the bandwidth needed in 2017 is 42.5 GB/sec. Obviously, if we relax this requirement by a factor of two, the need is ~20 GB/sec.

8.2.5.5 Project Data

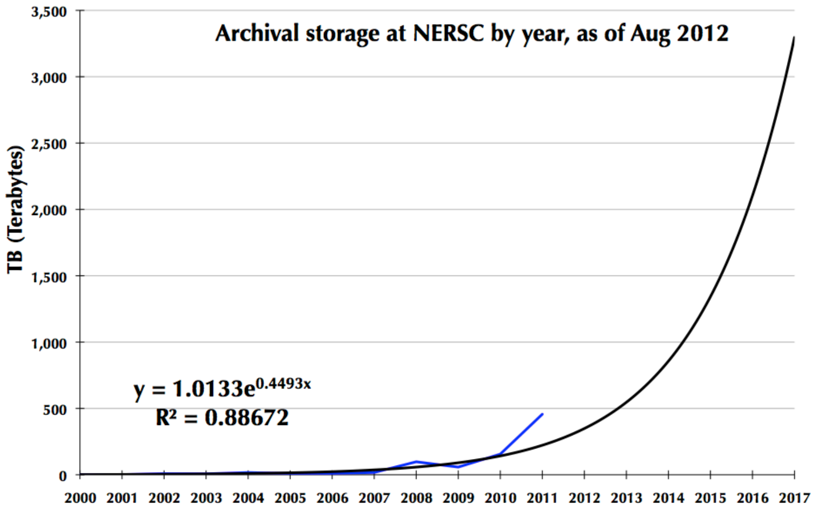
It is likely that the required space will scale as the increase in horizontal resolution of the production model. Currently, the ccsm1 directory has a 10 TB quota, with usage of 4 TB. Thus, in 2017 we would anticipate a usage of 64 TB (factor of 16 for resolution increase), so a 100 TB quota is reasonable.

8.2.5.6 Archival Data Storage

We can estimate the archival storage requirements using two methods. First, a 2012 end-to-end production experiment requires 84 TB (Table 2); currently we have 1,500 TB on the HPSS. Thus, at any time we could have as much as 15X (or more) on the HPSS as is required for a current end-to-end experiment. In 2017, a single end-to-end experiment is estimated to require 2,000 TB. Applying 15X, the requirement would be 30,000 TB, ***some 20 times the 2012 value***. As shown in the projection given in Figure 2 from past HPSS use, we expect to use 3,500 TB in 2017 following business as usual. Given the intensive dataset sizes for the high-resolution simulations, we would be forced to actively attempt to control the archive size unless NERSC is able to accommodate our accelerating data needs.



year	files	volume (TB)
2000	34,354	2
2001	33,690	2
2002	198,950	9
2003	89,454	7
2004	123,043	17
2005	52,256	9
2006	98,932	10
2007	71,272	17
2008	493,446	98
2009	144,137	58
2010	248,066	156
2011	808,572	456
2012		
2013		
2014		
2015		
2016		
2017		



8.2.5.7 Memory Required

Because the CESM disperses the component models onto both shared and non-shared nodes, the minimum value is difficult to determine and may not be useful. The main driver is the CESM-reported "pes min memory highwater (MB) 249.003" for a standard job. Historically, because of careful use of global arrays (minimizing them), the CESM needs cores for performance before the need of cores for memory. In other words, while memory per core is an important metric, sustained FLOPS per core is even more important.

8.2.5.8 Many-Core and/or GPU Architectures

In preparation for using the hybrid GPU architecture of Titan (Cray XK7) at OLCF, the AMIP version of CESM using CAM-SE has been designated a benchmark for acceptance of the machine. An OLCF team of application readiness software engineers has redesigned the SE dynamics for efficient use of the GPU hardware, and has achieved an overall 2.6x speedup of the CAM-SE dynamics. We are encouraged by these results and work will continue into 2013. CAM physics have not yet been evaluated for GPU use. CESM software engineers are beginning to look at designing a GPU enabled version of POP.

8.2.5.9 Software Applications and Tools

Beyond standard libraries which currently exist at NERSC, it is important that CESM have available the Parallel I/O Library[4], the NCAR Command Language (NCL) for post-processing, and the NetCDF Operators (NCO).

8.2.5.10 HPC Services

Obtaining information from our production staff (and software engineers) about items such as I/O bandwidth and memory usage has been difficult (and is still in progress). We would like a tool that could be executed at the end of a job to provide information about memory use, per node and aggregate.

8.2.5.11 Time to Solution and Throughput

As in the past, with a large allocation, and because our production jobs are serial (because of forward time integration), it is important at times to have access to a rapid resubmit queue as opposed to waiting in the standard queue. NERSC has been sensitive to this need, and with their help we have achieved high turnaround rate when needed for crucial, time critical simulations.

8.2.6 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Million)	34.4	1,000
Typical number of cores* used for production runs	2,000	20,000-30,000
Maximum number of cores* that can be used for production runs	2,000	30,000-50,000
Checkpoint data written per run	4 GB	64 GB
Checkpoint bandwidth needed	4.5 GB/sec	42.5 GB/sec
Data read and written per run (excluding checkpoint data)	30 GB	500 GB
Maximum I/O bandwidth (excluding checkpoint data)	4.5 GB/sec	42.5 GB/sec
Project directory space	10 TB	100 TB
Archival data	1,542 TB	30,000 TB

[4] http://web.ncar.teragrid.org/~dennis/pio_doc/html

Minimum memory per node	0.25 GB	0.25 GB

8.3 20th Century Reanalysis

Principal Investigator: Gilbert P. Compo (University of Colorado CIRES and NOAA Earth System Research Laboratory)
NERSC Repository: m958

8.3.1 Overview and Context

To have confidence in projected changes of weather and climate extremes in the 21st century, climate model simulations must be able to reproduce daily historical records of such changes throughout the Nineteenth and Twentieth centuries. However, this calibration of models is not possible directly because satellite measurements only commenced in the 1970s and upper-air observing records extend back to just the 1920s. The 20th Century Reanalysis is the first attempt to address this issue in detail, seeking to reconstruct the state of the Earth's weather and climate every six hours dating back to 1871.

This reanalysis will create a record that reaches far enough into the past to reveal processes that have shaped natural climate cycles, like the El Niño Southern Oscillation, as well as the drivers of man-made climate change. The project does so by eschewing satellite observations, relying only on monthly sea surface temperature and daily pressure data to reconstruct the weather and climate in six-hour chunks from 1871 to 2010.

Any such daily data must also have quantified estimates of uncertainty to allow a fair assessment of the simulations. The 20th century reanalysis dataset permits such a quantitative evaluation. More broadly, understanding climate variability and change requires global daily data to put current extreme weather and climate in a historical context. The reanalysis dataset accounts for the uncertainty of the time-changing observation network by running 56 different numerical weather prediction model simulations for each six hour period, requiring millions of compute hours.

The second version of the dataset produced on supercomputers at NERSC and Oak Ridge National Lab (Franklin and Jaguar) spanned 1871 to 2010. It has already been used to better understand the US Dust Bowl, heat waves, and cold spells. It has also been used to detect previously undocumented hurricanes and improve hurricane predictions in the North Atlantic. Studies have shown trends and variations in the major modes of atmospheric variability.

The next version of the dataset will be produced in partnership with Texas A&M University and will include a reconstruction of the ocean state using the Simple Ocean Data Assimilation system. With recent and expected future increases in computing power, in 2017 the resolution of this combined Ocean Atmosphere Reanalysis for Climate Applications dataset should be 12 times higher and span 1830 to 2017. It will require hundreds of millions of compute hours to generate 64 global, three-dimensional estimates of the atmosphere and ocean.

8.3.2 Scientific Objectives for 2017

While the 20th Century Reanalysis has made progress possible in understanding important changes of extreme weather and climate, improved understanding of the observed variations in hurricanes and tropical cyclones, severe storms and floods, droughts and heat waves, and small polar storms will require computer models with significantly higher spatial resolution. At least 4 times the horizontal resolution and 3 times the vertical resolution will be needed to achieve this. Approximately 300 times the computer power is required to make such improvements. Additionally, extending the dataset back to allow maximal use of the observational record back to the 1830's alone will need 50% more computer time than used in the 20CR.

Climate change studies are increasingly focused on moving beyond understanding and predicting global scale changes to regional scale changes, especially changes in the statistics of extreme weather and droughts. With this new version and the accompanying ocean reanalysis, more detailed comparisons will be possible between climate model simulations and these observational estimates of the extreme weather and climate events that have severe socio-economic consequences. Evaluating the models used to make projections of the influence of humans on the climate and weather requires such detailed data constructed with the state-of-the-art methods that make maximal use of the available observations.

8.3.3 Computational Strategies

8.3.3.1 Approach

The 20th Century Reanalysis is generated using an Ensemble Kalman Filter (EnKF) technique. Our implementation of the EnKF involves simultaneously running 56 short-term forecasts of the winds, temperature, and humidity from the surface of the earth to the stratosphere to serve as a first guess at the state of the weather at a particular time. The EnKF then optimally updates that first guess state with the available surface and sea level pressure observations. This updated state is called an analysis. We can run several decades of analyses simultaneously using 1000s of cores.

8.3.3.2 Codes and Algorithms

ensda_gfs_psonly.x: The data assimilation system is based on estimation theory, which combines aspects of statistical, dynamical systems, and signal processing theory. The goal is to estimate the state of a system (in this case the atmosphere) given measurements, a dynamical model of the system, and estimates of the errors in both the measurements and the model. In our case, the error dynamics are approximately linear, and the errors are approximately Gaussian distributed, so that the optimal solution to the estimation problem is the Kalman Filter. The Ensemble Kalman Filter Data Assimilation (EnsDA) system is a monte-carlo approximation to the full Kalman Filter.

global_fcst: The atmospheric dynamical model is a numerical weather prediction model, which is based on a discretization of the compressible Navier-Stokes equation on a sphere, with parameterizations for unresolved physical processes, such as boundary-layer turbulence and moist convection. Linear terms of the model integration are evaluated in the space of spherical harmonics. Total wavenumbers (T) of 62 have been used for the 20th Century Reanalysis. The discretized Navier-Stokes equation is transformed with the spherical harmonic transform (Legendre transform and Fast Fourier Transform). Nonlinear terms are evaluated in the model grid space on the sphere. The model uses semi-implicit integration of a coupled set of partial nonlinear differential equations. It also uses linear interpolation and Fast Fourier Transform algorithms.

Simple Ocean Data Assimilation uses the Parallel Ocean Program ocean climate model as the dynamical model of the system. A simplified form of the model error is used to combine sea surface temperature observations, the 20th Century reanalysis estimates of heat, water, and momentum fluxes to the ocean, and the ocean state.

8.3.4 HPC Resources Used Today

8.3.4.1 Computational Hours

The project used approximately 12 million (NERSC MPP) hours in 2012 at NERSC.

8.3.4.2 Compute Cores

For 20CR: Every 5 years of sequential time of reanalysis (a stream), say for 1901 to 1905, or 1906 to 1910, …, 1991 to 1995, can be run as either a single job or as part of a large job of many 'streams'. For example, 1901, 1906, 1911,…, 1991 can all be run at the same time. Running many years simultaneously as a large job is much more efficient from a job management perspective. The scaling for this is nearly perfectly linear and 1 to 1. Each typical stream uses 336 cores. At 4X the number of cores (1,344) there is a 50% speed up.

Over the course of the year, I have had 3-5 simultaneous jobs using 336 to 3,360 cores with one using 13,440 cores for a few weeks to test the scaling of running 10 streams simultaneously using 1,344 cores per stream.

For the new atmospheric data assimilation system we are developing (20CR version 3, double the horizontal resolution of 20CRv2 and almost 3 times the vertical resolution), we use 1,536 cores per stream and will start to parallelize to use multiple streams in one job submission.

In 2013, we expect to use 1,536 cores per stream and reanalyze 32 streams spanning 1,850 to 2,013. This could use 49,152 cores simultaneously.

For Simple Ocean Data Assimilation: We are currently using 2,400 cores on hopper. We can also utilize 4,800 but this is starting to enter the region where our code does not scale linearly (we lose about 20% efficiency). We do not run multiple jobs concurrently as we are running for long sequential integrations of an ocean model.

We usually have four to five jobs running concurrently, because the turnaround time in the queues is faster. We would prefer to run one or two larger jobs.

8.3.4.3 Data and I/O

Each run generates about 1 TB of data. We're currently using close to 100 TB on permanent "project" file space at NERSC and close to 1 PB of archival storage. About 10 minutes of every hour of run time is taken up performing I/O to checkpoint and data files.

8.3.5 HPC Requirements in 2017

8.3.5.1 Computational Hours Needed

670 Million Hours

8.3.5.2 Number of Compute Cores

Using 128 cores per ensemble member * 64 members * 10 streams (simultaneously running reanalysis years) = 81,920 cores.

3 jobs currently using 81,920 cores would probably be ideal, but one job using the full amount would be possible also.

The model should scale up to 254 cores per ensemble member * 64 members * 32 streams (simultaneously running reanalysis years) = 520,192 cores

8.3.5.3 Data and I/O

Our I/O requirements per run will increase by a factor of about 4, to 4 TB. We will need about 800 TB of "project" data storage (fast access) and 8 PB of archival storage.

8.3.5.4 Memory Required

We anticipate needing up to 16 GB of shared memory per node and 1 TB of memory globally.

8.3.5.5 Many-Core and/or GPU Architectures

We are waiting for a programming model that supports using compiler directives to program for many core architectures.

8.3.5.6 Software Applications and Tools

netCDF4, python, IDL, climate data operators, NCL

Intel compiler

Science gateways and OPeNDAP

8.3.5.7 HPC Services

We will need a system to post-process data from our runs. Today Carver serves this purpose well.

We expect to continue extensive use of the NERSC Science Gateway – both HPSS and online OPeNDAP. There have already been several papers published by our direct collaborators and by other groups using the OPeNDAP access.

We will need consulting help to move to many-core architectures.

8.3.5.8 Time to Solution and Throughput

If the 20CR and SODA systems, or the combined Ocean Atmosphere Reanalysis for Climate Applications, are supporting the IPCC, then deadlines would be commensurate with that.

8.3.5.9 Data Intensive Needs

We will need sufficient I/O bandwidth to support our analysis project. We will have to sustain I/O rates on the order of many GB/s. This is extremely important to our project.

8.3.6 Requirements Summary

	Used at NERSC in 2012	Needed at NERSC in 2017
Computational Hours (Million)	11.7	670
Typical number of cores* used for production runs	2,016	81,920
Maximum number of cores* that can be used for production runs	30,912	520,192
Checkpoint data written per run	0.8 TB	3.2 TB
Checkpoint bandwidth	0.4 GB/sec	1.6 GB/sec
Data read and written per run (excluding checkpoint data)	0.12 TB	0.8 TB
Maximum I/O bandwidth (excluding checkpoint data)	400 MB/sec	3.2 GB/sec
Project directory space	90 TB	800 TB
Archival data	1,005 TB	8,000 TB
Minimum memory per node	0.024 GB	16 GB
Aggregate memory	0.008 TB	1 TB

8.4 Improving the Characterization of Clouds, Aerosols and the Cryosphere in Climate Models

Principal Investigator: Philip J. Rasch (PNNL)
Case Study Author: Jin-Ho Yoon (PNNL)
NERSC Repository: m1199

8.4.1 Overview and Context

Our overarching goals are (i) to improve the representation of aerosols, clouds, and their interaction in global climate and Earth system models, and then (ii) to use the resulting improved tool to characterize the impact of changing aerosol emission (both past and future) on climate. Our tools of choice are the CAM5 and CESM community models. We use a variety of other approaches in the project that also require substantial computational resources, including: 1) Cloud Resolving Models (CRMs) and Large Eddy Simulation (LES) models (running at very high resolution over limited domains) 2) a process-oriented and very costly multi-scale model (the PNNL-Multiscale Modeling Framework, MMF) to provide insight into ways of improving CAM (costing approximately 300 times the standard model configuration per integration interval); We estimate uncertainty in processes by comparing more detailed models with those used in GCMs, and explore a range of parameterizations of varying complexity. We have developed a strategy for Uncertainty Quantification (UQ) that helps identify solution sensitivity to parameter choices and optimal values for those parameters. The UQ procedures are quite expensive computationally, requiring ensembles of simulations. We test and evaluate model performance with these improvements through simulations in past, present, and future conditions.

8.4.2 Scientific Objectives for 2017

Our research addresses two sets of core questions:

1. Can climate models robustly simulate effects of aerosols on the energy balance of the climate system?

How well can models simulate aerosol indirect effects? What approaches to representing aerosols and their interactions with clouds are most effective for inclusion in climate models?

2. How have variations over the last century in aerosol emissions influenced major climate features, and how will future changes in aerosol emissions change climate?

We are focused on: a) specific improvements to parameterizations, particularly for CESM and CAM, and b) the impact of parameterization improvements on climate and climate change. The topics are studied in the context of fundamental parameterization development; we also integrate, compare, and evaluate parameterizations in the context of their impact on other parts of the climate system and assess the impact of those contributions on climate or climate change. Our treatment of the chemical mechanisms for Organic Aerosols is very simple, and will probably increase in complexity by 2017.

In summary, our scientific goal is to develop new capabilities for CESM and CAM and to use these to further understand the role of aerosol in the Earth system. Therefore, high-end computing and storage is the key for success of our scientific activities.

8.4.3 Computational Strategies

8.4.3.1 Approach

Our tool of choice is the CESM community model. This climate model represents the physics of the earth system (both the fluid flow and the diabatic processes that are important for the climate system). We develop and test various new physical representations of aerosols, and clouds, and the components of the climate system that they effect for CAM5/CESM and then evaluate the model performance in various frameworks. After building confidence of new physical representations, we use the model to understand the role of aerosols in the Earth system.

8.4.3.2 Codes and Algorithms

Our primary tool is the CAM5 and CESM community model in various incarnations. This model is well supported by the NERSC computational facility. It uses various finite volume and spectral element approaches for solving the fluid flow, and a variety of discretization techniques for treatment of the diabatic processes in the model. It scales very well to about 10^5 cores at high resolution, but at the usual resolution used by climate models it reaches its practical limit at a few thousand cores. We perform climate simulations over time intervals of a few days to a few centuries. We can (but usually do not) run the model as a “train” with parallel ensemble members that use a lot of cores simultaneously (this is the mechanism often employed at the ORNL LCF).

8.4.4 HPC Resources Used Today

8.4.4.1 Computational Hours

We requested 15 M hours at the NCAR CISL facilities for the period of Oct 2012 – Jan 2014. That proposal, entitled “Exploring Climate Modeling using CESM1” (PI: Philip J. Rasch), included not only a request in support of the DOE BER project (about 30% in the total request), but also other projects, including Earth System Models (EaSM) projects.

At NERSC during Allocation Year 2011 members of this repository ran about 10,500 jobs at NERSC on Franklin and Hopper that consumed 17.6 M hours. During AY 2012 12 M hours were consumed. The author of this case study has also used an additional 0.5 M hours as part of the mp9 (NCAR) repository.

8.4.4.2 Compute Cores

We have been using 1.9 x 2.5-degree resolution of CESM1/CAM5, which is relatively low resolution. In the fully coupled configuration, about 3,000 cores are used. In the atmospheric model only, 960 cores are used. The code and configuration we are using have a limit on scalability. After various tests and consultation with colleagues at NCAR and ORNL, we settled on a configuration of cores and load balancing that achieves reasonable throughput.

The NERSC job logs show that during AY2012, of the ~8,700 jobs run for this repository, about 60% use configurations consisting of 3,000-3,456 cores.

8.4.4.3 Data and I/O

Climate modeling produces large amounts of output. Minimum output (monthly and very few daily outputs) would consist of about 20 GB per one simulated year with CESM. Our typical run is 150 years long and with 10 different realizations. We checkpoint using restart files at the end of the run and once per year for most simulations. At our usual model resolutions each checkpoint set requires about 0.1 TB. In certain situations we archive more frequently. We currently have nearly 160 TB of data archived to the NERSC HPSS system.

8.4.4.4 Project Data

We have one project directory, PNNL-PJR, that currently contains about 20 TB (i.e., about five times the normal quota). We have been notified it is no longer backed up due to its size. We do back this up to HPSS by ourselves.

8.4.5 HPC Requirements in 2017

8.4.5.1 Computational Hours Needed

In our first year of operation (2011) we requested 20 M and used 17 M hours. In 2012, we requested 25 M, were granted 10 M initially, and used about 12 M hours out of our final allocation of 14 M. We expect future needs to increase at least ten to twenty times more than our 2011-2012 NERSC usage. There are several reasons we expect this much increase. First, higher spatial resolution in CESM will be needed. We include some of these higher spatial runs in our estimation for 2013. Second, more aerosol species and chemical tracers will be used. Currently we are involved in a couple of research tasks that make more aerosol species (e.g., the 4 mode version of the Modal Aerosol Module, MAM4) and more chemical tracers (e.g., more detailed representations of Secondary Organic Aerosol in CAM5). These configurations will increase the cost by an additional 10% to 50%.

8.4.5.2 Number of Compute Cores

At current resolutions CESM1/CAM5 stops scaling efficiently at about 3,000 cores. However, we are expecting that the next generation of CESM/CAM would be much more scalable due to introduction of new dynamical cores (e.g., Spectral Element, SE), the Model for Prediction Across Scales (MPAS), and the higher resolutions that we intend to use.

8.4.5.3 Data and I/O

Data input and output will be one of the big issues in the next generation of climate models. We are interested in regional and episodic events, and more frequent output at higher resolution will be required. Output can easily increase by one order of magnitude, and we anticipate this will be a real issue for some of our simulations. Writing restart files in CESM/CAM at the end of the run is relatively economical.

8.4.5.4 Project Data

We expect at least a two-to-four-fold increase. Currently we have been allocated 20TB, which is larger than the normal project allocation but even so, we regularly also have to utilize all the scratch spaces available to us on Hopper for some post processing of our model output.

8.4.5.5 Archival Data Storage

We anticipate an increase of about one order of magnitude in storage needs. Demand for high capacity and fast archival data storage becomes increasingly important to us. .

8.4.5.6 Memory Required

Running climate models with larger memory is beneficial.

8.4.5.7 Many-Core and/or GPU Architectures

Early tests of climate model show little benefit from GPUs but there are ongoing efforts in a SciDAC project to improve this situation, and we are participants in that activity. We will take advantage of this capability if the project achieves success.

8.4.5.8 Software Applications and Tools

CESM is evolving. This model activity is coordinated through NCAR and other DOE-funded projects. Most of our own research is focused on “science use” rather than “computational performance” but we will take advantage of gains in HPC performance as they become available, and will contribute where we can.

8.4.5.9 HPC Services

Large amounts of climate model output are expected to be available for the research community in the future. Transferring data at higher speed nodes and releasing data through the Earth System Grid (ESG) would be very beneficial to us. We want to stress that ESG and other forms of data sharing will become more important as data set sizes grow and we have more users.

8.4.5.10 Data Intensive Needs

Analyzing climate model output requires machines like “Euclid” for various short and I/O intensive jobs.

8.4.5.11 Additional Comments

Climate modeling is evolving into high-resolution in space and time, which means our data needs easily produce terabytes now and will produce petabytes in 2017. Large space and faster I/O for post-processing these data are necessary.

We would like to see an improvement in the way that changes to NERSC system software are communicated to users so that we can better associate differences we observe with NERSC changes.

Typically the validation of the CESM is done by NCAR but in the future we might need help with load balancing and optimization for specific machines. We also need an analysis system with large memory and fast I/O for interactive post processing (we need a resource like this). One of the reasons we have our large project space is for sharing data between Hopper and Euclid at NERSC.

8.4.6 Requirements Summary

	Used at NERSC in 2012	Needed at NERSC in 2017
Computational Hours (Million)	12.0	200
Typical number of cores* used for production runs	960 - 3,000	1,000 -10,000
Maximum number of cores* that can be used for production runs	960 - 3,000	1,000 - 10,000
Checkpoint data written per run	0.1 TB	1 TB
Data read and written per run (excluding checkpoint data)	30 TB (20*150*10)	300 TB
Project directory space	20 TB	80 TB
Archival data	158 TB	2 PB
Max Memory per node	2 GB	6-10GB

8.5 CLIMES and IMPACTS

Principal Investigator: William Collins (LBNL and UC Berkeley)
NERSC Repositories: CLIMES, m1204 and IMPACTS, m1040

8.5.1 Overview and Context

The "Investigation of Magnitudes and Probabilities of Abrupt Climate Transitions" (IMPACTs) project is exploring "tipping points" in Earth's climate that could quickly alter our natural environment and has two primary goals:

(1) Projecting the risk of abrupt climate change over the 21st Century, which involves several potentially significant types of abrupt climate change, including:

a) Disintegration of marine ice sheets

b) Melting permafrost leading to releases of CO_2 and CH_4

c) Destabilization of methane deposits in Arctic-circle oceans

d) Large-scale mega-droughts in North America

(2) Enhancing global models of these rapid climate transitions.

Our present focus in IMPACTs is to perform the first coupled projections of Earth's methane cycle; the first sea-level rise projections including Antarctica; and simulations of the future of western forests.

The Center at LBNL for Integrative Modeling and Measurement of the Earth System (CLIMES) investigates some of the central issues for the global environment and has two primary goals:

(1) Advancing simulations of climate forcing, response, and feedback, which has four key components:

a) Ultra high-resolution global climate simulation

b) Frameworks for robust regional climate modeling

c) Quantification of critical uncertainties in the carbon cycle

d) Representation of clouds, aerosols, and the cryosphere in climate models

(2) Advancing projections of climate mitigation measures, via

a) Improved representations of human-Earth system interactions

b) Integrated assessment model development, intercomparison, and diagnostics.

8.5.2 Scientific Objectives for 2017

By 2017 we expect to:

1. Develop probabilistic risks of abrupt climate change
2. Conduct local and regional projections of extreme rainfall

3. Simulate the $CO_2/CH_4/N_2O$ feedbacks in a warmer climate
4. Develop more integrated scenarios for climate mitigation

Doing this will require performance enhancement to support developments in:

- SLR: >10X for high-resolution land-ice / ocean models
- Extremes: 10 to 30X for superparameterized models
- Chemistry: 10X for reactive chemistry and transport
- Scenarios: 10 to 100X for scenario development.

8.5.3 Computational Strategies

8.5.3.1 Approach

We simulate climate computationally using models that solve the Euler equations, constituent equations, and thermodynamics for ocean, atmosphere, and ice.

The primary code we use is the DOE-NSF joint Community Earth System Model (CESM), found at http://www.cesm.ucar.edu/.

Distinctive features of our simulations include:

- Duration: Centuries to millennia
- Time steps: Minutes (atmosphere) to hours (ocean)
- Experiments: Response to time-evolving boundary conditions
- Metrics: Non-deterministic statistics of the solutions

8.5.3.2 Codes and Algorithms

CESM is comprised of five components and a coupler (numbers in parenthesis indicate resolution):

- Atmosphere: (200 x 288 x 30 = 1.7 Million grid points)
- Ocean: (180 x 360 x 40 x 0.7 = 1.8 Million grid points)
- Sea & land ice: (same as ocean/land grids)
- Land: (200 x 288 x 10 x 0.3 = 170 K grid points)

The dynamical frameworks are / are evolving to:

- Atmosphere: Spectral element dycore on cubed sphere (SNL)
- Ocean: Unstructured mesh/Voronoi tessellation (LANL)

Implementation of parallelism:

- Choice of MPI, OpenMP, MPI/OpenMP hybrid throughout.
- Components run in arbitrary mix of serial and parallel processor layouts.
- Parallel NetCDF for I/O.

Our biggest computational challenges are:

- Ensemble sizes required for uncertainty quantification (1,000s)
- 100x increase in throughput required for cloud/eddy-resolving models

- Barrier to long-time integrations from flat trends in clock rates

Current implementations exhibit scaling to O(10^5) processors:

- CESM scales to 30K Cray / 60K Blue Gene cores (Dennis et al, 2012)
- Spectral element dycore scales to 256K processors (Taylor et al, 2011)

Major changes anticipated by 2017:

- GPU implementation of CESM components (underway for Titan at Oak Ridge)
- New atmospheric/ocean dycores: focus on refinement, scalability
- Implementation of stochastic parameterizations in atmosphere/ocean
- AMR techniques for land ice

8.5.4 HPC Resources Used Today

8.5.4.1 Computational Hours

Machines currently running CESM:

- Major Facilities: NERSC, NCCS / OLCF, ALCF, NCSA, NCAR
- Architectures: Cray XE/XT, IBM Power Series, IBM Blue Gene, Linux cluster

Hours used in 2012:

- IMPACTS (m1040): 9.7 Million core hours at NERSC
- CLIMES (m1204): 4.1 Million core hours at NERSC
- Use at other facilities: 200 M core hours

8.5.4.2 Compute Cores

Typical parallel concurrency and run time, number of runs per year:

- Hopper cores: 2,064 (for 1-degree resolution)
- Hopper core-hours: 2,063 core-hours per year of simulation
- Hopper throughput: 24.01 simulation years per wall clock day
- Number of years/year 4,000 to 10,000 simulation years / calendar year

8.5.4.3 Data and I/O

Data read/written per run and data resources used

- In IPCC AR5 production runs, 56.2 GB/simulated month and 675 GB/simulated year
- Storage system: HPSS, 900 TB, ~2 million files, and 800 TB of I/O

Memory used globally

- 135 GB (e.g., the 1850 carbon/nitrogen compset (as reported in the Intel Benchmark, for HPC Advisory Council)

At the end of 2012 there are project directories at NERSC used by these projects:

- m1204 has a standard (4-TB quota with essentially no usage; and
- m1040 has a standard quota (4 TB) that is almost entirely consumed.

Archival storage at NERSC in 2012:

- m1040: 608 TB
- m1204: 268 TB

8.5.4.4 Necessary software, services or infrastructure

- UNIX like operating system (LINUX, AIX, OSX)
- csh, sh, perl, and xml scripting languages
- subversion client version 1.6.11 or greater
- Fortran 90 and C compilers. pgi, intel, and xlf are recommended options.
- MPI (although CESM does not absolutely require it for running on one processor only)
- netcdf 3.6.2 or greater
- Earth System Modeling Framework (ESMF) (optional) 5.2.0p1
- pnetcdf (optional) 1.1.1 or newer

8.5.5 HPC Requirements in 2017

8.5.5.1 Computational Hours Needed

We estimate that we will need 150 M hours at NERSC in 2017. This is four times our 2012 request and more than ten times the hours awarded at NERSC.

8.5.5.2 Number of Compute Cores

We anticipate using 2K to 20K processors per integration and we could use up to 100K in principle

8.5.5.3 Data and I/O

By 2017 we will be producing 200 GB of data per simulated month and 1.4 TB per simulated year. This translates to 21 PB in one calendar year.

8.5.5.4 Memory Required

Memory required for an entire simulation should be 4 * 135 GB = 540 GB.

8.5.5.5 Many-Core and/or GPU Architectures

Our strategy for many-core architectures involves collaboration with FASTMATH and SUPER Institutes via SciDAC climate apps:

- Transport and advection (led by ORNL)
- Land ice (LANL)
- Multiscale physics and integration w/new dycores (LBNL)

To date we have prepared for many core by implementing capabilities for arbitrary hybrid MPI / OpenMP parallelism and end-to-end MPMD architecture.

The CESM project is committed to porting to MIC machines, including the TACC Stampede machine based on Knights Corner.

8.5.5.6 Software Applications and Tools

Our codes require parallel NetCDF and we use the visualization and data analysis tools GrADS, IDL, MATLAB, NCAR, R, and VisIt.

8.5.5.7 HPC Services

The following "expanded" HPC resources are important for our project:

- Integration of provenance tracking throughout software / project / data cycles;
- "Rendering engines" (hardware and/or MPP software) for exabyte data sets;
- Multi-terabyte/second networks to key partners including LCFs, NCAR, etc.

8.5.5.8 Additional Comments

A "leading without bleeding" procurement strategy at NERSC works well for our applications, since we can leverage substantial DOE and NSF investments in performance portability.

Over the next five years with access to 32X our current NERSC allocation we could make significant scientific progress in:

- Advances towards global eddy-resolving projections of sea-level rise
- Initial exploration non-hydrostatic cloud-system-resolving climate models
- Development of regional-to-global carbon/water/climate analyses of the Earth system
- Integrated climate/energy scenarios for the Sixth IPCC report and national assessments

8.5.6 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Millions)	13.8 Total	150
CLIMES (NERSC repo m1204)	4.1	
IMPACTS (NERSC repo m1040)	9.7	
Typical number of cores* used for production runs	2 K	20 K
Maximum number of cores* that can be used for production runs		100 K

Data produced per year	5 PB	21 PB
Archival data	876 TB	4 PB

8.6 Climate Science for a Sustainable Energy Future

Principal Investigators: James Boyle and David Bader (LLNL); Mark Taylor (Sandia National Labs)

Case Study Authors: David Bader and Mark Taylor

NERSC Repository: mp193 ("Program for Climate Model Diagnosis and Intercomparison")

8.6.1 Overview and Context

This work involves modeling the Earth's climate system, which requires running coupled atmosphere/ocean/land/ice models. In the CSSEF and related Cloud-Associated Parameterizations Testbed (CAPT) research at Lawrence Livermore National Laboratory, the emphasis is on the atmospheric component. The modeling system used is the Community Earth System Model (CESM). The processes governing Earth's climate system exhibit a wide range of time and space scales spanning many orders of magnitude. The multiscale nature of the scientific problem makes it extremely challenging to accurately represent all the relevant scales of motion in mathematical and numerical models, particularly with regard to treating the process of phase change, or more generally, the processes governing Earth's hydrological cycle. It is generally recognized that our ability to numerically model climate and climate change is fundamentally limited by a lack of understanding of the interaction of hydrological processes and the large-scale radiation field, particularly with respect to clouds. The goal of this project is to make significant progress in the simulation of the hydrological cycle of the global climate system.

8.6.2 Scientific Objectives for 2017

By 2017, we will produce a global atmosphere model with about 10-km horizontal resolution with integration rates suitable for multi-century climate modeling and a companion testbed that can be used to further improve the model. We envision that the 10-km global atmosphere model — a so-called "weather-resolving climate model" — will become a formal released configuration of the Community Atmosphere Model (CAM) and be suitable for global climate integrations as part of the Community Earth System Model (CESM).

8.6.3 Computational Strategies

8.6.3.1 Approach

The computational components of this project include (1) quantifiable model-observation comparison techniques embedded with systematic and reproducible optimization of perturbed physical parameters; (2) utilization of the latest water cycle observations from ground-based and satellite instruments; (3) testing of the three-dimensional (3-D) climate model; and (4) computational efficiency to permit parameter optimization of high-resolution simulations. The atmospheric testbed consists of two components: a calibration platform and a validation platform. The calibration platform is where Uncertainly Quantification (UQ) techniques are used to calibrate the model against local data sets. The validation platform will test the hypothesis that model calibration to one or more fixed-site data sets yields improved global simulations of water cycle processes. In this platform, a global model with uniform resolution is integrated in "climate mode," driven only by observed sea-surface temperatures and sea-ice distributions. In model development cycles, these integrations are routinely performed for ~20 years of simulated time to demonstrate the fidelity of a model's climate. Due to its computational expense, in the first years of the project, only limited ensembles of such runs are possible at the desired 1/8° global resolution.

8.6.3.2 Codes and Algorithms

The dynamical core to be implemented in the atmospheric model is the variable resolution Community Atmosphere Model with the Spectral Element dynamical core (CAM-SE). This is a scalable dynamical core that is available as an option in CAM. The High-Order Methods Modeling Environment (HOMME) dynamical core introduces a new horizontal discretization in CAM based on the spectral element method, which is a type of continuous-Galerkin finite element method. The spectral element configuration is fourth-order accurate and designed for fully unstructured quadrilateral meshes. CAM-SE is the first unstructured grid dynamical core integrated into CAM. At (1/4°) global resolution, CAM-HOMME has nearly perfect scalability on a Blue Gene/P system (Intrepid, ANL), scaling out to 86,000 cores, representing one element per core.

The spectral element method has been designed from inception to support adaptive grids and grids with regional refinement, and all the numerical properties of the method (conservation, fourth-order accuracy) are retained on such grids. In 2D, local mesh refinement with spectral elements has been shown to be effective at reducing the global solution errors when refinement is used over dynamically significant regions such as localized topography.

8.6.4 HPC Resources Used Today

8.6.4.1 Computational Hours

Users of the PCMDI repository consumed about 7.3 M hours at NERSC during the 2012 allocation year (through mid-November) and about 6.7M hours during AY2011. However, most 2012 entries in the table below are left empty because although the reference point is the PCMDI project, the 2017 science described is a completely different project.

8.6.4.2 Compute Cores

The current configuration of CESM does not scale well. Runs on Hopper for science of interest to this repository have routinely used 7,680 cores, a value found to offer reasonable efficiency. Such runs use 1,280 MPI processes, four per node, plus six OpenMP threads per MPI process. The result is one year of simulation in about 24 wallclock hours.

8.6.4.3 Data and I/O

CESM restart files for current runs are about 25 GB. Data output can vary a great deal depending on the frequency of output (e.g., monthly, daily, or six-hour) and the number of variables needed. Typical monthly data is about 11 GB, so a 20-year run would produce about 11 GB * 12 * 20 = 2640 GB.

This project currently has a NERSC project directory "CAPT" with about 4TB of data stored in it and a 5-TB quota.

8.6.5 HPC Requirements in 2017

8.6.5.1 Computational Hours Needed

For our estimates of HPC requirements in 2017 we will assume the bulk of the simulations will be fully coupled with 1/8-degree atmosphere and 1/10-degree ocean. We base our estimates on similar experimental configurations that we have previously run at these same resolutions. Based on benchmark numbers for fully coupled simulations using CAM-SE with CAM4 physics and the POP ocean model out to O(200K) cores, as well as atmosphere-only simulations on up to 64K cores using CAM-SE with CAM5 physics (which are significantly more expensive than CAM4 physics), we estimate fully coupled CAM5 simulations would cost roughly 3M core hours per simulated year. This number can be reduced significantly with the upcoming introduction of prognostic aerosols in CAM, and will also be reduced with anticipated scalability improvements in the CESM flux coupler. However, by 2017 additional complexity may be introduced into the model, which would raise the cost in terms of core-hours. In addition, HPC systems in 2017 are expected to have many more cores per node, but we don't expect significant gains in the performance per core. Thus for simplicity, our 2017 estimates will be based on a cost of 3M core hours per simulated year.

8.6.5.2 Number of Compute Cores

The CESM with CAM-SE configuration runs effectively at this resolution on 100,000 cores.

8.6.5.3 Data I/O

Typical CESM output includes monthly daily and hourly averages with additional high-frequency output added for specific simulations, up to 1.2 TB per simulated year, requiring an average bandwidth of 6 TB/day (73 MB/s) when running at a typical throughput of five simulated-years-per-day. This also implies about 1.2 TB per simulated year for archival storage. Based on our current rate of increase for archival storage, we conservatively estimate needing 7.5 PB at NERSC in 2017.

The CESM is currently using the PIO aggregation library and parallel-netcdf.

Most likely, the default level of "project" storage will be sufficient.

8.6.5.4 Memory Required

We estimate needing 2 GB per MPI task.

8.6.5.5 Many-Core and/or GPU Architectures

The CESM has long relied on a hybrid MPI/OpenMP programming model that runs well on systems with moderate core counts and can hopefully be extended to take advantage of many-core systems. Personnel from Cray, ORNL, Nvidia, and NREL have been modifying CAM-SE to utilize the GPUs on the DOE Titan system. Current results show 2-3x speedup over running on Titan without using the GPU, but we hesitate to extrapolate these gains to the full CESM.

8.6.5.6 Software Applications and Tools

Fortran 90 compiler

MPI library mvapich2

Parallel-netcdf or HDF5/NetCDF4 library

Data analysis utilities (NCL, NCO, cdat, paraview)

8.6.5.7 HPC Services

Simulation data needs to be archived and made available onsite. Data is often used by many researchers for several years after it is produced, looking at many different quantities often requiring significant amount of post-processing

8.6.5.8 Time to Solution and Throughput

We expect to be running at a throughput of five simulated-years-per-day, requiring four days of compute time for a typical 20-year simulation.

8.6.5.9 Data Intensive Needs

Post-processing is the most data intensive operation, as the data are continuously reprocessed by different researchers.

8.6.6 Requirements Summary

Note: for the table below, I took "run" to represent 1 simulated taking 5 hours of wall clock time.

	Used at NERSC in 2012	Needed at NERSC in 2017
Computational Hours (Million)	7.8	150
Typical number of cores* used for production runs		75 K
Maximum number of cores* that can be used for production runs		400 K
Checkpoint data written per run		150 GB
Checkpoint bandwidth		1 GB/s
Data read and written per run (excluding checkpoint data)		1.2 TB
Maximum I/O bandwidth (excluding checkpoint data)		5 GB/s
Project directory space		350 GB

Archival data	960 TB	7.5 PB
Minimum memory per node		2 GB
Aggregate memory		16 TB

8.7 Projecting Ice Sheet and Climate Evolution at Extreme Scales

Principal Investigator: William Lipscomb (LANL); Phil Jones (LANL) – Acting P.I.

Case Study Author: Stephen Price (LANL)

NERSC Repository: m1343 - Projections of ice sheet evolution using advanced ice/ocean models (W. Collins (LBL) P.I.)

8.7.1 Overview and Context

During the past decade, mass loss from ice sheets has raised global mean sea level by 1 mm/yr, roughly equal to the contributions from ocean thermal expansion and the melting of smaller glaciers and ice caps. If recent trends continue, ice sheets will make a dominant contribution to 21st century sea-level rise. Although ice sheet models have improved in recent years, much work is needed to make these models reliable and efficient on continental scales, to couple them to earth system models, and to quantify their uncertainties.

The Projecting Ice Sheet and Climate Evolution at Extreme Scales (PISCEES) project will continue development of two ice sheet dynamical cores targeted for DOE HPC platforms: (1) BISICLES, a finite-volume core on a structured mesh, using the Chombo adaptive mesh refinement (AMR) software framework, and (2) FELIX, a finite-element core on an unstructured mesh, using the Model for Prediction Across Scales (MPAS) framework and the Trilinos software library. Both will include a hierarchy of solvers applied at variable resolution and in different regions of dynamical complexity and will be engineered to be highly scalable and to optimize performance on new high-performance computers with heterogeneous architectures. PISCEES will also develop new methods and tools for ice sheet model initialization, verification and validation (V&V), and uncertainty quantification (UQ), allowing for confidence ranges on projections of sea-level rise from ice sheets. These improved models and tools will be implemented in the Community Ice Sheet Model (CISM) and the Community Earth System Model (CESM). The outcome of PISCEES will enable quantitative predictions of coupled ice-sheet/climate evolution using a new generation of high-performance computers and computational tools.

8.7.2 Scientific Objectives for 2017

By 2017 we aim to have high-resolution, fully coupled simulations of ice sheet and climate evolution (e.g. sea-level rise) with uncertainty quantification. We will run stand-alone CISM model simulations as well as runs fully coupled (ocean-atmosphere-sea ice) to CESM. CISM will use the BISICLES and FELIX dynamical cores, both with adaptive mesh refinements and hierarchical, 3-D, higher-order momentum balance solutions (including and up to nonlinear Stokes).

8.7.3 Computational Strategies

8.7.3.1 Approach

Ice sheet flow is most accurately described by the nonlinear Stokes-flow equations. While lower-order approximations may be adequate in some cases (ice sheets can often be treated as low-aspect ratio flows) the dominant problem to be solved is that of a nonlinear, elliptic PDE for the components of the velocity field. The nonlinearity, arising from the power-law viscous rheology of ice, is treated using standard iterative methods (e.g., Newton- and Picard-based methods). Conservation of energy is described by a standard advective-diffusive heat equation. Conservation of mass follows from the treatment of ice as an incompressible fluid. Because the velocity solution at any time step can be diagnosed (i.e. there is no time tendency term for the momentum balance equations), a sequential solution approach is usually taken during any time step; velocities are diagnosed from the current geometry, boundary conditions, and temperatures (the viscosity is also temperature dependent) and those velocities are then used to advect heat and mass over the same time step. Currently, an explicit forward-Euler time stepping scheme is used.

8.7.3.2 Codes and Algorithms

CISM (the Community Ice Sheet Model) is the ice sheet model component of CESM (the Community Earth System Model). Currently, CESM is partially coupled (land and atmosphere only – ocean coupling work is ongoing) to a version of CISM that contains a crude dynamical core (the momentum balance is incompletely described by column physics only). At least three new, significantly more advanced dynamical cores will soon be coupled to CISM (and hence to CESM). These include (1) SEACISM, a finite-difference based dynamical core on a structured, regular grid; (2) BISICLES, a finite-volume based dynamical core with block-structured, adaptive mesh refinement capabilities; (3) FELIX, a finite-element based dynamical core on a fully unstructured mesh. All three dynamical cores are, or will become, fully 3d with higher-order accurate treatments of the momentum balance for ice sheets. All three dynamical cores are fully parallel (using MPI). The primary computational bottleneck and limit to scalability for all dycores in CISM is (and will likely remain) the velocity solve, which requires the solution to a large (order ~2-4 x the number of grid cells), sparse, nonlinear, elliptic system of equations. Conservation of energy and mass are currently handled through column physics (e.g. vertical heat diffusion) and explicit advection (e.g. heat and mass advection), both of which are scalable using existing algorithmic approaches. SEACISM is currently coupled to CISM and work is ongoing to couple this version of CISM to CESM in time for the spring 2013 CESM 1.1.1 release. Work on BISICLES and FELIX are ongoing. Both will be integrated into CISM and coupled to CESM within the next ~2 yrs. SEACISM and FELIX use the Trilinos solver library and BISICLES uses the Chombo libarary. FELIX also uses the MPAS (Model for Prediction Across Scales), unstructured-mesh, climate modeling framework.

8.7.4 HPC Resources Used Today

8.7.4.1 Computational Hours

Total ice sheet model use on Hopper over past year: ~900 K hours

Total ocean model** use on Hopper over past year: ~1,400 K hours

Total ice sheet use on Jaguar over past ~2 years: ~3,275 K hours

** This is only ocean model use related to ice-sheet/ocean model coupling experiments.

8.7.4.2 Compute Cores

A "standard" current run for Greenland at 5 km spatial resolution with ~10 vertical levels (~650K grid cells and ~1.3 x 10^6 DOFs) uses 1-2 K cores for our 3D codes.

The same code scales reasonably well up to ~6 K cores. A typical run uses less than the maximum because we can always run fewer cores for a bit longer rather than waste more cores because of poor scaling (while performance work on the code continues – in the long term, we expect scaling out to order 10 K cores).

In the past, we've had somewhere between 1-5 jobs running concurrently (or at least in the queue). These would be similar runs with, e.g. different parameters settings or different climate forcing time series applied.

8.7.4.3 Data and I/O

Currently, a standard restart file (i.e., a "checkpoint" file – one record of the model state needed for restarting a run at some later date) for a uniform, 5-km resolution Greenland run is ~60 MB. For uniform 5-km resolution Antarctica, a similar file would be approximately 500 MB.

For a 100-year run (e.g., 2000-2100), recording the equivalent of a checkpoint file at 1 year intervals gives an output file of ~6 GB, for 5-km resolution Greenland, or ~50 GB for 5-km resolution Antarctica.

Currently, very little of the total runtime is spent in I/O (approximately 5% in I/O, 90% in velocity solve). This is an estimate from some initial profiling of the code and is likely to increase in the future. Climate-forced runs in stand-alone mode may require regular input of at least several 2D fields. For this reason, I/O might be as large as 15% for some runs.

8.7.5 HPC Requirements in 2017

8.7.5.1 Computational Hours Needed

Ice sheet model only runs*

100 100-year stand-alone Greenland: 100 * 55 K = 5.50 M

- Minimum needed to satisfy UQ requirements = 11 M

100 100-year stand-alone Antarctica: 100 * 435 K = 43.5 M

- Minimum need to satisfy UQ requirements = 87 M

Coupled runs**

3 ice-sheet/ocean Greenland: 3 * (500 K + 55 K) = 1.67 M

3 ice sheet/ocean Antarctica: 3 * (5 M + 435 K) = 16.3 M

2 ice-sheet/ocean/atmos Greenland: 2 * (2*500 K + 55 K) = 2.1 M

2 ice-sheet/ocean/atmos Antarctica: 2 * (2*5 M + 435 K) = 20.9 M

1 ice-sheet/ocean/atmos/sea-ice Greenland: = (3*500 K + 55 K) = 1.6 M

1 ice-sheet/ocean/atmos/see-ice Antarctica: (3*5 M + 435K) = 15.4 M

Total: 156 M

* Includes assumptions that about 20 model optimization (deterministic inverse) problems are done for both Antarctica and Greenland (i.e. using an efficient adjoint-based code, which cost ~100X a forward model solve. 100 forward model solves (for a 1-year time step) is approximately equal in cost to a 100 year forward model run.

** Climate-coupled runs are assumed to be conducted with MPAS-ocean/atmosphere/sea ice, at high spatial resolution and on a regional domain, allowing for ~10X savings over, e.g. current hi-res (10th-degree POP). Original estimates are for Antarctic sub-domain. Greenland sub-domain is assumed to be ~10X cheaper.

*** Note that the estimate for the total number of stand-alone ice sheet simulations is extremely optimistic in terms of addressing the UQ aspects of the proposed work. In particular, it assumes success in characterizing and forward propagation of uncertainty using “linearized UQ” (linear adjoint approaches to avoid sampling and prohibitively large and expensive numbers of forward model runs), and/or the use of (computationally efficient) emulators to sample the parameter space efficiently without large numbers of samples and full (i.e., PDE-based) forward model runs, and/or gradient/Hessian informed MCMC methods. None of which are currently in common use or have even been adequately tested on analogous problems. For these reasons, it should also be noted that sampling based (e.g. MCMC) methods of UQ could easily increase the CPU requirements noted here by 1-2 orders of magnitude (and possibly more). We assume a ~2X increase in the number of stand-alone ice sheet model runs in order to allow for a nominal amount of UQ.

8.7.5.2 Number of Compute Cores

While our code is currently only scaling to order ~1 K cores, we have run it on as many as 20 K cores, while still seeing performance increases (although with far less then ideal scaling). Assuming increased performance of the code over the next few years, it seems reasonable to assume we will have the code running regularly (and scaling) on order ~10 K cores.

In the future, we expect the number of concurrent runs to increase significantly since we will be doing parameter optimization (e.g. for finding optimal initial conditions), sensitivity analysis, and uncertainty quantification. Thus, at times we might aim to run order 10-100 jobs concurrently.

8.7.5.3 Data and I/O

Scaling up our current estimates based on anticipated spatial resolution for standard runs in the future, a single “checkpoint” (restart) file would be approx. 360 MB for a Greenland ice sheet run and 3 GB for an Antarctic ice sheet run (~6x larger than at present). The stored output fields for a 100 year run would be ~33 GB for Greenland and ~270 GB for Antarctica.

We would also be storing checkpoint files and model output for a relatively small number of runs (e.g., one dozen) in which the ice sheet, the ocean, and/or the atmosphere, and/or sea ice are coupled together. For Antarctica, a high-resolution, regional ocean simulation for 100 years would require ~800 GB of storage for restart files and 1,000 GB of storage for model output. For runs that also include a coupled atmosphere, double these numbers (1,600 GB and 2,000 GB, for restarts and output, respectively), and for runs that also include coupled sea-ice, triple these (2,400 GB and 3,000 GB). For Greenland, reduce all of these numbers by a factor of ~8.3x.

Currently, very little of the total runtime is spent in I/O (approx. 5% in I/O, 90% in velocity solve). This is an estimate from some initial profiling of the code and is likely to increase in the future. Right now, this involves writing approximately one "checkpoint" (restart) sized file per year for a century scale long run. Climate-forced runs in stand-alone mode may require regular input of at least several 2D fields. For this reason, I/O might be as large as 15% for some runs. Ideally time spent in I/O would be kept to a minimum, but we don't have enough experience yet for a baseline estimate of what a "normal" amount of time might be, but keeping I/O at or below 5% of total run time would be ideal.

8.7.5.4 Project Data

Stand alone runs (neglecting restarts)

100 100-year stand-alone Greenland: 100 * 33 GB = 3.3 TB

100 100-year stand-alone Antarctica: 100 * 270 GB =27 TB

Coupled runs (neglecting restarts)

3 ice-sheet/ocean Greenland: 3 * (33 GB + 120 GB) = 460 GB

3 ice sheet/ocean Antarctica: 3 * (270 GB + 1,000 GB) = 3,810 GB

2 ice-sheet/ocean/atmos Greenland: 2 * (33 GB + 2 * 120 GB) = 545 GB

2 ice-sheet/ocean/atmos Antarctica: 2 * (270 GB + 2 * 1,000 GB) = 4,540 GB

1 ice-sheet/ocean/atmos/sea-ice Greenland: 33 GB + 3 * 120 GB= 390 GB

1 ice-sheet/ocean/atmos/see-ice Antarctica: 270 GB + 3 * 1,000 GB= 3,270 GB

Total: 43 TB

All project data will need to be archived. Over five years (2012-2017), the accumulated archival storage need will be about an additional 250 TB.

8.7.5.5 Many-Core and/or GPU Architectures

At the moment, there is no obvious place in the code where GPUs could be put to use (the bulk of the computation time for any time step is the elliptic solve for the velocity field). In later years of the project (2014-2017) project members at ORNL and LBL will be looking at the potential for code performance improvement through the use of GPUs. At present, however, the utility of GPUs for performance improvements of ice sheet codes remains an open question.

8.7.5.6 Software Applications and Tools

Our project needs are currently fairly standard, including netCDF, Trilinos, and Chombo libraries, standard suites of compilers, and NCO.

8.7.5.7 HPC Services

Considering the anticipated increase in output file size and the difficulties associated in moving these files from platform to platform for post processing, some support for data analytics and/or visualization tools may be necessary.

8.7.5.8 Time to Solution and Throughput

The International Panel on Climate Change (IPCC) assessment reports (AR) historically occur at five-year intervals (e.g., AR4 in 2007/8, AR5 in 2012/13). Assuming that format continues into the future, 2017/18 would coincide with AR6. In that case, you should assume a spike in climate-related computing activity from ~2015-2017.

8.7.6 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Million)	2.3	156
Typical number of cores* used for production runs	1,000	10,000
Maximum number of cores* that can be used for production runs	6,000	20,000
Data written per run	47 GB	270 GB
Project directory space	70 GB	43 TB
Archival data	58 TB	300 TB

* “Conventional cores.” For GPUs and accelerators, please see section 4.8.

8.8 Anthropogenic Climate Change Using Multi-scale Modeling Frameworks and Super-Parameterization

Principal Investigator: Cristiana Stan (GMU and IGES-COLA)

NERSC Repository: m1441

8.8.1 Overview and Context

Current state-of-art numerical models used for climate prediction include only a statistical representation of weather events. There is increasing evidence suggesting that the current gap between weather- and climate-prediction models has to be closed, and the two components have to be part of a seamless prediction system. The processes that determine weather are closely and nonlinearly related to the processes that determine climate.

The primary goal of the proposed research is to conduct and analyze simulations of anthropogenic climate change based on a version of the Community Climate System Model (CCSM) in which representation of cloud processes in the atmosphere model is based on the "super-parameterization," referred to as SP-CCSM. In the super-parameterization approach, atmospheric convection is explicitly represented to improve the simulation of cloud processes. SP-CCSM combines processes that determine the weather and climate in a unified framework and therefore can provide accurate and reliable predictions of regional climate change, including statistics of extreme events and high impact weather, which are required for both local and global adaptation strategies.

The proposed research takes advantage of increasing computational capability to push the resolved scales into the cloud regimes in the atmosphere. Because of cost, super-parameterized experiments to date have implemented a 2-D representation of atmospheric convection. The one-dimensional horizontal sub-grid that results leads to unrealistic features that go away with a 3-D representation.

8.8.2 Scientific Objectives for 2017

The objectives of the anthropogenic climate change experiments are to answer questions like 1) Does the explicit representation of clouds change the projected global mean warming and the associated increase in global mean precipitation estimated from simulations with conventional cloud parameterizations? 2) Do patterns of change projected by the current generation of models depend on the representation of cloud processes? 3) Is the El Niño Southern Oscillation (ENSO) interannual variability sensitive to cloud representation? 4) Does global warming in this model lead to a change in the ENSO teleconnections?

Our present focus is on simulations in which the cloud-resolving model is 2D. Our need for 2017 is to conduct simulations in which the cloud-resolving model is 3D and with improved physics to fully represent a three-dimensional view of cloud dynamics.

8.8.3 Computational Strategies

8.8.3.1 Approach

The project will make use of the SP-CCSM code, which includes a point-wise modification of the atmospheric component of the CCSM (CAM). The ocean, land, ice and coupler components of SP-CCSM are the baseline codes from CCSM. The atmospheric component of SP-CCSM, referred to as SP-CAM, is based on the finite volume CAM in which the call to the physics package is replaced by a cloud-resolving model (CRM) that represents atmospheric convection explicitly. Conceptually, each vertical column of the global model – i.e., every horizontal grid point – interacts with its own CRM by passing a small number of global model tendencies arising from processes other than clouds. The CRM then uses those inputs to integrate, running with a much smaller time step than the global model time step (e.g., 20 seconds *vs.* 1800 seconds). Once complete, the output of the CRM is a vertical profile of physics tendencies that represent the impact of convective-scale cloud processes on the global atmospheric model. The CRM model is embarrassingly parallel, comprised of a relatively compact set of finite differences codes. CRMs can be parallelized through both message passing and OpenMP threads.

8.8.3.2 Codes and Algorithms

The 3-D CRM model we propose for use in the experiments has horizontal 12x12 mesh grid points with 30 levels. For the 0.9 x 1.25 resolution proposed for the experiment, the level of SP-CAM parallelism available is simply the number of horizontal grid points, 55,296 MPI tasks. Current simulations with the 2-D CRM run on 4,096 cores. A four-fold increase in computational resources is required just to allow us to maintain the same model integration rate.

8.8.4 HPC Resources Used Today

8.8.4.1 Computational Hours

During AY2011 the PI had a NERSC ERCAP project m1441, "Simulations of Anthropogenic Climate Change Using a Multi-scale Modeling Framework," that used 5.3 M hours at NERSC. A total of about 1,550 jobs were run, all on Hopper, primarily at a concurrency of 176 nodes (4,224 cores) but also using as many as 326 nodes (7,824 cores). During AY2012 the PI has two projects at NERSC. The m1441 project was initially allocated 2M hours and then increased to 5 M hours but this allocation was entirely used by October, with 469 176-node Hopper jobs. In 2012 the PI was also awarded a new ASCR Leadership Computing Challenge (ALCC) allocation, m1576, "Reducing Uncertainty of Climate Simulations Using the Super-Parameterization," that through early-December had used 15 M hours with 347 (mostly) 176-node Hopper jobs. There was also usage of the NSF Kraken machine during 2012 but this allocation ended in September. There is also a 2.7 M-hour allocation on the NCAR Yellowstone machine but this system is not available yet.

8.8.4.2 Compute Cores

The number of cores used today is chosen based on the following type of analysis. The atmospheric and ocean components exist simultaneously, the former using 4,096 cores (1 MPI task per) and the latter using 128 cores. The coupler and ICE components exist simultaneously, the former using 2,880 tasks and the latter using 1,024. So the maximum at any time is 4,224. This assumes the following grid sizes:

ATM: Nx=288, Ny=192, Nz=30

LRM: Nx=32, NZ=28

LND: Nx=288, Ny=192

OCN: Nx=320, Ny=384, Nz=60

ICE: Nx=320, Ny=384

For this, the number of Cloud Resolving Models (CRMs) is 288x192=55,296; this number is limited by available memory. The maximum number of MPI processes used in the latitude-vertical decomposition is 64x4=256 and there are 13.5 CRM calculations per core.

8.8.4.3 Data and I/O

A typical run will write 4.7TB overall and about 250 GB per checkpoint file. We currently have 10.6 TB archived on HPSS.

8.8.4.4 Project Data

There is currently a project directory for m1441 with about 2.3 TB of data stored in it. It is used to store static files shared by members.

8.8.5 HPC Requirements in 2017

8.8.5.1 Computational Hours Needed

In 2017 we expect about a two-fold increase in concurrency per run. There is a need to do two kinds of runs, control runs and actual climate change runs using future IPCC scenarios. We would run 100 simulated years for both but with one ensemble for the control and four ensembles for the climate change runs for uncertainty quantification, four being a small number but a good compromise based on how expensive the runs are. Ten ensembles would probably be a good number.

8.8.5.2 Number of Compute Cores

See above (14.4.2).

8.8.5.3 Data I/O

Checkpoint data written is likely to increase to about 4TB per run and overall output from a run is expected to increase from about 4 TB to about 70 TB.

8.8.5.4 Project Data

We do not envision that this will change very much from 2012.

8.8.5.5 Archival Data Storage

Increasing resolution will probably mean that we will require an aggregate of 150TB of HPSS storage in 2017.

8.8.5.6 Memory Required

The memory requirements for the proposed experiments are likely to be tractable, depending on trends in available memory per core. First, the amount of memory used by the CRM is relatively small today – order MBs per core. Fourteen 2D-CRMs per core are known to run on the Cray XT5, which has only 1.33GB per core. The proposed 3D-CRMs are only about four times larger.

8.8.5.7 Many-Core and/or GPU Architectures

There are some efforts to port the super-parameterization on the GPU architectures. In five years this option might become available for production runs. We believe that the CAM-SE dycore should work for our studies by 2017.

8.8.5.8 Software Applications and Tools

pgi/Fortran

netCDF/pnetCDF

MPICH MPI Library

Cray LIBSCI

8.8.6 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Million)	5.2	55
Typical number of cores* used for production runs	4,224	~16,000
Maximum number of cores* that can be used for production runs	16,512	~64,000
Checkpoint data written per run	0.25 TB	4.5 TB
Checkpoint bandwidth	2.5 GB/sec	2.5 GB/sec
Data read and written per run (excluding checkpoint data)	4.5TB	70TB
Maximum I/O bandwidth (excluding checkpoint data)	2.5GB/s	2.5GB/s
Project directory space	4 TB	6 TB
Archival data	10.6 TB	150 TB
Memory per node	1.33 GB	8 GB

* “Conventional cores.” For GPUs and accelerators, please fill out section 4.8.

8.9 Development of Frameworks for Robust Regional Climate Modeling

Principal Investigator: Lai-Yung (Ruby) Leung (PNNL)

NERSC Repository: m1178

8.9.1 Overview and Context

The objective is to apply a hierarchical framework to evaluate three dynamical approaches to modeling regional climate through global high-resolution models, global variable resolution models, and nested regional climate models, all sharing a common physics package. The Global high-resolution model consists of CAM Spectral Eulerian and HOMME; the Global variable resolution mode is CAM-MPAS; and the nested regional climate model is WRF. Our present focus is analysis of aquaplanet simulations and AMIP style simulations to assess the impacts of dynamical framework, dynamical core, and model resolution

8.9.2 Scientific Objectives for 2017

By 2017 we expect to:

- Move towards higher resolution, including cloud resolving simulations
- Have more focus on coupled simulations
- Evaluate interactions among dynamical framework, dynamical core, and model resolution in the context of scale-aware physics parameterizations
- Apply models to understand water cycle variability and extremes

8.9.3 Computational Strategies

8.9.3.1 Approach

We approach this problem computationally at a high level by utilizing existing software/hardware to perform idealized and real world simulations with different models at low resolution (1°), high resolution (0.25°), and variable resolution (1° → 0.25°).

8.9.3.2 Codes and Algorithms

The codes we use include offline and coupled atmosphere/ocean models:

- CAM Spectral Eulerian, POP
- CAM HOMME, POP
- CAM MPAS-A, MPAS-O
- WRF, ROMS

MPAS-A and MPAS-O are characterized by these algorithms:

- The codes are fully explicit - no global reductions or large linear system solvers are used.
- Scaling to large processor counts is dependent upon our ability to transfer "halo" data between processors in a local communication pattern (i.e., one processor sending messages to less than eight processors) on each time step.
- The data model is structured in the vertical, but unstructured in the horizontal – directly address data in the vertical, but require indirect addressing to find neighboring data in the horizontal.
- Data are laid out with the vertical index first in order to exploit this structured index - largely mitigates the inefficiencies incurred due to using unstructured addressing in the horizontal.

WRF is characterized by these algorithms:

- The ARWRF solver uses a time split finite difference scheme.
- The code has two levels of domain decomposition (patch and tile) designed to run over distributed as well as shared memory.
- Scaling to larger processor counts is limited by I/O and communication (little is gained beyond 100 grid points per tile).

In general, our biggest computational challenges are:

- Completing long integrations require frequent submission of sequential jobs, but long wait time in the queue limits productivity
- Not getting the amount of resources requested limits what can be accomplished
- WRF not able to utilize a large number of processors limits efficiency
- Our parallel scaling is limited by I/O for WRF

We expect our computational approach and/or codes to change by 2017 in this way:

- MPAS uses common approaches to high performance computing that exploit large, massively parallel computing systems.
- Researchers at LANL are exploring mixed parallelism in the form of MPI-OpenMP for MPAS-O and will test parts of the code on accelerators (GPUs) over the next year.

8.9.4 HPC Resources Used Today

8.9.4.1 Computational Hours

Machines currently used:

- MPAS: NERSC Hopper, LANL Mustang
- WRF: NERSC Hopper, NCCS Jaguar

During AY 2012 this project had an initial 6.4 M hour allocation that was later extended to its current 11.5 M-hour allocation. The project has used ~5.9 hours at NERSC, running about 1,700 jobs, about 1,400 on Hopper and ~300 on Carver. An additional approximately 1.5 M hours have been consumed on Jaguar.

8.9.4.2 Compute Cores

MPAS currently uses approximately 4,000 cores per run. The maximum number of cores that have been used is 6,000. Typically MPAS-O/MPAS-A jobs are not run with multiple jobs concurrently. WRF typically uses 1,296 cores per run. A global quarter-degree tropical channel WRF simulation typically costs about 200 processor hours per model run day.

The MPAS model is designed to disallow any global arrays – this somewhat mitigates thin nodes by spreading the problem over more nodes, even if this does not improve time-to-solution.

MPAS typically uses a large suite of compilers (pgf, ifort, gfortran, xlf, etc) to test robustness of the code. It also uses various flavors of MPI (mpich and openmpi) and the NCAR PIO tool for parallel output. WRF typically uses pgf90 and MPI.

8.9.4.3 Data and I/O

For MPAS, a typical checkpoint is 10 GB. Typically about 20% of wall clock time is used for I/O. Of this, checkpointing accounts for about 1/4 of I/O time.

In WRF about 50% of the time is spent on IO (writing six hourly data). The MPAS code typically writes about 100GB per job submission. The NCAR PIO (Parallel I/O) tool is used - about 10X faster than serial I/O for MPAS applications.

The output from a 5-year, quarter-degree tropical channel WRF run is about 40 TB. MPAS has about 50 TB stored on HPSS; WRF has about 100 TB.

8.9.4.4 Project Data

There is an m1178 project directory but with negligible usage.

8.9.5 HPC Requirements in 2017

8.9.5.1 Computational Hours Needed

100 M hours will be needed.

8.9.5.2 Number of Compute Cores

We believe that for MPAS, typical jobs will use ~25K processors. Small numbers (~5) of jobs might be run concurrently. For WRF, typically jobs will use 1,200 – 2,400 processors. Up to eight such simulations may run concurrently.

8.9.5.3 Data and I/O

MPAS: Checkpoint file size will be approximately 100GB. We expect to use community-supported parallel I/O solutions, such as the NCAR PIO. Typical single job runs will generate 250GB of data. Typical simulations will require 20 to 100 job submissions.

WRF: With 12 hour runs, about 30 resubmissions are required to complete a run.

With our compute needs increasing by a factor of 16 between 2012 and 2017, we expect to need a minimum of 3.2 PB of archival storage by 2017 (16X).

8.9.5.4 Memory Required

MPAS: Data-intensive problems that carry on order 100 tracer constituents (for biogeochemistry) with optimal scaling would require ~10 GB per processing unit. For machines with, say, 24 procs per node, this would require approximately 256 GB of memory per node.

8.9.5.5 Many-Core and/or GPU Architectures

MPAS is currently not using GPUs, but we plan to port the code to Titan during this calendar year. We expect to be able to utilize directive-based accelerators in 2013 and beyond. A significant part of a current SciDAC project (Multiscale Earth Modeling, PI-Collins) at LANL is directed toward computational efficiency, including the use of accelerators. Researchers at LANL are also exploring mixed parallelism in the form of MPI-OpenMP and will test parts of the code on accelerators (GPUs) over the next year.

8.9.5.6 Software Applications and Tools

MPAS would benefit from using parallel I/O tools that are used by the broader community.

8.9.5.7 Additional Comments

Improvements in NERSC computing hardware, software and services will lead to improved ability to perform high resolution climate simulations to better predict water cycle changes in the future and establish more robust frameworks for high resolution modeling. With a 32-fold increase in computing time we could perform cloud-resolving simulations over large regions for evaluating cloud parameterizations and sensitivity to model resolution. Increased memory per node, more nodes, larger storage capacity, and shorter wait time in queues would be of great benefit to our research. We also need the PIO library from CESM.

8.9.6 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Million)	6.2	100
Typical number of cores* used for production runs		up to 125,000
Data read and written per run (excluding checkpoint data)		25 TB
Project directory space	Small	small
Archival data	200 TB	3.2 PB
Memory per node		10 GB / task

* “Conventional cores.” For GPUs and accelerators, please fill out section 4.8.

Subsurface Science and Simulation

Principal Investigator: Tim Scheibe (PNNL)

NERSC Repository: m749

8.9.7 Overview and Context

This research is focused on numerical simulation of subsurface porous media flow and transport processes related to a wide range of applications including: 1) contaminant transport, remediation, and risk assessment; 2) geological carbon sequestration; 3) biogeochemical cycling of carbon and other nutrients in soils; 4) geothermal energy production; 5) fossil fuel recovery.

While numerical models of subsurface processes are abundant and widely applied, their reliability for field-scale prediction and design is poor. The primary reason for this is the challenge of scale – that is, the physical and temporal scales at which processes are best understood are typically orders of magnitude smaller than that at which predictions are needed. Direct simulation of small-scale fundamental processes at application scales is currently infeasible due to limitations in 1) our ability to characterize the subsurface environment and 2) computational resources. Approaches to this challenge range from ad hoc (empirical) model calibration to sophisticated (but nevertheless non-unique) optimization to upscaling and multiscale simulation. Our work focuses on the latter, multiscale approaches to simulation that combine models at fundamentally different scales to capture relevant processes in large-scale simulations.

BER supports my high-end computing efforts through three research projects, two of which are ending this fiscal year: 1) A SciDAC-2 subsurface science application and two related science application partnerships that were funded through FY11 but for which some work continued on carryover funds into FY12; 2) a university-led Subsurface Biogeochemical Research (SBR) project led by the University of Massachusetts, which ends in FY12, and 3) PNNL's SBR Scientific Focus Area (SFA) project, which is continuing. In all three of these projects, we have performed work on coupling pore-scale and continuum-scale models of porous media flow, transport and reactions, but with varying applications. I am also involved in two projects that also address these issues, but are funded by internal PNNL investment (LDRD) and focus on carbon cycling and geological sequestration.

8.9.8 Scientific Objectives for 2017

The key scientific goal for 2017 is the development of a production-level set of codes that provide a capability of directly coupling pore- and continuum-scale simulations within a single hybrid framework. We have developed individual at-scale codes during the past 5-6 years, and have performed research-level work on hybrid multiscale coupling, but for this to be more widely applicable the hybrid multiscale methods must be generalized and integrated directly into widely accepted simulation workflow processes. The grand challenge that this objective targets is to narrow the gap between simulations based on fundamental process descriptions and applied simulations with practical engineering applications, that is: science-based predictive modeling.

8.9.9 Computational Strategies

8.9.9.1 Approach

The key computational problem is how to efficiently couple models that are defined at fundamentally different spatial and temporal scales. In our case, we consider pore-scale simulators that describe explicitly the geometry of solids and liquid phases at the scale of individual grains and pores as scientifically sound representations of fundamental processes (multiphase flow, transport, and reaction). We couple pore-scale simulations with traditional continuum-scale (also called Darcy-scale) simulators that represent a porous medium as an effective continuum with macroscopic properties such as porosity, permeability, and dispersivity that do not exist at the pore scale. The key problems involve 1) the mathematical/algorithmic approach to coupling, which depends on the nature of the problem being considered, and 2) the logistics of coupling multiple codes, already each of which is highly complex, in a multiscale simulation workflow.

8.9.9.2 Codes and Algorithms

We work with three primary codes, two at the pore scale and one at the continuum scale.

SPH (Smoothed Particle Hydrodynamics, pore scale): SPH is a discrete particle-based approach to solution of pore-scale flow, transport and reaction processes. Our parallel code was developed at PNNL. We note that it can also be used at the continuum scale but is generally best suited to pore-scale problems that involve moving interfaces such as mineral precipitation/dissolution reactions, biofilm dynamics, and multiphase flow. The algorithm uses a local smoothing function applied to discrete particles to define sums of local forces and rates of mass transfer, which it then steps forward in time. It is implicitly a transient method with (at least slightly) compressible fluids. The primary computational demand lies in the searches required to identify neighboring particles, since there is no connecting grid and particles are free to move; these are made efficient through tree searching algorithms. The method does NOT involve solution of linear systems of equations (except perhaps when multi-component reactions are considered, in which case such solutions would be entirely local). Our code uses the Global Arrays (GA) Toolkit for parallel communications and data management. For details of the code see Palmer, B. J., V. Gurumoorthi, A. M. Tartakovsky, and T. D. Scheibe, "A Component Based Framework for Smoothed Particle Hydrodynamics Simulations of Reactive Fluid Flow in Porous Media," International Journal of High Performance Computing Applications 24(2):228-239, 2010.

TETHYS (Transient Energy Transport Hydrodynamics Simulator): TETHYS is a general Navier-Stokes CFD code developed at PNNL, and has been applied extensively to pore-scale problems with complex fluid-solid boundary geometries. It is a single-phase flow solver, and can be used to solve either transient or steady-state problems. Standard finite volume numerical methods are utilized, and are implemented using the GA Toolkit and PETSc.

STOMP (Subsurface Transport Over Multiple Phases): STOMP (and the scalable version eSTOMP) is a continuum-scale multiphase flow and reactive transport simulator used widely for environmental and energy applications. eSTOMP was recently recoded using GA and PETSc, and has demonstrated high scalability to petascale platforms. It uses standard finite difference numerical methods, with linear solvers at the heart and non-linear problems solved by Newton iteration.

We have explored a variety of multiscale coupling approaches, including both hierarchical and concurrent methods. Our current work focuses on a hierarchical approach that uses short bursts of microscale (pore-scale) simulation to inform (update parameters for) larger time steps of the macroscale simulator. Our approach is described in Tartakovsky, A. M. and T. D. Scheibe, "Dimension reduction method for advection-diffusion-reaction systems," Advances in Water Resources, 34(12): 1616-1626, doi:10.1016/j.advwatres.2011.07.011, 2011. In this approach, we are loosely coupling the SPH and STOMP simulators using the SWIFT workflow management tool and a series of custom python scripts. The workflow management system requests a group of compute nodes from the system (e.g., Hopper) and then manages the process of executing multiple SPH and STOMP simulations on those nodes and exchanging information through I/O files as the overall simulation proceeds.

8.9.10HPC Resources Used Today

8.9.10.1 Computational Hours

In addition to our NERSC allocation, we also utilize (as appropriate given project scopes and availability) smaller allocations on local PNNL machines, specifically the PNNL Institutional Computing system "Olympus" and the EMSL supercomputer "Chinook." In particular we have utilized Chinook for SPH code development and testing, and then use NERSC allocations for larger production runs once the codes have achieved sufficient scalability.

eSTOMP development was performed largely by others under different project funding, and they have a large allocation on the ORNL Jaguar machine. However, my project and direct collaborators do not have access to that allocation.

During AY2012 this work used 3.7 million hours at NERSC in approximately 1,700 jobs, run primarily on Hopper,

8.9.10.2 Compute Cores

Parallel concurrencies are in the range 1-1,024 nodes (24 – 24,576 cores).

8.9.10.3 Data and I/O

The project has about 6 TB of data stored on HPSS.

8.9.10.4 Project Data

None.

8.9.11HPC Requirements in 2017

An interesting and potentially transformative approach is a new paradigm for subsurface modeling – directly coupling pore- and continuum-scale codes in a single simulation domain. This approach spans the scale gap between fundamental process representations and applications, maintains reasonable efficiency, and takes advantage of multiple levels of concurrency.

8.9.11.1 Computational Hours Needed

Significant advances would be made with about 120 million hours of compute time.

8.9.11.2 Number of Compute Cores

The runs will span multiple levels of concurrency.

8.9.11.3 Data and I/O

.I/O and data requirements will be larger, but still relatively small; requiring perhaps on the order of 100-200 TB of archive storage.

8.9.11.4 Many-Core and/or GPU Architectures

SPH could take advantage of GPUs but at the current time we do not and codes are not ready.

8.9.11.5 Software Applications and Tools

For parallel I/O (SPH) we are currently using H5PART. Many of our tools rely on GA (Global Arrays) and PETSc libraries.

8.9.11.6 HPC Services

Visualization while the simulation is running is important.

8.9.11.7 Data Intensive Needs

Workflow and data management will be critical to our hybrid simulations, as they will involve running many individual simulations in coordinated manner, with exchange of information between different simulations either through loose coupling (file I/O, current approach) or tight coupling (direct communication, potential future approach). We have found in our current test runs that visualization costs are high – we like to generate multiple plots at each time output point for evaluation during- and post-execution and these tend to take a long time, perhaps because of inefficiency in the workflow management tool we are using?

8.9.12 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Million)	3.7	120
Typical number of cores* used for production runs	SPH: 100-2,000 TETHYS: 4,000 STOMP: 100-1,000 Hybrid: 1,000	SPH: 1,000-2,000 TETHYS: 4,000 STOMP: 1,000-10,000 Hybrid: 50,000
Maximum number of cores* that can be used for production runs	SPH: 9,000 TETHYS: 5,000 STOMP: 131,000 (on Jaguar) Hybrid: 2,000	SPH: 50,000 TETHYS: 50,000 STOMP: 200,000 Hybrid: 200,000+
Checkpoint data written per	SPH: N/A (Transient	All < 1 TB

run	output used for restart if needed) TETHYS: N/A (Transient output used for restart if needed) STOMP: Small (2.5 GB)	
Data read and written per run (excluding checkpoint data)	SPH: 0.6 TB TETHYS: < 1 TB STOMP: < 1 TB Hybrid: 0.3 TB	SPH: 5 TB TETHYS: < 1 TB STOMP: < 1 TB Hybrid: 10 TB
Maximum I/O bandwidth (excluding checkpoint data)	SPH: 10 GB/sec TETHYS: Unknown STOMP: 0.8 GB/sec	Unknown
Project directory space	< 10 TB	1,000 TB
Archival data	6 TB	200 TB

* “Conventional cores.” For GPUs and accelerators, please fill out section 4.8.

9 Biological Sciences Case Studies

9.1 Overview

Dr. Susan Gregurick, Program Manager, Biological Systems Science Division, DOE

The Biological Systems Science Division supports a diverse portfolio of fundamental research and technology development to achieve a predictive systems-level understanding of complex biological systems to advance DOE missions in energy and the environment. By integrating genome science with advanced computational and experimental approaches, the Division seeks to gain a predictive understanding of living systems, from microbes and microbial communities to plants and other whole organisms. This foundational knowledge serves as the basis for the confident redesign of microbes and plants for sustainable biofuel production, improved carbon storage and contaminant remediation. NERSC is the flagship provider of HPC resources in support of these efforts.

Research into systems biology and the DOE Genomic Science program aimed at identifying the foundational principles that drive biological systems. These principles govern the translation of genetic codes into integrated networks of catalytic proteins, regulatory elements, and metabolite pools underlying the functional processes of organisms. It is these dynamic interactions of nested subsystems that ultimately determine the overall systems biology of plants, microbes, and multispecies communities. The ultimate goal of the Genomic Science program is to achieve sufficient understanding of the fundamental rules and dynamic properties of these systems to develop predictive computational models of biological systems and tools for rational biosystems design.

The Genomic Science program research also brings the -omics driven tools of modern system biology to bear on analyzing interactions between organisms that form biological communities and with their surrounding environments. Understanding the relationships between molecular-scale functional biology and ecosystem-scale environmental processes illuminates the basic mechanisms that drive biogeochemical cycling of metals and nutrients, carbon biosequestration, and greenhouse gas emissions in terrestrial ecosystems or bioenergy landscapes.

The major objectives of the Genomic Science program are to:

- Determine the molecular mechanisms, regulatory elements, and integrated networks needed to understand genome-scale functional properties of microbes, plants, and interactive biological communities.
- Develop -omics experimental capabilities and enabling technologies needed to achieve dynamic, systems-level understanding of organism and/or community function.
- Develop the knowledgebase, computational infrastructure, and modeling capabilities to advance predictive understanding and manipulation of biological systems.

NERSC supports fundamental research in the redesign of microbial metabolic and regulatory process to harness their potential in the conversion of biomass to biofuels. This work requires the sequencing and annotation of complete microbial and plant genomes, elucidation of metabolic pathways and simulations of complex biological processes.

Computational simulations run at NERSC are unraveling functional annotations of unstructured proteins from analysis across genomic and structural relationships. This work requires the comparison across large datasets as well as the molecular dynamic simulation of complex cellular processes. NERSC provides the resources to allow researchers to understand protein dynamics and the role these play in creating large nanoassemblies. NERSC also provides key support for the computational infrastructure of the DOE Joint Genome Institute and the DOE Systems Biology Knowledgebase. The fully functional Knowledgebase will include storage, retrieval, data management, and integration of systems biology data, and enable new knowledge acquisition and management through free and open access to data, analytical software, modeling tools, and information for the research community.

In summary, the HPC requirements for computational biology and bioinformatics are those that will enable biological simulations to be performed with both greater accuracy and complexity so as to guide experimentation that leads to discovery of new properties for biofuel production or understanding environmental processes. Computations at NERSC advance our ability to predict an organism’s phenotype from a genomic sequence and require an integration of computational modeling, algorithm and software development with new advances in hardware architecture.

9.2 Molecular Simulation in the Energy Biosciences

Principal Investigator: Jeremy Smith (ORNL and UTK)

Case Study Author: Loukas Petridis (ORNL)

NERSC Repository: m906

9.2.1 Overview and Context

High-performance computer simulation has a significant role to play in the energy biosciences in obtaining an understanding of physical processes leading to biological function. Molecular mechanical techniques can provide atomic detailed insight into processes at the core of research into bioenergy, bioremediation, carbon capture, neutron scattering and other critical research missions.

Work in progress has derived computational simulation models for use in industrial hydrolysis of plant biomass to glucose for the production of fuels and chemicals via microbial fermentation. This effort has already shed light on the phenomenon of 'biomass recalcitrance,' or resistance to hydrolysis into sugars, which is a bottleneck in biofuel production. Further biofuel research involves simulation work on the functioning of molecular machines participating in cell signaling in microbes and plants, lignocellulosic biomass and enzyme complexes that hydrolyze cellulose chains. The fate of mercury in streams contaminated in DOE sites is strongly influenced by bacterial enzymes, and current research is aimed at understanding how these enzymes function. Neutron scattering will play a significant role in the energy-related materials and biosciences, and the synergy between neutron scattering and high-performance computing is critical in examining biological function over a range of relevant time and length scales.

9.2.2 Scientific Objectives for 2017

Critical issues in 21st century biosciences concern the complex interplay of the molecular systems within cells. However, understanding how structure and dynamics relate to function requires spatiotemporal characterization spanning decades of time and length scales. The overarching aim of our project is to employ simulation and neutron scattering techniques to obtain high-resolution spatial and temporal information on biological processes and thus demonstrate the role that the interaction between the members of complexes plays in defining their function. Examples of systems to be studied in 2017 include: Large multi-subunit complexes related to cell signaling pathways; the interaction between bacteria and lignocellulosic biomass and single proteins in aqueous solution or in membrane environment simulated for tens of microseconds.

9.2.3 Computational Strategies

9.2.3.1 Approach

Molecular Dynamics (MD) simulations, involving stepwise integration of the equation of motion for a system of classical particles through an empirically-derived potential energy function, will be performed on biomolecular systems of a wide range of sizes: (i) single proteins and/or single plant cell wall polymers. Such systems involve multi-subunit molecular machines and/or ensemble techniques that enhance the sampling of biologically relevant rare events and require long-timescale simulations reaching tens of microseconds. (ii) multi-component systems such as enzymes interacting with biomass. Quantum chemical calculations will be performed when necessary in the optimization of the force fields. Improvements in the strong scaling of MD codes that allow longer timescales to be simulated are critical for the future success of biological MD simulation.

9.2.3.2 Codes and Algorithms

Algorithms:

- Integration of coupled differential equations (equations of motions): velocity verlet;
- *N*-body algorithms with neighbor-list;
- Grid-based electrostatics: Particle Mesh Ewald (FFT) or multigrid (multi-level real-space);
- Domain-decomposition and force decompositon for multi-level parallelization.

Codes: We mostly use Gromacs and NAMD. Gromacs is the fastest code on a single-core basis. NAMD and LAMMPS scale better than Gromacs but aren't as fast. Gromacs also has good analysis tools and built-in ensemble methods.

9.2.4 HPC Resources Used Today

9.2.4.1 Computational Hours

We have a 30-M hour allocation at Jaguar through an INCITE award. During AY 2011 our allocation at NERSC was 11 M hours and 100% of it was used. In AY2012 our allocation is 7.5 M hours and we used 7.7 M hours (using job discount queues). During AY2012 over 25,000 jobs were run at NERSC.

9.2.4.2 Compute Cores

The number of cores that can be used depends on simulated system size: large systems (~10M atoms) display better strong scaling than smaller systems (~100k atoms). We have scaled our code up to 300,000 cores for a 100M-atom system. For production runs at NERSC today (on Hopper) we typically use between 200 and 2,000 cores for system sizes ranging from 50k to 500k atoms. We sometimes have multiple jobs running concurrently, usually between 5 and 10.

9.2.4.3 Data and I/O

In single-precision (Gromacs) runs checkpoint file sizes are given by: number of atoms * 6 * 4 bytes. Therefore, file size range from about 1.2 MB to 12 MB. If double-precision data are written (NAMD), the maximum file size is about 24 MB. Usually less than one percent of our total runtime is spent writing checkpoint files.

For a 1-hour run data in the range 0.2 to 2.0 GB are produced. However, output can be written in a compressed format, which requires: (number of atoms) * 3 * (1 byte) * (0.3 frames/s). Therefore, bandwidth ranges from about 50 KB/sec to 500 KB/sec.

We have 47 TB of data stored in the NERSC HPSS system in 2012, more than double what we had stored in 2011 and ten times more than in 2010.

We have an m906 project directory in the NERSC Global File System, but it currently has very little stored in it (~125 GB).

9.2.5 HPC Requirements in 2017

9.2.5.1 Computational Hours Needed

To continue our current work and perform large-system/ensemble simulations will require 360 M hours in 2017. We do not know if some of this time will be available through an INCITE award.

9.2.5.2 Number of Compute Cores

We expect to be able to use 2 K – 20 K cores for conventional jobs and up to 600 K cores for large-scale/ensemble simulations. We expect to have ten jobs running concurrently for our conventional simulations and up to 500 for ensemble calculations.

9.2.5.3 Data and I/O

In 2017 we expect that checkpoint files will consist of 12 to 120 MB for conventional jobs and 1.2 GB for large-systems. We expect that a minimum of 20 MB/s will be required to support checkpointing; we are willing to devote no more than 1% of total run time to this.

In 2017 we expect to write output in a more efficient way, e.g., by writing the coordinates of water molecules less frequently than solute. One-hour jobs will probably write about 2 – 20 GB of data when complete. Our Project Directory needs will probably expand to 2 TB and archival storage will be about 100 TB.

9.2.5.4 Memory Required

Memory requirements are minimal for biological MD simulations.

9.2.5.5 Many-Core and/or GPU Architectures

Our codes are currently compatible with hybrid CPU/GPU architectures. The GROMACS code is ready and has a very efficient implementation compared to other MD codes. On current hardware the speedup when comparing a CPU+GPU node to a single-CPU node is <2x. Thus, currently two traditional CPUs are faster than CPU+GPU. By 2017, however, we expect CPU+GPU nodes to be about twice as fast as two traditional CPUs. Although a speedup of a factor of two will be beneficial, we do not expect that the use of accelerators will revolutionize biological MD simulations.

9.2.5.6 Software Applications and Tools

Applications: Gromacs, NAMD, VMD

Development: C++, Boost, libxml, Cmake, Git, FFTW, Cuda (or equivalent), Eclipse/PTP

9.2.5.7 HPC Services

Consulting and account support.

9.2.5.8 Time to Solution and Throughput

Long-timescale simulations typically take about three months of CPU time and are therefore executed as a series of many dependent jobs. Therefore, for the simulations to be completed in a timely manner, a scheduling policy that allows dependent jobs to "age" in the queue is required.

9.2.5.9 Additional Comments

Strong scaling of MD simulations is limited by network latency. Therefore a NERSC cluster of size similar to that of Carver and with network with lowest latency available would speed up biological MD simulations considerably.

MD simulations are globally synchronized. Therefore, the lowest network connection is slowing down the entire simulation. A task placement that ensures that computing nodes allocated to the job can communicate with low network latency would thus speed-up simulations.

In terms of hardware, tightly integrated (ideally shared cache) CPU + GPU/Manycore systems that enable a fine-grain split of work between CPU and GPU would provide considerable speed up in our calculations.

9.2.6 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Million)	7.7	360**
Typical number of cores* used for production runs	200 - 2,000	2,000 – 20,000 (up to 600,000)
Maximum number of cores* that can be used for production runs	300,000	1 M
Checkpoint data written per run	1.2 to 24 MB	12 to 120 MB
Checkpoint bandwidth	2 MB/sec	20 MB/sec
Data read and written per run (excluding checkpoint data)	0.2 – 2 GB	2-20 GB
Maximum I/O bandwidth (excluding checkpoint data)	500 KB/sec	5 MB/sec
Project directory space		2 TB
Archival data	47 TB	100 TB

* "Conventional cores." For GPUs and accelerators, please fill out section 4.8. **Less if we receive an INCITE award.

9.3 Computational Predictions of Transcription Factor Binding Sites

Principal Investigator: Harley McAdams (Stanford)
Case Study Author: Mohammed AlQuraishi (Stanford)
NERSC Repository: m926

9.3.1 Overview and Context

The project is developing a novel way of computationally predicting where proteins bind on the genome, with accuracy approaching experimental measurement. This is a classic problem in biology, since many cellular processes are regulated by DNA-binding proteins that control gene activity. Prediction of protein binding sites with useful accuracy has been elusive, forcing scientists to rely on labor-intensive and costly laboratory experiments. With reliable computational prediction, it will be possible to determine genetic regulatory pathways in any organism, including humans, much more quickly. These advances can lead to new approaches to other types of molecular interactions, including medically important enzymatic reactions and metabolic pathways important in energy production and synthetic biology.

We have developed an algorithmic approach for predicting protein-DNA interactions based on the structures of protein-DNA complexes. This approach computes the binding energies of a protein to different DNA sequences, and uses these computed energies to predict the DNA-binding motif of a protein. The key innovation of our approach is how the underlying energies are computed. We employ a compressed sensing approach to derive a model of protein-DNA energetics strictly from empirical data. This is contrast to existing approaches that relied on theoretical models of biophysical interactions. Our compressed sensing approach is both data- and compute-intensive, requiring the storage of large amounts of sequence and structural data, and the ability to perform convex optimization and Monte Carlo simulations on a large scale.

9.3.2 Scientific Objectives for 2017

Our five-year objectives are two-fold. To broaden the applicability of our approach to the full spectrum of protein-DNA interactions, including those involving protein families that are not part of the helix-turn-helix superfamily, which we are currently restricted to. The second objective is to broaden our approach to tackle protein-protein interactions, including those involved in forming multi-protein complexes.

To tackle a broader array of protein-DNA interactions largely represents a data challenge. A broader data set of protein-DNA structures will be used to derive a general model of protein-DNA energetics that is not specific to a particular protein family. Doing so will increase the data storage requirements as well as the computing requirements.

Tackling protein-protein interactions will present additional algorithmic challenges. Unlike the case of protein-DNA interactions, in which the region of the protein responsible for DNA-binding is known a priori, in protein-protein interactions the regions of the proteins responsible for binding are unknown. Consequently predicting protein-protein interactions involves first identifying the binding regions involved in an interaction, and then computing the binding energy of the proteins based on the predicted binding regions. The algorithmic challenges involved in predicting protein-protein interactions will present significant increases in our computing needs.

9.3.3 Computational Strategies

9.3.3.1 Approach

There are three basic computational problems in our project. The first is to derive an energy model of protein-DNA or protein-protein interactions from empirical data. To do so involves solving a convex optimization problem, specifically a constrained version of logistic regression with L1 regularization. We use standard convex optimization approaches to solve this problem, but the matrices involved can be quite large, requiring significant memory resources.

The second computational problem is to use an energy model to compute the binding energy of new protein-DNA complexes. This involves generating large numbers of in silico DNA molecules, and then computing the binding energy of the protein to these molecules. In the future, this may also involve running short MD simulations to energetically relax the protein-DNA molecule.

The third computational problem is the identification of protein regions that are involved in protein-protein interactions. Our approach is to search large databases of protein sequences and structures for overrepresented motifs, while simultaneously minimizing a variety of global objective functions that encourage finding sparse solutions. This problem is non-convex, so we use Monte Carlo sampling techniques to approximately solve it.

9.3.3.2 Codes and Algorithms

We use the R computing environment and Matlab + CVX to solve convex optimization problem. We use Mathematica to carry out general computing procedures and visualization. Finally we use custom C code for running MC simulations.

9.3.4 HPC Resources Used Today

9.3.4.1 Computational Hours

During AY2012 members of this repository ran about 100 jobs on Carver, consuming 536,000 hours and using between one and 13 nodes (up to 104 cores). However, a significant amount of computation (primarily via Mathematica) was also performed on the NERSC Euclid analytics resource, for which usage is not measured.

9.3.4.2 Compute Cores

Depending on the type of code running, it can be from a few cores all the way to 384 cores. Typically the runs are around 80 to 88 cores. In general we have multiple jobs running, because we found that it is easier to manage the jobs if they are split up into smaller pieces. We usually have 4-6 jobs running simultaneously.

9.3.5 Data and I/O

A typical run writes about 50GBs over a 24-hour period, so about 0.5MB/sec.

9.3.6 HPC Requirements in 2017

9.3.6.1 Computational Hours Needed

We anticipate needing about 30 million hours at NERSC in 2017.

9.3.6.2 Many-Core and/or GPU Architectures

No our codes are generally not. We have only begun recently to look into GPUs.

9.3.6.3 Software Applications and Tools

Mathematica, Matlab, R.

9.3.6.4 Time to Solution and Throughput

Will probably have multiple jobs running for ease of management.

9.3.6.5 Data Intensive Needs

Access to significantly more disk space would be very helpful. The current 20TB constraint is very limiting.

9.3.6.6 Additional Comments

Primarily would like more disk space, integrated checkpointing services, and better support/more licenses for Mathematica and Matlab.

9.3.7 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Million)	0.536	30
Typical number of cores* used for production runs	80	3,000
Maximum number of cores* that can be used for production runs	384	6,000
Checkpoint data written per run	0.05TB	0.5 TB
Checkpoint bandwidth	1 GB/sec	10 GB GB/sec
Data read and written per run (excluding checkpoint data)	0.05TB	0.5 TB
Maximum I/O bandwidth (excluding checkpoint data)	0.5 MB/sec	5 MB/sec
Archival data	0 TB	0 TB
Memory per node	2 GB	6 GB

Aggregate memory	0.6 TB	6 TB

* “Conventional cores.” For GPUs and accelerators, please fill out section 4.8.

9.4 Joint Genome Institute

Principal Investigators: Edward Rubin and Victor Markowitz (LBNL)
Case Study Authors: Shane Canon, Rob Egan, David Goodstein, Victor Markowitz (LBNL)
NERSC Repositories: m342 and m1045

9.4.1 Overview and Context

In the next decade JGI will adopt new genomic capabilities while continuing to maintain and expand its massive-scale sequencing capabilities in order to accommodate expected improvements in sample throughput and increase in demand for sequence generation. As genomic datasets increase in scale and complexity, their systematic biological interpretation is critical for enabling scientific studies.

Genome and metagenome "raw" sequence data are transformed into biologically meaningful information using computational tools and pipelines. A comprehensive sequence data interpretation process employs the integrated data context of an expanding universe of genome and metagenome sequence datasets, and involves incorporation of complementary 'omics' technologies for validating the coherence of biological information. Data interpretation is also inherently iterative, since repeating one or several of the processing stages in the presence of ever-growing datasets gradually improves the breadth and depth of biological information.

Sequence data interpretation and integration processes must be scalable to cope with the increase in the rate of sequencing of genomes and metagenomes, the size of metagenome datasets generated using new sequencing platforms, and the diversity of 'omics' datasets. The estimated size of datasets generated with new genome sequence technology platforms are expected to grow faster than the computing resources available to JGI. Addressing this challenge requires leveraging computing resources and developing scalable and efficient data processing tools.

Analysis of high throughput sequence data in an 'omics' context requires High Performance Computing (HPC) capabilities set in a High Throughput Computing (HTC) environment. JGI relies on Lawrence Berkeley National Lab's National Energy Research Scientific Computing Center (NERSC) for supporting its High Performance & Throughput Computing (HPTC) needs, with a compute cluster and large capacity distributed file system maintained by NERSC at its core. Leveraging HPC platforms at NERSC and other DOE Leadership Computing Facilities will require refactoring sequence data processing and integration pipelines to run efficiently on these platforms.

9.4.2 Scientific Objectives for 2017

In the next decade, DOE JGI will evolve from a production sequencing facility to a next-generation genome center, offering a diversity of capabilities that will complement massive-scale sequence production to meet the scientific needs of energy and environmental researchers. Key areas of new development and expansion include:

- **Large-scale rapid DNA synthesis and genomic manipulations**. To accelerate the linking of sequence to function, JGI will develop new approaches for designing and creating DNA fragments encompassing genes and larger segments of DNA. These capabilities will be made available to users for testing of genomics-derived hypotheses, creation of synthetic pathways and for the functional exploration of metagenomic and other sequence data sets.

- **Massive-scale and customizable sample processing**. The exponential growth in sequence data generation fortunately mirrors the expansive needs of future large-scale environmental and systems-based science. JGI will develop custom large-scale sample-processing capabilities including the implementation of automated DNA/RNA extractions able to process tens of thousands of samples and large-scale single-cell and single-chromosome isolation techniques.
- **Comprehensive genome annotation and data integration**. JGI will develop advanced data processing and integration techniques enabling data interpretation across the rapidly expanding universe of genomic, metagenomic and functional genomic datasets. These capabilities will allow refining both structural annotations (the location of functional elements within sequences) and functional annotation (the function of these elements in the context of biological systems), raising the level of "interpreted" data provided to JGI users.
- **Organization of mission-oriented user communities**. As genome and functional genomic projects become larger and more complex, JGI will play an expanded role in organizing communities around problems of central importance to DOE. JGI will help coordinate activities of diverse groups of scientists, ensure access to state-of-the-art genomics capabilities and strategies, and facilitate data sharing and integration in order to speed progress toward solving DOE's most pressing challenges in alternative fuels, carbon management and climate and environmental remediation.

9.4.3 Computational Strategies

9.4.3.1 Approach

Eukaryotic and isolate prokaryotic genome analysis focuses on the use of massive amounts of relatively small (tens to hundreds of base pairs) DNA and RNA sequence fragments to a) reconstruct large-scale (ideally chromosome-scale) drafts of the organism's genome, b) identify and correctly model the structure of protein-encoding genes and other conserved functional elements in the genome, using homology (similarity to known elements in closely related species) and ab-initio methods, c) infer the biological function of the elements identified in (b), using homology to sequences of known function or correlated expression (RNA-seq data) with elements known to participate in particular biological processes, d) model the evolutionary history of the organism's genome using phylogenetic methods, and e) analysis of the genomic diversity found in different individuals representing the same organism, arising from differences in climatic, geographic, nutrient or toxin availability, and lineage.

The computational approach taken by the JGI for most of these analyses involves using wherever possible existing third party applications, integrating them into workflows and pipelines that exploit the "embarassingly parallel" nature of the vast majority of the relevant algorithms (the exceptions being eukaryotic genome reconstruction (assembly) and some phylogenetic methods, which require access to "complete" data sets during certain calculations). The main difference between eukaryotic and prokaryotic analyses is the driver in computational scale. Eukaryotic genomes are typically hundreds to thousands of times larger than prokaryotic genomes, which are arranged into multiple chromosomes, each of which may have undergone complete or partial duplication and rearrangement in the evolutionary history of the organism. Eukaryotic genomes also tend to have extensive repetitive content, which greatly complicates their reconstruction. The eukaryotic gene complement is also typically five to ten times larger than the corresponding prokaryotic set, and each gene has a potentially more complex and difficult to model structure (i.e., the presence of intronic, non-coding regions). Eukaryotic computational complexity is thus driven by the size, duplication history, and repetitive nature of the genome, as well as the number and structural complexity of typical eukaryotic genes.

Prokaryotic computational complexity has two main drivers: the high degree of evolutionary divergence between species (driving up the size of the homolog databases that need to be searched in most analysis steps), and the sheer number of genome projects taken on by the prokaryotic genome program (on the order of 1,000 - 3,000 per year compared to 5-10 plants and 25-50 fungal genomes). Because the vast majority of prokaryotes cannot be cultured, the only way to analyze the genomes of these organisms is by sampling the environment and either sequencing the entire community together and/or isolated cells individually. The size, complexity and variability of a community of organisms poses an assembly and analysis challenge even greater than that of some eukaryotic genomes, primarily because the mixture of highly similar individuals, strains, species and families introduce more errors and uncertainty into the assembly and annotation.

9.4.3.2 Codes and Algorithms

The majority of work performed in the analysis of isolate genomes and microbial communities is based on sequence similarity algorithms. They either rely on pairwise sequence similarity comparisons (e.g. blast, usearch, Smith & Waterman), mapping of sequences on reference genomes (e.g. bwa, bowtie) and comparison of sequence to models such as Hidden Markov Models and Covariance Models (e.g. hmmsearch, cmsearch). The code used for these comparisons is provided by third parties, either as open source or licensed. In all cases a query sequence (either protein sequence or nucleotide sequence) is compared to a reference database.

Additional algorithms used for the comparative analysis of multiple genomes/proteomes include multiple sequence aligners (e.g., clustalw, muscle, VISTA) and phylogenetic tree builders (e.g., RAxML, Tree-Puzzle, Mr. Bayes). A summary of code characteristics appears in the table below.

Code	What it does	Used for	scaling with sequence number/length	HighMem?
BLAST, USEARCH	fast pairwise alignment of moderately dissimilar sequences	protein-to-genome alignment to seed gene finding, input to gene family construction, homolog identification	the algorithm scales linearly with the size of query sequences for a given database size. Typically though databases grow linearly with time resulting in an exponential computational time growth	The memory requirement depends on the size of the database. USEARCH hash higher memory requirements than blast (approx 2X the size of the database) Efforts to parallelize blast

				typically result in splitting the databases in smaller pieces which result in smaller memory needs
Smith-Watermann	accurate pairwise alignment of dissimilar sequences	identifying protein homologs		
bowtie, cufflink, bwa	pairwise alignment of highly similar sequences	gene expression (RNA-Seq) analysis, diversity (resequencing) analysis	the algorithm scales linearly with the size of query sequences for a given database size. The database size is typically stable for a given project	Memory requirment depends on the size of the database.
Hmmer	Hidden Markov Model creation/alignment	motif/domain identification, detection of very distant homologies	the algorithm scales linearly with the size of query sequences for a given database size.	N
Meraculous	de Bruijn graph traversal	genome assembly	Scales with complexity of the genome & graph	N - Distributed Computing
AllPaths	de Bruijn graph traversal	genome assembly	Scales with complexity of the genome & graph	Y - SMP
SOAP DeNovo	de Brujn graph traversal	genome assembly	Scales with complexity of the genome & graph	Y - SMP
ABySS	de Bruijn graph traversal	genome assembly	Scales with complexity of the genome & graph	N - Distributed Computing
Ray	de Bruijn & Overlap graph traversal	genome assembly	Scales with complexity of the genome & graph	N - Distributed Computing
clustalw, muscle	multiple sequence alignment	input for building Hidden Markov Models, input for tree building,		
RAxML, Tree-Puzzle,				
Mr. Bayes	maximum likelihood estimation of	analysis of evolutionary history of gene		Y

	phylogenetic trees	families and species		
Infernal	Covariance model creation and search	identification of nucleotide sequences by structural similarity to a given database	the algorithm scales linearly with the size of query sequences for a given database size.	N

9.4.4 HPC Resources Used Today

9.4.4.1 Computational Hours

During AY2012 users of NERSC respository m342 (JGI) ran approximately 2,100 jobs, mostly on Hopper but also on Carver, that consumed about 11 M hours, and used as many as 3,000 Hopper nodes (72,000 cores) per job. During the same period users of repository m1045 (IMG) ran approximately 1,100 jobs that consumed about 21 M hours, all on Hopper, using as 4,097 nodes (98,328 cores) per job.

In addition to the NERSC allocation, the JGI also has a private computational resource managed by NERSC called genepool. For the majority of FY2012 the genepool cluster was comprised of 530 compute nodes representing 4,544 cores. The configuration of the machine for FY2013 will include the addition of 3,520 cores (total 8,064 cores). One important reason that the JGI uses genepool for a great deal of its computational needs is that genepool has much more memory / core than the other computational resources JGI has access to at NERSC. All nodes on genepool have at least 5 GB / core of RAM, whereas the commodity nodes on carver have 2GB / core. Furthermore there are 24 nodes that have increased memory capacity allowing the JGI to perform some calculations that use extreme amounts of on-node memory (one node has 2 TB of RAM).

From June 2012 - August 2012 the JGI used about 5.2M CPU hours on genepool, for an estimated usage of 20.5 M hours per year. We view this usage as much lower than expected usage because the cluster was being constructed during this period from several smaller clusters. There are a large variety of codes executed on genepool.

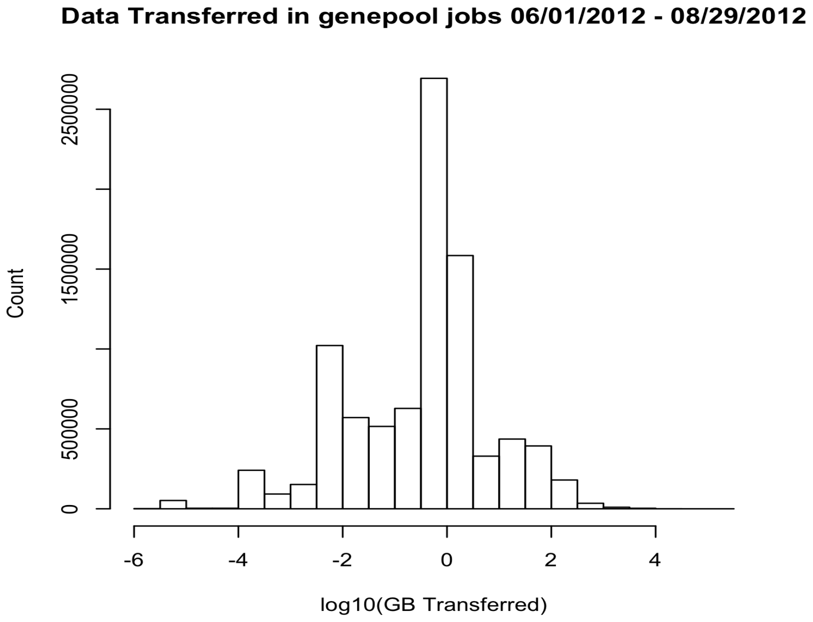
9.4.4.2 Compute Cores

Some 97% of the jobs executed on genepool use a single core and most are presently implemented as sequential codes. To make efficient use of the system, datasets are subdivided and run in parallel. Genepool often has thousands of serial jobs running simultaneously. The parallel jobs on genepool almost exclusively use on-node shared-memory parallelism (p-threads/OpenMPI). Some of the jobs that are presently executed as serial codes do have some threaded capabilities, but are operated sequentially due to inefficiencies in the parallel implementation.

The jobs executed on the NERSC systems, Hopper and Carver, typically used much more parallelism. The JGI codes, however, are quite diverse - most using just 24 cores (a single hopper node) for many instances of a serial code, with a few codes scaling up to 76,000 cores.

9.4.4.3 Data and I/O

We have not kept good I/O data per type of program run on the Genepool system to this point. The JGI team is building a workflow tool that will track various performance metrics, including I/O. The plot below is a measure of the total bytes read and written during the course of a job, but does not account for the time. We can see from the plot that most jobs run on Genepool read and write approximately 1GB of data. The maximum for a job run in June, July and August was 212,852 GB.



9.4.4.4 Project Data

JGI has its own entire project file system, with 2012 usage of about 700 TB.

9.4.5 HPC Requirements in 2017

9.4.5.1 Computational Hours Needed

JGI currently expects its minimum sequencing output to increase an average of 30-50% per year over the next five years, though these numbers are extremely hard to estimate. Historically, the genomic output has increased 500% per year but we believe that certain market pressures on the dominant Illumina technologies have stabilized and will track closer to "Moore's Law" in the future. The IT budget for JGI is expecting to remain flat over this same time period. Therefore, typical "Moore's Law" improvements should allow JGI to keep pace with essential data processing from this growth. However, we are increasingly seeing new types of data analysis emerge, like Pacific Biosciences and Oxford Nanopore. As these new technologies emerge we will expect them to improve at similar rates to Illumina in the first five years (i.e., 5x per year). This growth is often downstream of the typical analysis pipelines where data is being synthesized and analyzed in new ways.

9.4.5.2 Archival Data Storage

The JGI leverages NERSC's HPSS for every byte of raw data that is produced by the sequencer as both supplemental storage and disaster recovery. Immediately after the initial, automated, processing this raw data is archived at HPSS for disaster recovery. When that raw data is eventually aged off of the spinning disks, it is then verified in HPSS, duplicated in HPSS for disaster recovery and then purged from disk. If the data is required for subsequent analysis in the future, it is automatically retrieved back from HPSS. For 2012 we estimate storing about 500 GB to HPSS, and assuming a modest 30% growth in sequencer output on a flat sequencing budget, this would translate to 1.8 PB stored in 2017, for a total of 6.2 PB in additional tape storage. We currently have 1.3 PB in HPSS, so our total 2017 need is 7.5 PB.

Additionally the JGI has recently enacted a new data management plan that requires that older project data be archived in HPSS before any new space is allocated on our recently purchased file system. This has had a dramatic effect in curtailing the exponential growth of data stored on our most expensive disks, primarily because it encourages users to clean up temporary and working files before storing the relatively small final product and documentation to tape.

9.4.5.3 Memory Required

The two most difficult computations that require large amounts of memory are large-scale assembly and phylogenetic calculations. While there exist distributed memory applications for both algorithms, it is generally agreed that the best assemblers and best phylogenetic tree builders are limited to the memory and cores within a single machine. The JGI has purchased a handful of very large memory machines ranging from 256GB to 2TB in order facilitate the assembly of eukaryotes, fungi and metagenomes. Looking towards 2017, either there will have to be a new surge in the development and adaptation of these algorithms into a distributed framework or even larger machines will need to be purchased in order to tackle the upcoming datasets. JGI continues to evaluate distributed memory assemblers. These cannot yet replace the current state-of-the-art shared memory assemblers for all projects, but there are encouraging signs.

9.4.5.4 Many-Core and/or GPU Architectures

A growing number of tools and applications used in genomics are being threaded in order to take advantage of processors containing many core. Examples include BLAST, uclust, and HMMER. In addition, a small number of tools have been ported to GPU, including BLAST, and HMM. However, many of these implementations do not achieve similar efficiencies compared to serial executions. Many of the tools require investment to adequately leverage multicore processors and computer architectures containing elements like GPUs and the Intel Xeon Phi. Given the growing computational demands of bioinformatics, it is critical that the underlying tools keep pace with the technology. While JGI has a number of expert staff that could contribute to this effort, the problem is much larger than one center. A coordinated effort, such as SciDAC, is needed to make progress on this front.

9.4.5.5 Software Applications and Tools

For the most part every piece of software that the JGI uses and/or creates requires a Linux compatible system with gcc and access to a large file system.

Additionally the convenience of the "Cloud" as a deployable VM appliance has convinced many software packages to be delivered as such (PacBio, OpenGenome, Galaxy, K-base), and some, such as contrail, exclusively within an existing cloud framework. So if this trend continues, it will become increasingly important to be able to both execute an arbitrary Virtual Machine Image and to have many of the now-common cloud based services available, such as a Map-Reduce framework, a block storage system and a Key-Value storage system.

9.4.5.6 Time to Solution and Throughput

JGI will continue to depend on NERSC for development and optimization of batch processing strategies and software allowing submission and tracking of large numbers of independent, concurrent job streams.

9.4.5.7 Additional Comments

The following items are some possible additional services that JGI may require in the future:

- User controllable services
- Support for non-relational database services (NoSQL, Key-Value stores)
- Continued support for Hadoop
- Support for cloud-like interfaces making it easy to shift between a cloud service and NERSC, possibly for both computation and for HPSS.
- Expert Consulting, architecture design, reference designs.
- Enhanced support for Data Intensive and High-Throughput Workloads (enable more workloads to move from Genepool to the big systems)

9.4.6 Requirements Summary

	Used at NERSC in 2012	**Needed at NERSC in 2017**
Computational Hours (Millions)	32	400
Typical number of cores* used for production runs	1	Goal: 100s (Many-core)
Maximum number of cores* that can be used for production runs	100s	1,000s of threads (GPU)
Projectb directory space	700 TB	10 PB
Archival data	1.3 PB	7.5 PB
Memory per node	Up to 2 TB	>2 TB

* “Conventional cores.” For GPUs and accelerators, please fill out section 4.8. **Including an estimate of “Genepool” usage.

9.5 KBASE Systems Biology Knowledge Base

Principal Investigators: Shane Canon (NERSC)
Case Study Authors: Tom Brettin (ORNL) and Shane Canon (LBNL)
NERSC Repository: kbase

9.5.1 Overview and Context

The goal of KBase is to enable predictive biology. To achieve this, KBase integrates commonly used tools and their associated data, and builds new capabilities on top of the combined data. New functionality allows users to visualize data, create powerful models, and design experiments based on KBase-generated suggestions. KBase is composed of core biological analysis and modeling functions, including an application programming interface that can be used to connect different software programs within the community. KBase is supported by a computing infrastructure that combines traditional clusters at centers like NERSC along with clusters based on the OpenStack cloud system software distributed across the core sites.

NERSC is an active partner in the KBase project. NERSC has provided allocations through NISE (a NERSC discretionary allocation) and has contributed hardware resources to the effort. NERSC also hosts portions of the KBase infrastructure at its Oakland Scientific Facility. Most importantly, NERSC staff members are "matrixed" to the KBase project and are working to develop a special web-based interface, termed the Cluster Service, to enable KBase to easily leverage NERSC resources using the KBase allocation. HPC resources like NERSC will primarily be used to support the most computational intensive analysis required by KBase. This will include on-demand analysis as well as periodic re-analysis. As compute intensive applications are identified, they are being ported to NERSC and integrated into the Cluster Service.

9.5.2 Scientific Objectives for 2017

KBase will maximize understanding of microbial system function, promote sharing of data and findings, and vastly improve the planning of effective experiments. Early efforts will target enabling the reconciliation of metabolic models with experimental data. The ultimate aim is manipulating microbial function for applications in energy production and remediation. In order to accomplish this, we will help users expand on a strong foundation of quality genome annotations, to reconstruct metabolism and regulation, to integrate and standardize 'omics data, and to construct models of genomes.

A high priority within the plant research community is linking genetic variation, phenotypes, molecular profiles, and molecular networks, enabling model-driven phenotype predictions. A second goal will be to map plant variability onto metabolic models to create model-driven predictions of phenotypic traits. Initial work will focus on creating a workflow for rapidly converting sequencing reads into genotypes. We will also build tools for data exploration, and the linking of gene targets from phenotype studies such as genome-wide association studies, with co-expression, protein-protein interaction, and regulatory network models. Such data exploration will allow users to narrow candidate gene lists by refining targets, or be able to visualize a subnetwork of regulatory and physical interactions among genes responsible for a phenotype in question. Users can also highlight networks or pathways impacted by genetic variation.

Through comparative analysis of metagenomes acquired over different spatial, temporal, or experimental scales, it is now possible to define how communities respond to and change their environment. KBase will provide the computational infrastructure to research community behavior and to build predictive models of community roles in the carbon cycle, other biogeochemical cycles, bioremediation, energy production, and the discovery of useful enzymes. We are building the next-generation metagenomic platform that provides scalable, flexible analyses, data vectors for models, tools for model creation, data quality control, application programming interfaces, and GSC-compliant data and standards for data collection. Initial efforts will target the development of bio-prospecting and experimental design tools.

9.5.3 Computational Strategies

9.5.3.1 Approach

The computational approach for KBase is implicit in its goal of integrating diverse data types to derive new insights. A service-oriented architecture is being utilized to expose data and analysis methods and using Cluster Services to expose high-end computational resources like those at NERSC.

9.5.3.2 Codes and Algorithms

Genomics analysis typically draws from a large collection of community-developed applications, libraries, and tools. The most CPU intensive applications are being exposed through cluster services so they can be invoked from the KBase services. We will briefly summarize some of the applications that are being targeted. Many of these tools are not MPI-capable. Therefore, the TaskFarmer developed at NERSC is being used to run these tools in parallel. However, a few applications, such as Kiki, are implemented in MPI.

- BLAST/BLAT – BLAST and BLAT both perform local alignments of sequences against a reference. Both rely on heuristics for this alignment. BLAST provides more accurate alignments but is more computational expensive. Both applications can be run in a multi-threaded manner, but the threading implementation is typically not as efficient as the serial version. There are MPI implementations of BLAST, but currently this is not being used for several reasons. The TaskFarmer is being used instead.
- Kiki – Kiki is an MPI-enabled parallel assembler. Sequencers are read in parallel and indexed. A designated number of nodes are then used to serve the sequence using the index hash as a lookup. Seed sequences are selected and extended by querying the sequence servers. The sequences that are returned are then run through a consensus algorithm. This process is continued until the sequence can no longer be extended. At that point a new seed is selected and the process is repeated. This is done until all of the sequences have been exhausted. Kiki has been ported to Hopper and has been run to thousands of cores.
- Gap Analysis – A key new functionality of KBase will be the ability to automatically reconcile models with experimental data. The Gap Analysis tool identifies missing functions from a model and guides the user in modifying a model to fit the experimental data.

9.5.4 HPC Resources Used Today

9.5.4.1 Computational Hours

KBase continues to ramp up. To date, the usage has been low since this project is still in development mode and NERSC has been used only for development and early testing. In the coming year, we anticipate KBase transitioning to early production. This should lead to increased need for storage and computational resources.

9.5.4.2 Compute Cores

The number of required cores depends both on the specific application and the input data sets. The TaskFarmer, which is used to create parallel instances of many serial applications, is currently capable of scaling to around 32K cores under the right circumstances. Improvements are needed to the TaskFarmer in order to run efficiently at larger scales. The underlying applications are embarrassingly parallel in most cases, so the scaling issues typically originate from load imbalance, I/O contention, or overloading the server that orchestrates the worker in the TaskFarmer.

9.5.4.3 Data and I/O

The amount of I/O is highly variable. The largest metagenomic datasets can be on the order of 100 GB and reference datasets can often be tens of gigabytes. Typically reference datasets must be read in on all nodes and the query data set distributed across the cores. Most applications do not currently provide application-based checkpoint methods. However, the TaskFarmer does maintain a checkpoint at the task level.

9.5.4.4 Project Data

KBase has a project directory (kbase) which is used to store applications, reference datasets, and long-lived outputs. It is currently at the default size, and uses only about 20% of the quota, but this will likely need to grow significantly as KBase transitions into production.

9.5.5 HPC Requirements in 2017

9.5.5.1 Computational Hours Needed

It is difficult to project how much the computational demands may increase for KBase. The project is still several months away from production, so the baseline demand from users is still not understood. However, the potential demand could be very high. The sequencing technology that helps drive the demand continues to advance. In addition to advancements in short read technology, single molecule based technologies are expected to enter the market during this time period. This has the potential to generate a huge spike in demand as sequencing becomes further commoditized. Using JGI as a reference point, it is not unlikely that KBase could require 100M core hours in this time period. KBase will expand its own infrastructure to handle the growth for data storage and web services. The analysis demands could easily exceeds its dedicated capacity. The ability to leverage NERSC (and other ASCR resources) is a key strategy to addressing this gap.

9.5.5.2 Number of Compute Cores

Many of the key applications will likely continue to be throughput oriented. Tools similar to the TaskFarmer will continue to be used to run these applications at scale. However, we expect the number of MPI-enabled genomic applications to increase. For example, a workshop was recently held to investigate the feasibility of developing a Sequencing Analysis Library that could exploit the capabilities of capability class computing systems.

9.5.5.3 Data and I/O

Genomic applications tend to be data intensive. However, the absolute data rates are modest. For example, a similarity tool will at a minimum require reading in the entire query and reference sequences. These are typically tens to hundreds of gigabytes in size. The outputs can approach terabytes over a run that last tens of hours. This leads to rates of only gigabytes per second. However, additional processing is often needed to re-order and sort the output. These can lead to significantly higher I/O rates (many gigabytes per second). Most genomic application do not support application-level checkpointing, so fault tolerance is typically managed at the task level.

9.5.5.4 Project Data

KBase currently has a standard amount of project space. We expect this need to increase at least 30x by 2017. However, based on user demand of KBase, the needs could be larger.

9.5.5.5 Memory Required

Genomic tools and applications have a broad range of memory requirements that are heavily dependent on the input data set. Assemblers typically have the most demanding memory requirements since they need to store a large hash of the sequences. For large metagenomes this can require a terabyte of memory or more. Most similarity tools (i.e. BLAT, BLAST) work best when the entire reference data set can be stored in memory.

9.5.5.6 Many-Core and/or GPU Architectures

While many of the current applications are threaded, the current implementations are typically inefficient. Significant investment is needed to improve these implementations and prepare them to fully exploit emerging many-core architectures. These improvements are not in the current scope of the KBase project, which is focused on building service-oriented architectures, defining data standards for biological data, and developing applications to integrate data from multiple sources to provide new insight.

A limited number of applications have been ported to GPUs (e.g, HMMer, BLAST), but these implementations achieve only modest speed-ups compared to the non-accelerated versions. NVidia is developing a toolkit for bioinformatics that could accelerate development and adoption of GPUs for bioinformatics. We expect continued growth in the number of algorithm implementations that take advantage of GPUs but this project's needs will track more slowly than the growth because only a select few algorithms will be incorporated into the KBase services architecture.

9.5.5.7 Software Applications and Tools

Compilers and tools that can assist in exploiting many-core and accelerators will be valuable. Tools to facilitate integrating data stored on file systems and archival storage could also be of value.

9.5.5.8 HPC Services

KBase is deploying specialized hardware to host KBase services. This is partially driven by some of the unique requirements of the project. A more generalized solution at NERSC that would allow projects to deploy customized services a separate address space could simplify the deployment of these types of services.

9.5.5.9 Time to Solution and Throughput

By 2017, KBase should be deep into production use. A key requirement will be the ability to trigger near real-time analysis based on web-oriented user requests. While some delay can be tolerated depending on the complexity of the analysis, the ability to quickly service request will be critical to providing a good user experience.

9.5.5.10 Data Intensive Needs

KBase is major driver and demonstration of data intensive computing. For data sharing, workflow management, data analysis, and specialized visualization methods, KBase will require addressing all aspects of data intensive computing. Any advanced capabilities that NERSC can provide to help address these needs will be of value. However, these capabilities will need to be provided in a manner that can be easily integrated into the larger KBase project.

9.5.6 Requirements Summary

	Used at NERSC in 2011	**Needed at NERSC in 2017**
Computational Hours (Million)	0.017	100
Typical number of cores* used for production runs	2,400 - 9600	90,000
Maximum number of cores* that can be used for production runs	32 k	200 k
Checkpoint data written per run	N / A	N / A
Checkpoint bandwidth	N / A	N / A
Data read and written per run (excluding checkpoint data)	5 TB	100 TB
Maximum I/O bandwidth (excluding checkpoint data)		100 GB/S
Project directory space	5 TB	150 TB
Archival data	1 TB	2,000 TB
Minimum memory per node	40 GB	512 GB

Aggregate memory	1 TB	10 TB

Appendix A. Attendee Biographies

Application Scientists

Mohammed AlQuraishi is a Systems Biology Fellow at Harvard Medical School. He received his Ph.D. in Genetics from Stanford University under the supervision of Harley McAdams and Lucy Shapiro. His current research interests lie at the intersection of systems and structural biology. He aims to obtain a systems-level understanding of biological processes through a molecular-level understanding of biological structures and their interactions. Towards that end he is developing computational methods for predicting the binding partners and quantitative binding affinities of biological molecules from their atomic structure. His work combines recent advances in machine learning and artificial intelligence with concepts from statistical mechanics and biophysics.

Thomas Bettge is currently a member of NCAR's Climate Change Prediction Program under UCAR's Cooperative Agreement with the DOE/BER. For 35 years he has specialized in high performance computing, weather, climate and ocean modeling, and data analysis applications. As an associate scientist with the Climate Change Prediction Group from 1986-2002, he assembled the DOE Parallel Climate Model (PCM). From 2003-2009 he served as Director of Operations and Services (OSD) within the Computational and Information System Laboratory (CISL) at NCAR, managing the computational, data management, and research needs of the atmospheric and related earth science communities.

William Collins's research is focused on the changes in the energy balance of the Earth system and the implications of those changes for the future of our climate. He trained in physics and astrophysics at Princeton University and the University of Chicago. He is Senior Scientist and Department Head Professor in Residence Professor, University of California, Berkeley and a member of the Earth Sciences Division at Berkeley Lab.

Gilbert ("Gil") Compo is a CIRES research scientist studying atmospheric and oceanic variations ranging from climate change to storm tracks using climate models, observational and reanalysis datasets. He co-leads the development of historical reanalysis ensemble-based techniques and the recovery of the historical observations to extend reanalysis back to the 18th century. With these and other datasets, he studies the role of the oceans in observed climate change and climate variability and the influence of El Niño/Southern Oscillation on these variations. He is particularly focused on developing and improving techniques to assess changes in the risk of extreme and high-impact weather. He holds a Ph.D. in Astrophysical, Planetary, and Atmospheric Sciences from the University of Colorado, Boulder.

Rob Egan is a software developer in the Research and Development department of the DOE Joint Genome Institute. He primarily develops MPI and other cluster-based software targeted at assembling and analyzing terabase scale metagenomic datasets.

David Goodstein is bioinfomatician and software manager at the Joint Genome Institute and Center for Integrative Genomics at the University of California, Berkeley. He holds a Ph.D. in Physics from Cornell University.

David Goodwin is the DOE Program Manager for NERSC, was the NERSC Allocation Manager in High Energy and Nuclear Physics for over 16 years, and holds degrees in physics and engineering.

Ruby Leung is a recognized leader in regional climate modeling. She has been actively involved in the modeling of regional and global climate, developing subgrid cloud parameterizations, and coupling land and atmosphere models. She has applied regional and global climate models and hydrology models to understand the impacts of climate variability and change on water resources in the United States and East Asia. Dr. Leung develops subgrid cloud parameterizations to represent the influence of subgrid-scale terrain variations on orographic precipitation, and turbulence and cloud-radiation interaction effects on stratocumulus clouds. She holds a Ph.D. in Atmospheric Science from Texas A&M University.

Victor Markowitz is Chief Informatics Officer & Associate Director at DOE Joint Genome Institute and head of Lawrence Berkeley National Laboratory's Biological Data Management and Technology Center. He received his M.Sc. and D.Sc. degrees in computer science from Technion - Israel Institute of Technology. Dr. Markowitz has authored articles and book chapters on various aspects of data management and served on review panels and program committees for database and bioinformatics programs and conferences.

Loukas Petridis obtained a Ph. D. in theoretical physics from Cambridge University in 2006. He was postdoctoral fellow at Oak Ridge National Laboratory (ORNL) from 2007 to 2009. Since 2010 he has been a Staff Scientist at ORNL. Petridis has performed research in high-performance computer simulation of biological macromolecules, neutron scattering in bioenergy research and polymer physics.

Stephen Price is a staff member of Los Alamos National Lab’s Climate, Ocean and Sea Ice Modeling (COSIM) group, whose mission is to develop and apply high-performance, multi-scale models of the Earth’s climate for studying the role of ocean and ice systems in high-latitude climate change. He has expertise in glaciology and large-scale numerical modeling of glaciers and ice sheets in the climate system. He is a lead developer of the Community Ice Sheet Model (CISM), a co-PI and Science Team Lead for the new DOE PISCEES project, an acting co-chair for the Community Earth System Model (CESM) Land Ice Working Group, and a member of the U.S. CLIVAR working group on ice sheet / ocean interactions in Greenland.

Tim Scheibe is a hydrogeologist in the Environment and Energy Directorate at Pacific Northwest National Laboratory. He holds a Ph.D. in Civil Engineering from Stanford University. His research is focused on integrating models of fluid flow, material transport, and biogeochemical reactions in the subsurface from the pore scale to field applications. Application areas of focus include contaminant bioremediation and microbial transport, enhanced geothermal systems, geological carbon sequestration, and microbially-mediated carbon cycling in soils.

Jeremy Smith holds the Governor's Chair at University of Tennessee, Department of Biochemistry, Cellular and Molecular Biology and is Director of the UT/ORNL Center for Molecular Biophysics. He holds degrees in biophysics from the University of Leeds and the University of London and a completed a postdoctoral fellowship at Harvard University in Chemistry.

Cristiana Stan is an assistant professor in the Department of Atmospheric, Oceanic and Earth Sciences at George Mason University. She received her Ph.D. in Atmospheric Sciences from Colorado State University. Her research interests center on climate modeling with a focus on the dynamics and predictability of tropical variability. Specific topics include the role of cloud representation in modeling the tropical cyclone activity, monsoon circulations, Madden-Julian Oscillation and the El-Nino Southern Oscillation under current conditions and future climate change scenarios.

Mark Taylor contributes to the development of CAM, the atmosphere model component of the Community Earth System Model (CESM). Mark is a co-chair of the CESM Atmospheric Model Working Group, as well as co-editor of the Springer book 'Numerical Techniques for Global Atmospheric Models'. Mark Taylor received his Ph.D. from New York University's Courant Institute of Mathematical Sciences in 1992. From 1992 to 1998, Mark was a post-doc and then a software engineer at the National Center for Atmospheric Research (NCAR). From 1998 to 2004 he was a staff member at Los Alamos National Laboratory before joining Sandia National Laboratories in 2004.

Jin-Ho Yoon conducts climate physics research at Pacific Northwest National Lab. His areas of interest include Climate modeling with global and regional climate models, diagnostic analysis of model output and observational data, climate variability and change, seasonal climate prediction, climate change impact on human and hydro-ecosystems, and climate sensitivity and feedback processes. He has a Ph.D. in Meteorology from Iowa State University.

Editors and NERSC Application Support Personnel

Shane Canon leads the NERSC Technology Integration Group (TIG). He joined NERSC in 2000 to serve as a system administrator for the PDSF cluster. He left LBNL to take a position as Group Leader at Oak Ridge National Laboratory returned to NERSC to lead the Data Systems Group in 2008. In 2009, he transitioned to leading the newly created TIG in order to focus on the Magellan Project and other strategic areas. Shane has a Ph.D in Physics from Duke University and B.S. in Physics from Auburn University.

Richard Gerber is NERSC Senior Science Advisor and User Services Deputy Group Lead and, with Harvey Wasserman, 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. Workshop Agenda

Thursday, May 26		
Time	Topic	Presenter
8:00am	Arrive, informal discussions	
8:30	Welcome, introductions, workshop goals, Workshop outline, logistics, format, procedures	Yukiko Sekine, Harvey Wasserman, Richard Gerber
8:45	BER Program Office Research Directions	Susan Gregurick,
9:15	NERSC Role in BER Research	Kathy Yelick
10:00	Break	
10:15	Case Study: Community Earth System Model (CESM)	Tom Bettge
10:45	Case Study: CLIMES and IMPACTS	William Collins
11:15	Case Study: PISCEES Ice Sheet Modeling	Stephen Price
11:45	Case Study: Development of Frameworks for Robust Regional Climate Modeling	Ruby Leung
12:15	Working Lunch	
12:45	Case Study: Characterization of Clouds Aerosols and the Cryosphere	Jin-Ho Yoon
1:15	Case Study: Anthropogenic Climate Change Using Super-Parameterization	Cristiana Stan
1:45	Case Study: Climate Science for a Sustainable Energy Future (CSSEF)	David Bader
2:15	Case Study: Sparse Input Reanalysis	Gilbert Compo
2:45	Break	
3:00	Subsurface Flow and Reactive Transport	Tim Scheibe

3:15	Case Study: Joint Genome Institute	David Goodstein, Victor Markowitz
4:00	Open Discussions	
5:00	Adjourn for the day	

Friday, May 27		
8:00am	Arrive, informal discussions	
8:30	Introduction to the New NERSC Director	Sudip Dosanjh
9:00	Case Study: Molecular Dynamics Simulations of Protein Dynamics and Lignocellulosic Biomass	Loukas Petridis
9:30	Case Study: Computational Predictions of Transcription Factor Binding Sites	Mohammed AlQuraishi
10:00	Case Study: DOE Systems Biology Knowledgebase	Tom Brettin, Shane Canon
10:15	Break	
10:45	Review, Report schedule and process	Harvey Wasserman
11:00	Q&A, general discussions, consensus findings	Richard Gerber
12:00	Working lunch	
1:00pm	Adjourn	

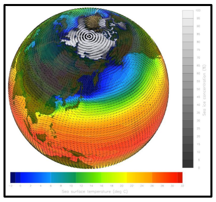
Appendix C. Abbreviations and Acronyms

20CR	20th Century Reanalysis
ALCC	ASCR Leadership Computing Challenge
ALCF	Argonne Leadership Computing Facility
AMIP	Atmospheric Model Intercomparison Project
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
CAM	Community Atmosphere Model
CAM-SE	CAM-Spectral Element
CCP	Climate Change Prediction Group (at NCAR)
CCSM	Community Climate System Model
CESM	Coupled Earth System Model
CG	Conjugate Gradient
CICE	Community Ice Code
CISL	Computational Information Systems Laboratory (at NCAR)
CLIVAR	World Climate Research Programme (WCRP) project that addresses Climate Variability and Predictability
CMIP	Coupled Model Intercomparison Project
CRM	Cloud Resolving Model
CSL	Climate System Laboratory (at NCAR)
CSSEF	Climate Science for a Sustainable Energy Future
CUDA	Compute Unified Device Architecture
EMSL	Environmental Molecular Sciences Laboratory at PNNL
ESG	Earth System Grid
ESnet	DOE's Energy Sciences Network
FEM	Finite Element Modeling
FFT	Fast Fourier Transform
GA	Global Arrays
GPGPU	General Purpose Graphical Processing Unit
GPU	Graphical Processing Unit
GHG	Greenhouse Gas
HDF	Hierarchical Data Format
HOMME	High-Order Methods Modeling Environment
HPC	High-Performance Computing
HPSS	High Performance Storage System
IPCC	Intergovernmental Panel on Climate Change
I/O	input output
IDL	Interactive Data Language visualization software
INCITE	Innovative and Novel Computational Impact on Theory and Experiment
JGI	Joint Genome Initiative
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LLNL	Lawrence Livermore National Laboratory

MD	Molecular Dynamics
MPAS	Model for Prediction Across Scales
MPI	Message Passing Interface
NCAR	National Center for Atmospheric Research
NERSC	National Energy Research Scientific Computing Center
NetCDF	Network Common Data Format
NGF	NERSC Global Filesystem
NISE	NERSC Initiative for Science Exploration
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
OS	operating system
PCMDI	Program for Climate Model Diagnosis and Intercomparison
PDE	Partial Differential Equation
PDSF	NERSC’s Parallel Distributed Systems Facility
PISCEES	Projecting Ice Sheet and Climate Evolution at Extreme Scales
PNNL	Pacific Northwest National Laboratory
POP	Parallel Ocean Program
SC	DOE's Office of Science
SciDAC	Scientific Discovery through Advanced Computing
SLAC	SLAC National Accelerator Laboratory
SODA	Simple Ocean Data Assimilation
SP-CCSM	Super-parameterized CCSM
SPH	Smoothed Particle Hydrodynamics
UQ	Uncertainty Quantification
WRF	Weather Research and Forecasting Model

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.



Earth's climate system (DI02467). Image courtesy Gary Strand.This image depicts a single month from a simulation of the 20th century by the NCAR-based Community Climate System Model, [version 4]. The CCSM and its successor, the Community Earth System Model (CESM), represent two of the world's most powerful computer models for simulating the complex interactions of Earth's climate system, including the atmosphere, oceans, sea ice, and land surface. This image captures wind directions, ocean surface temperatures, and sea ice concentrations. Image copyright University Corporation for Atmospheric Research.

Montage depicting research activities within the DOE Office of Biological and Environmental Research (http://science.energy.gov/ber) and BER's Genomic Science program (http://genomicscience.energy.gov). The original montage was created by Betty Mansfield, Group Leader of the Biological and Environmental Research Information System at Oak Ridge National Laboratory. Image credits, from top, left: Plant cell wall image courtesy of Advanced Cell Wall Characterization Team, National Renewable Energy Laboratory and DOE BioEnergy Science Center. Green leaf and DNA researcher images courtesy of DOE Joint Genome Institute. 3D visualization of pore-scale fluid flow computed using the parallel Smoothed Particle Hydrodynamics code developed in the Computational Hybrid Integration of Physical Processes across Scales (CHIPPS) project, Tim Scheibe, PNNL. Atmospheric instruments image courtesy of U.S. Department of Energy's Atmospheric Radiation Measurement Climate Research Facility. Globe image courtesy of Gary Strand, National Center for Atmospheric Research. Aerial landscape image courtesy of David F. Karnosky, Michigan Technological University. Leaf autoradiograph image courtesy of Richard Ferrieri, Brookhaven National Laboratory. Subsurface sampling image courtesy of Oak Ridge National Laboratory. Molecular image courtesy of EMSL facility, Pacific Northwest National Laboratory. Protein image courtesy of Lawrence Livermore National Laboratory.