

Advanced Scientific Computing Research Network Requirements

ASCR Network Requirements Review Final Report

Conducted October 4–5, 2012



Lawrence Berkeley
National Laboratory



Office of Science

DISCLAIMER

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

Advanced Scientific Computing Research Network Requirements

Office of Advanced Scientific Computing Research, DOE Office of Science
Energy Sciences Network
Germantown, Maryland — October 4-5, 2012

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR). Vince Dattoria is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract DE-AC02-05CH11231.

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

This is LBNL report LBNL-XXXX

Participants and Contributors

Charles Bacon, ANL (ALCF computing)

Greg Bell, ESnet (Networking)

Shane Canon, LBNL (NERSC computing)

Eli Dart, ESnet (Networking)

Vince Dattoria, DOE/SC/ASCR (ESnet Program Manager)

Dave Goodwin, DOE/SC/ASCR (ASCR Program)

Jason Lee, LBNL (NERSC networking)

Susan Hicks, ORNL (ORNL networking)

Ed Holohan, ANL (ALCF networking)

Scott Klasky, ORNL (OLCF computing)

Carolyn Lauzon, DOE/SC/ASCR (ASCR Program)

Jim Rogers, ORNL (OLCF)

Galen Shipman, ORNL (SNS)

David Skinner, LBNL (NERSC computing)

Brian Tierney, ESnet (Networking)

Editors

Eli Dart, ESnet — dart@es.net

Brian Tierney, ESnet — bltierney@es.net

Table of Contents

1	Executive Summary.....	6
2	Findings	7
3	Action Items	10
4	Review Background and Structure	11
5	Office of Advanced Scientific Computing Research	12
6	Argonne Leadership Computing Facility (ALCF)	17
7	National Energy Research Scientific Computing Center (NERSC)	26
8	Oak Ridge Leadership Computing Facility (OLCF)	39
9	Appendix A – The ESnet OSCARS Service.....	55
10	Appendix B – The ESnet 100G Testbed	59
11	Glossary.....	60
12	Acknowledgements.....	63

1 Executive Summary

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the U.S. Department of Energy (DOE) Office of Science (SC), the single largest supporter of basic research in the physical sciences in the United States. In support of SC programs, ESnet regularly updates and refreshes its understanding of the networking requirements of the instruments, facilities, scientists, and science programs that it serves. This focus has helped ESnet to be a highly successful enabler of scientific discovery for over 25 years.

In October 2012, ESnet and the Office of Advanced Scientific Computing Research (ASCR) of the DOE SC organized a review to characterize the networking requirements of the programs funded by the ASCR program office.

The requirements identified at the review are summarized in the *Findings* section, and are described in more detail in the body of the report.

2 Findings

2.1 General Findings

The Argonne Leadership Computing Facility (ALCF), the National Energy Research Scientific Computing Center (NERSC), and the Oak Ridge Leadership Computing Facility (OLCF) are significantly expanding the number and capabilities of data transfer node (DTN) systems. All three centers expressed a desire to collaborate on the design and configuration of the next generation of DTNs.

This INCITE (Innovative and Novel Computational Impact on Theory and Experiment) year saw a marked increase in the number of INCITE proposals that requested time at both the ALCF and the OLCF. At the review, the discussion indicated that some codes run better on one architecture or another (the Cray systems at the OLCF are quite different from the IBM BlueGene systems at the ALCF); because of this, several projects want to run at both Leadership Computing Facilities (LCFs). Tight coupling that would require co-scheduling of the two leadership-class systems is not yet envisioned, but the output of codes run at one LCF would serve as inputs to codes run at the other LCF, requiring the transfer of large data sets between LCFs.

There are often large data transfers associated with the startup of an INCITE project at an LCF. Sometimes the transfer is from another DOE facility, but this is not always the case.

As data scale increases, the issue of data integrity becomes more important. Larger data sets are statistically more likely to experience errors in storage or transmission. Data integrity was identified as a concern, and data integrity technologies should be considered for inclusion in workflow tools, file systems, and other capabilities in the data-intensive era.

In the future, advanced petascale or exascale machines will be expected to be able to conduct realistic simulations of internal combustion engines. The data sets produced would be of interest to a wide range of users, both within DOE and elsewhere. The transfer of that data may be a challenge in some cases (one simulation is expected to generate multiple exabytes of results data).

Experimental facilities (e.g., the Advanced Light Source [ALS] at Lawrence Berkeley National Laboratory [LBNL] and the Spallation Neutron Source [SNS] at Oak Ridge National Laboratory [ORNL]) have been working with computational centers to stream experimental data to computational resources in real time or near-real time. This streaming allows beamline scientists to receive analysis results while conducting the experiment, and adjust or improve the experiment in response to the analysis. The coupling of experimental facilities and computational facilities is likely to expand in the future, resulting in significant additional network demand because of the very high data rates (1 GB/sec within two years) produced by next-generation beamline instruments. Network enabled real-time coupling between experimental and computational facilities also creates a paradigm shift in the scientific view of networks and compute resources;

rather than disjoint tools for post-experimental analysis, they become an extension of the experimental apparatus itself.

The setup and use of On-Demand Secure Circuits and Advance Reservation System (OSCARS) virtual circuits is not well understood by the ASCR computational facilities. Multiple participants expressed a desire for ESnet to work with the ASCR computational facilities to increase the deployment and usage of virtual circuits. In addition, a desire was expressed for ESnet to provide guidelines to assist scientists and facilities in determining when virtual circuits might be useful to particular projects. A description of OSCARS is included in Appendix A of this report (see section 9).

It is a common practice for many scientists at LCFs to download only the results of their work to their home institution. Increasingly, there is a need to download intermediate results or reduced data sets for further analysis on local resources. This will increase the number of locations that require good data-transfer performance to/from the ASCR HPC centers.

The need for real-time or near-real-time streaming analysis of data, either from a running simulation or a running experiment, is increasing. This is in contrast to a file-based methodology, which involves transferring the data files and then running the analysis after the data files have been transferred. In addition, remote input/output (I/O) techniques are being developed to provide alternative means of data access unencumbered by traditional file-based semantics. One reason for this is that when using file-based data-access methods, very large data sets must be transferred from one file system to another, which means that large-scale data sets take up space on multiple file systems. Remote I/O technologies are expected to place additional demands on the network.

Real-time experiment monitoring mechanisms used by running experiments can benefit from feedback from networks so that the quality of information returned from the experiments can adapt to the network. For example, double precision values could be reduced to single precision when time-to-solution becomes more important than the precision of the real-time feedback data.

Requests for co-scheduling of computing, storage, networking, and experimental resources are on the rise. The main driver is experimental workflows at user facilities that have a high-performance computing (HPC) component (typically analysis or visualization).

2.2 Findings for Specific Facilities

Experimental data from the Korea Superconducting Tokamak Advanced Research (KSTAR) fusion facility in South Korea could be analyzed far more effectively at the OLCF if network performance between the two facilities were improved. These improvements can likely be realized through collaboration among ESnet, KSTAR, the OLCF, and the Korean science networks.

A need was expressed by NERSC for additional bandwidth between NERSC and LBNL to support experimental workflows at the ALS — this need is generalizable to other experimental facilities.

In FY2015, NERSC will require between 2 and 4 100Gbps circuits to support the relocation of the NERSC facility to a new building. These circuits will facilitate the production operation of computational resources in both buildings during the transition, and will be needed for a period of 12 to 18 months.

There are some risks to the diversity of network connectivity to ORNL in the metro area and regional contexts.

The LaserPlasmaSim group is interested in exploring a DTN and the Science DMZ model.

3 Action Items

Several action items for ESnet came out of this review. These include:

- ESnet will collaborate with the HPC centers on the design, configuration, and tuning parameters for the next generation of DTNs.
- ESnet will work with KSTAR and the OLCF to improve data-transfer performance in support of the Fusion collaborations that analyze KSTAR experimental data at the OLCF, and its partners at LBNL and Princeton Plasma Physics Laboratory (PPPL).
- ESnet and ORNL will discuss the diversity of network connectivity to ORNL and the strategic solutions available.
- ESnet will work with the LCFs to assess the needs of INCITE projects that run at both LCFs and need to transfer data between the two. ESnet will help with performance tuning for other INCITE program data-movement needs (e.g., at INCITE project start-up).
- ESnet will work with the ASCR computational facilities on guidelines for virtual circuit use, and on the deployment and integration of virtual circuit capabilities at the computational facilities.
- ESnet will continue to work with experimental and computational facilities to support coupled experiments using Basic Energy Sciences (BES) facilities and ASCR computing centers.
- ESnet will continue to develop and update the fasterdata.es.net site as a resource for the community.
- ESnet will continue to assist sites with perfSONAR (PERformance Service Oriented Network monitoring ARchitecture) deployments and with network and system performance tuning.

In addition, ESnet will continue development and deployment of the ESnet OSCARS to support virtual circuit services.

4 Review Background and Structure

The strategic approach of ASCR and ESnet for defining and accomplishing ESnet's mission covers three areas:

1. Working with the DOE SC-funded science community to identify the networking implications of instruments and HPC resources, and the evolving process of how science is done
2. Developing an approach to building a network environment that will enable the distributed aspects of SC science and continuously reassess and update the approach as new requirements become clear
3. Continuing to anticipate future network capabilities to meet future science requirements with an active program of R&D and advanced development

For point (1), the requirements of the SC science programs are determined by:

- a. A review of the plans and processes of the major stakeholders, including the data characteristics of scientific instruments and facilities, to investigate what data will be generated by instruments and HPC resources coming online over the next 5-10 years. In addition, the future process of science must be examined: How and where will the new data be analyzed and used? How will the process of doing science change over the next 5-10 years?
- b. Observing current and historical network traffic patterns to determine how trends in network patterns predict future network needs

The primary mechanism to accomplish (a) is through SC Network Requirements Reviews, which are organized by ASCR in collaboration with the SC Program Offices. SC conducts two requirements reviews per year, in a cycle that assesses requirements for each of the six program offices every three years. The review reports are published at <http://www.es.net/requirements/>.

The other role of the requirements reviews is to ensure that ESnet and ASCR have a common understanding of the issues that face ESnet and the solutions that ESnet undertakes.

In October 2012, ESnet and the ASCR organized a review to characterize the networking requirements of ASCR-funded science programs, with an emphasis on high performance computing facilities. Participants were asked to codify their requirements in a case-study format that included a network-centric narrative describing the science; instruments and facilities currently used or anticipated for future programs; the network services needed; and how the network is used. Participants considered three timescales in their case studies — the near term (immediately and up to two years in the future), the medium term (two to five years in the future), and the long term (greater than five years in the future). The information in each narrative was distilled into a summary table, with rows for each timescale and columns for network bandwidth and services requirements. The case-study documents are included in this report.

5 Office of Advanced Scientific Computing Research

5.1 ASCR Overview

5.1.1 Mission

The Advanced Scientific Computing Research (ASCR) program's mission is to advance applied mathematics and computer science; deliver, in partnership with disciplinary science, the most advanced computational scientific applications; advance computing and networking capabilities; and develop, in partnership with U.S. industry, future generations of computing hardware and tools for science. In this way, ASCR supports the science goal of the Department of Energy (DOE) 2012 Strategic Plan to maintain a vibrant U.S. effort in science and engineering as a cornerstone of our economic prosperity and underpins the targeted outcome to "develop and deploy high-performance computing hardware and software systems through exascale platforms."

5.1.2 Background

Over the past decade, in both theory and experiment, computing has become a ubiquitous tool for science and engineering that allows researchers to delve deeper, think bigger, and explore regimes previously out of reach. ASCR and its predecessor organizations, in partnership with the National Nuclear Security Administration's Advanced Simulation and Computing (ASC) program, has led this computing revolution for the past decade, building on a foundation of over 50 years of research and collaboration. This partnership has delivered the scientific promise of high performance computers for national security, science, and engineering and has driven the world leadership of U.S. vendors in high performance computing.

Together, ASCR and ASC led the transition to parallel computing with interconnected commercial processors in the 1990s. In 2009, ASCR delivered the first petascale systems for open science that enabled new insights into: diseases such as Parkinson's and Alzheimer's, disasters such as earthquakes and hurricanes, and allowed industry to improve the energy efficiency of aircraft and long-haul trucks. In addition, ASCR supported software, such as the Message Passing Interface (MPI) built into all massively parallel software, has enabled the worldwide parallel computing industry—from dual core laptops to supercomputers. ASCR developed software, protocols, advanced storage technologies, data tools and math libraries are also used throughout industry and academia. ASCR's Scientific Discovery through Advanced Computing (SciDAC) program improved the performance of DOE applications up to 10,000 percent and have enabled dozens of applications to run at the petascale enabling new insights into: improving the efficiency of combustion engines, understanding the physical mechanisms of stress-corrosion cracking, reducing uncertainties in global climate models such as the transport of ice sheets, predicting the behavior of fusion plasmas, explaining the progression of supernovae, predicting structure and decay of novel isotopes, and calculating the subatomic interactions that determine nuclear structure.

5.1.2 Current Challenges

Growth in the use of computing, the demand for both capability and capacity computing and the impact on science and engineering continues to challenge and inspire the ASCR program. The potential from broad adoption of more advanced computing for our society, our economy, and the Department's missions is tremendous and has been well documented through numerous reports from DOE, DARPA, the National Academies, the Council on Competitiveness, and other workshops, studies and reports. ASCR must deliver this in the context of rapidly changing hardware and rapidly growing demands from data-intensive science. Together, these challenges make the transition to the next generation of high performance computing fundamentally different than the transition to parallel computing because the power required to move data between processors across an HPC system interconnect now dwarfs the power necessary for calculations and because the increases in parallelism are now multilayered—both on the chips and between them.

ASCR strategy to address these challenges, like the strategy that has underpinned the Department's leadership during the past half century, has three thrusts: world-class computing and network facilities for science; research in applied mathematics, computer science and advanced networking to define and enable the future; and partnerships to bring the first two thrusts together to transform science.

5.2 ASCR Facilities Overview

The High Performance Computing and Network Facilities subprogram delivers forefront computational and networking capabilities to scientists nationwide. These include high performance production computing at the National Energy Research Scientific Computing Center (NERSC) facility at LBNL and Leadership Computing Facilities (LCFs) at Oak Ridge and Argonne National Laboratories. These computers, and the other SC research facilities, generate many petabytes of data each year. Moving data to the researchers who need them requires advanced scientific networks and related technologies provided through High Performance Network Facilities and Testbeds, which includes the Energy Science network (ESnet). The Research and Evaluation Prototypes activity invests in long-term needs that will play a critical role in achieving exascale computing.

Computing resources are allocated through competitive processes. Up to 60% of the processor time on the LCFs is allocated through the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program, which is open to all researchers and results in awards to 20–30 large projects per year. The high performance production computing facilities at NERSC are predominately allocated to researchers supported by SC programs. Remaining processor time on the LCFs and NERSC is allocated through the ASCR Leadership Computing Challenge (ALCC). ALCC is open year-round to scientists from the research community in the national labs, academia, and industry for special situations of interest to DOE with an emphasis on high-risk, high-payoff simulations in areas directly related to the DOE's energy mission, for national

emergencies, or for broadening the community of researchers capable of using leadership computing resources.

Allocations on ASCR facilities provide critical resources for the scientific community following the peer reviewed, public access model used by other SC scientific user facilities. In addition, ASCR facilities provide a crucial testbed for U.S. industry to deploy the most advanced hardware and have it tested by the leading scientists across the country in universities, national laboratories, and industry.

5.2.1 High Performance Production Computing

NERSC

The National Energy Research Scientific Computing Center (NERSC) facility, located at LBNL, delivers high-end production computing services for the SC research community. Annually, approximately 5,200 computational scientists in about 500 projects use NERSC to perform basic scientific research across a wide range of disciplines including astrophysics, chemistry, climate modeling, materials, high energy and nuclear physics, and biology. NERSC enables teams to perform modeling, simulation, and data analysis on some of the most capable computational and storage systems in the world to address some of the biggest scientific challenges within the SC mission. NERSC users come from nearly every state in the U.S., with about 65% based in universities, 25% in DOE laboratories, and 10% in other government laboratories and industry. NERSC's large and diverse user base requires an agile support staff to aid users entering the high performance computing arena for the first time as well as those preparing codes to run on the largest machines available at NERSC and other SC computing facilities.

NERSC is a vital resource for the SC research community and it is consistently oversubscribed, with requests exceeding capacity by a factor of 3–10. This gap between demand and capability exists despite upgrades to the primary computing systems approximately every 3 years. NERSC regularly gathers requirements from SC programs through a robust process that informs NERSC upgrade plans. These requirements activities are also vital to planning for SciDAC and other ASCR efforts to prioritize research directions and inform the community of new computing trends, especially as the computing industry moves toward heterogeneous, multi-core computing.

5.2.2 Leadership Computing Facilities

The Leadership Computing Facilities (LCFs) enable open scientific applications, including industry applications, to harness the potential of leadership computing to advance science and engineering. The era of petaflop science opened significant opportunities to dramatically advance research as simulations more realistically capture complex behavior in natural and engineered systems. The success of this effort is built on the gains made in Research and Evaluation Prototypes and ASCR research efforts. LCF staff operates and maintains forefront computing resources. One LCF strength is the staff support provided to INCITE projects, ASCR Leadership Computing Challenge projects,

scaling tests, early science applications, and tool and library developers. Support staff experience is critical to the success of industry partnerships to address the challenges of next-generation computing.

ALCF

The Argonne Leadership Computing Facility (ALCF) provides a 10 petaflop IBM Blue Gene/Q with relatively low-electrical power requirements. The Blue Gene/Q was developed through a joint research project with support from NNSA, IBM, and ASCR's Research and Evaluation Prototypes activity. The ALCF and OLCF systems are architecturally distinct and this diversity of resources benefits the Nation's HPC user community. ALCF supports many applications, including molecular dynamics and materials, for which it is better suited than OLCF or NERSC. Through INCITE, ALCF also transfers its expertise to industry, including working with Proctor and Gamble to study the complex interactions of billions of atoms to determine how tiny submicroscopic structures impact the characteristics of the ingredients in soaps, detergents, lotions, and shampoos, as well as in fire retardants and foams used in national security applications.

OLCF

The Oak Ridge Leadership Computing Facility (OLCF) 20 petaflop Cray hybrid system is one of the most powerful computers in the world for scientific research. Through INCITE allocations, several applications, including combustion studies in diesel jet flame stabilization, simulations of neutron transport in fast reactor cores, and groundwater flow in porous media, are running at the multi-petaflop scale. OLCF staff is sharing its expertise with industry to broaden the benefits for the Nation. For example, OLCF worked with Boeing to significantly reduce the need for costly physical prototyping and wind tunnel testing to advance the development curve of the Ramgen CO2 compressor with their next generation rotor, and with BMI trucking to increase fuel efficiency in long-haul trucks.

Research and Evaluation Prototypes

The Research and Evaluation Prototypes activity addresses the challenges of next generation computing systems. These activities are coupled to the co-design centers to strengthen feedback loops in the portfolio. By actively partnering with the research community, including industry, on the development of technologies that enables next-generation machines, ASCR can ensure that the commercially available architectures serve the needs of the scientific community. Coupling this activity to the co-design centers ensures that application and software researchers can gain a better understanding of future systems to get a head start in developing software and models to take advantage of the new capabilities. Research and Evaluation Prototypes prepares researchers to effectively utilize the next generation of scientific computers and seeks to reduce risk for future major procurements.

DOE has been at the forefront of leadership computing for science and national security applications for decades. ASCR continues to invest in leadership class systems at Argonne and Oak Ridge, which play a key role in the health of the U.S. high performance computing industry. However, the next generation of computing hardware is expected to present new challenges for science and engineering applications—most notably from power demands that will restrict memory usage, effectively managing communication between billions of chips and accelerators, and from chip failures and silent errors. This activity supports research and development partnerships with vendors to influence and accelerate critical technologies for exascale, system integration research, development and engineering efforts that are coupled to application development to ensure Department applications are ready to make effective use of commercial offerings.

5.2.3 High Performance Network Facilities and Testbeds

ESnet

The Energy Sciences Network (ESnet) provides the national network and networking infrastructure connecting DOE science facilities and SC laboratories with other institutions connected to peer academic or commercial networks. This network allows scientific users to effectively and efficiently access, distribute, and analyze the massive amounts of data produced by these science facilities.

The costs for ESnet are dominated by operations, including refreshing switches and routers on the schedule needed to ensure the 99.999% reliability required for large-scale scientific data transmission. Additional funds are used to support the testing and evaluation of new technologies and services that will be required to keep up with the data volume of new DOE facilities and unique DOE scientific instruments.

6 Argonne Leadership Computing Facility (ALCF)

6.1 Background

Argonne National Laboratory, located just outside Chicago, is one of DOE's largest national laboratories for scientific and engineering research. UChicago Argonne, LLC, manages Argonne for DOE SC. The Laboratory's mission is to apply a unique mix of world-class science, engineering, and user facilities to deliver innovative research and technologies. Argonne's programmatic activities cover all aspects of the innovation ecology: basic research, technology development, and prototype development and testing.

The Argonne Leadership Computing Facility (ALCF) provides the computational science community with a world-class computing capability dedicated to breakthrough science and engineering. It began operation in 2006, with its team providing expertise and assistance to support user projects to achieve top performance of applications and to maximize benefits from the use of ALCF resources.

Awardees of time on the ALCF systems range from national laboratories and universities to corporations and international collaborators. As such, data to be processed must be transferred into the facility from its original home, and results of simulation and analysis are sometimes transferred back to home institutions, driving a range of networking requirements for the facility.

6.2 Key Science Drivers

6.2.1 Instruments and Facilities

ALCF1 — Blue Gene/P

ALCF1, which entered full production on February 2, 2009, supports five user resources. In current production are 42 racks of Blue Gene/P divided into three machines, as well as Eureka, a 100-node visualization cluster, and Gadzooks, a smaller four-node visualization cluster. The three Blue Gene/P systems — Intrepid (40 racks), Surveyor (one rack), and Challenger (one rack) — together support 31 INCITE awards.

The ALCF1 systems are interconnected to the storage infrastructure by a Clos switching network built with Myricom Myri-10G fabric technology. Using both the native Myricom MX protocol as well as 10 Gbps Ethernet, end-to-end connectivity is non-blocking, with full bisectional bandwidth between all compute and storage resources.

Online storage is provided by 10 PB of raw disk capacity served by a combination of the Data Direct Networks (DDN) 9550 and 9900 SAN architectures. A total of 128 file servers interconnect these InfiniBand disk tiers to the MX network and provide both the open-source Parallel Virtual File System (PVFS), and the IBM commercial General Parallel File

System (GPFS). The aggregate theoretical bandwidth available from the DDN 9900 SAN is 88 GB/sec.

Nearline storage is provided by a pair of Spectralogic T950 tape libraries. Space is available for 12,500 LTO4 tapes. With 48 tape drives online, we have an aggregate theoretical bandwidth of 5.8 GB/sec to tape. Two backup tools are used for storage and retrieval. The open-source Amanda tool is responsible for backups of critical host systems, and the DOE/IBM collaboration High Performance Storage System (HPSS) is used to manage user data stored to tape.

ALCF2 — Blue Gene/Q

ALCF2 is currently under construction and scheduled to enter full production status in October 2013. ALCF2 will bring an additional four user resources online. The new infrastructure will connect 50 racks of Blue Gene/Q divided into three machines, as well as a new visualization cluster, Tukey, with 96 nodes and 192 Nvidia Tesla GPUs.

For testing and debugging, two smaller Blue Gene/Q systems are available. Cetus, a single rack of 1024 compute nodes, is connected to the production storage fabric and allows debug runs to be performed against production file systems, speeding time to resolution for users troubleshooting a failure at scale. Vesta, a 2048 node two rack system, lives in a separate, isolated fabric. This allows experimental configurations to be tested that may involve unstable codes not fit for use on the production resource. Cetus has eight input/output nodes (IONs) connected with QDR (quad data rate — a four-lane, 10 Gbps signal rate, capable of 32 Gbps actual data rate) InfiniBand, giving a ratio of 1024:8, or 128:1 compute nodes per IO node. Vesta has been equipped with a total of 64 IONs, yielding a lower ratio of only 32:1, significantly increasing the capability of data IO experimentation on this system.

With 48 racks of Blue Gene/Q compute nodes, Mira is now the fourth fastest supercomputer in the world, according to the November 2012 Top500 List. For MPI communication, the system has a proprietary 5-D torus interconnect. As this is only used for internal system communication, it is not relevant to the discussion of LAN and WAN requirements. To reach external resources, such as the storage subsystem, Mira is equipped with 384 IONs connected to a QDR InfiniBand fabric. With all IONs operating at full capacity, Mira has an aggregate theoretical bandwidth of 1.5 TB/sec, providing plenty of room to grow as storage requirements increase over the next five years. Connectivity is provided by a fully connected network of four Mellanox IS5600 QDR IB switches. Each switch provides 324 ports of edge connectivity for hosts such as IONs and file servers, and 324 ports of core connectivity between switches. Total aggregate capacity of the fabric is over 5 TB/sec.

DDN again provides the storage infrastructure, using the new SFA12K-20E platform. In this platform, virtual machines running onboard the disk controller couplets act as file servers for the infrastructure. The rated performance of a single couplet is 20 GB/sec. With 16 controller couplets, the aggregate theoretical bandwidth from Mira to disk is

320 GB/sec. As of the time of this writing, tests have achieved performance in excess of 230 GB/sec.

A new cluster, Tukey, will accomplish visualization for Mira. With 96 nodes, each equipped with two Nvidia Tesla GPUs, Tukey will be equipped with 98,304 CUDA cores achieving nearly 128 teraflops of double-precision performance. Tukey is uplinked with 1 Gbps of public connectivity to every visualization node. We have designed this cluster to be heavily used for real-time data analysis and video streaming to remote sites, and we fully expect to consume up to 100 Gbps with live video data at peak utilization.

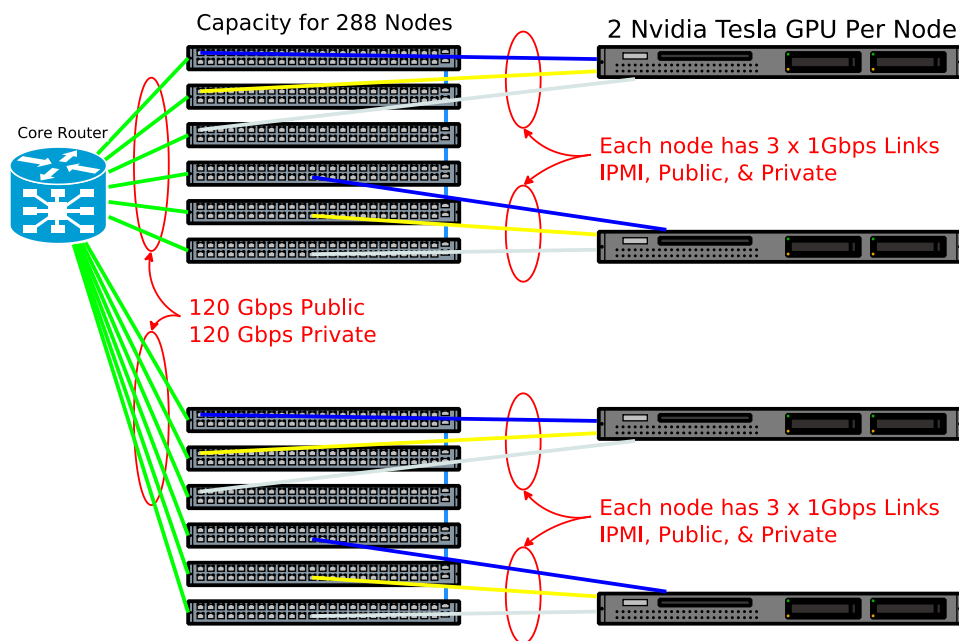


Figure 1.

Data Transfer Nodes

ALCF1 and ALCF2 each have a distinct set of GridFTP DTNs for their respective file systems. All data sets moved in or out of an ALCF resource, other than live visualization data, will be transferred using these dedicated DTNs.

In ALCF1, three DTNs provide the capability of up to 30 Gbps of continuous throughput. The isolated test and development system, Surveyor, has a dedicated DTN with 10 Gbps of public connectivity. Intrepid has a pair of DTNs allowing striped transfers at 20 Gbps.

In ALCF2, a much larger GridFTP cluster is being prepared for the transfer of data sets. At this time, 12 DTNs are available, each with 10 Gbps of public connectivity. We have planned this cluster to fully utilize the newly available 100 Gbps ESnet connectivity at

ANL. In addition, this cluster will be used for data migration between ALCF1 and ALCF2 resources as more INCITE projects grow to use the faster Mira system.

ALCF2 brings a new element of connectivity — interactive login to Blue Gene/Q I/O nodes. These login I/O nodes, or LIONS, are each equipped with dual-port Mellanox adapters. In addition to the QDR IB connection, each LION also provides a public 10 G Ethernet connection direct to the Blue Gene/Q hardware. This new dimension will allow for live data streams that cannot be encapsulated within a GridFTP block transfer, and may provide interesting future capabilities for direct supercomputer-to-supercomputer transfers that do not require traditional bulk data transfer to disk. Mira has 16 of these nodes, while Cetus and Vesta each have eight, for a combined theoretical 320 Gbps of public direct-to-compute bandwidth.

To facilitate data migration, the ALCF is currently implementing a dense wavelength division multiplexing (DWDM) infrastructure between the two HPC facilities at Argonne. At this time, four 10 G Ethernet waves and twelve 8 G Fibre Channel waves are being multiplexed between facilities. Ethernet connectivity will be made available for both administrative and user data, presenting our new GridFTP DTNs to the existing infrastructure over this dedicated link. Fibre Channel is being used to bridge the tape backup infrastructure — our existing T950 libraries are being expanded to support backups from the Mira support infrastructure and by bridging Fibre Channel fabrics, we are able to provide disaster-recovery diversity in our tape backup locations.

ALCF DWDM Connectivity

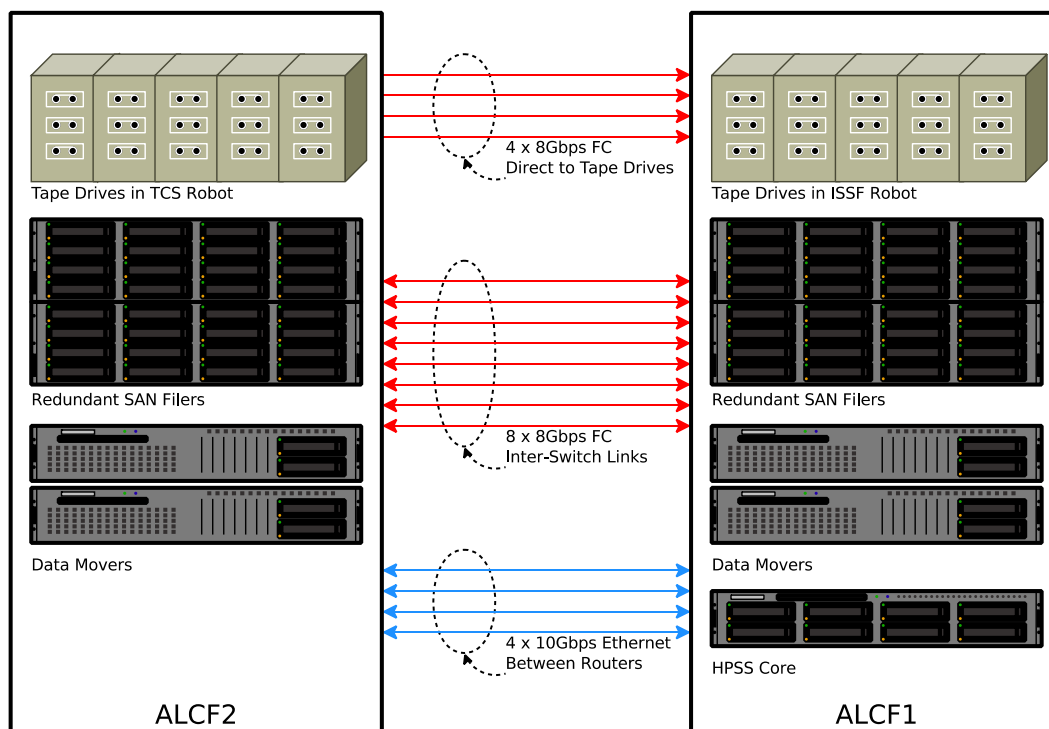


Figure 2.

6.2.2 Process of Science

Groups awarded time at the ALCF often transfer in their data sets to the facility at the beginning of their computing time. This comes in the January time frame for INCITE awards, and the July time frame for ASCR Leadership Computing Challenge (ALCC) awards. Groups also transfer out simulation results to their home facilities or to collaborators year-round. The INCITE process, which allocates 60% of the available time on the machine, is open to international proposals. In the past several years, there has been at least one trans-Atlantic awardee per INCITE year. Mira has close to a petabyte of memory; this feature will likely attract new projects that have large inputs from other sites and generate large outputs of interest to other sites. Below is a list of known, representative high-data users.

1. **Hardware/Hybrid Accelerated Cosmology Code (HACC).** Moved 20 TB of data from Los Alamos National Laboratory (LANL) using Globus Online. This workflow had an interesting issue related to cyber security — LANL prohibited the use of an automated transfer, and required the user to periodically indicate that the transfer was being monitored (Figure 3.)



Figure 3.

2. CyberShake

- The Southern California Earthquake Center (SCEC) computational research program makes extensive use of distributed, grid-based scientific workflows. The Pegasus workflow is used to reference logical file names to query data registry services such as Globus Replica Location Service, and supports staging of statically linked executables on demand. CyberShake requires approximately 250 K to 840 K individual tasks per geographic site for the analysis.
- From 2009: We had some challenges moving large files from National Science Foundation (NSF) systems (e.g., Kraken) to ALCF. We often need to transfer large (>5 TB) input files to ALCF and simulation results from ALCF with destinations at academic, NSF, and other DOE sites. A robust, easy to use implementation of GridFTP would make these data transfers significantly easier and faster for us.

3. **Climate End Station.** Climate End Station represents the kind of project that performs simulations whose results need to be shared, in contrast with projects with input data sets that need to be transferred. Our understanding is that their project may copy multiple petabytes of data to the National Aeronautics and Space Administration (NASA), requiring hundreds of terabytes of buffer space for files awaiting transfer.
4. **LaserPlasmaSim.** This project generates large amounts of data from simulations, but has relatively modest transfer requirements for analysis artifacts. Currently the transfers are about 8 MB/day. However, after a run is complete, the history dumps, the last one or two restart dumps, and the movie data are usually retained. This would represent approximately 30 TB of archival storage. The endpoint for the transfers is Lawrence Livermore National Laboratory.
5. **PHASTA.** PHASTA is an example of a project running visualization simultaneously with simulation. During the LAN scenario for live visualization of the PHASTA

simulation between Intrepid and Eureka, the network sustained ~54 GB/sec (432 Gbps). This was primarily limited by the network switch (Myrinet) capability. In another experiment in which we used Intrepid as a source and the 128 file servers + 100 Eureka nodes as sinks, we sustained ~84 GB/sec (672 Gbps). Their home institution is the Rensselaer Polytechnic Institute (RPI).

6. **General GridFTP use.** Susan Kurien's Multiscale Coupled Turbulence project transferred 21 TB over 30 separate dates, with a maximum of 8 TB in a day using GridFTP. The LatticeQCD project transferred 9 TB over 18 separate days, with a maximum of 3 TB in a day using GridFTP. One positive note on using GridFTP for data transfer: The server can be configured to keep logs of specific transfers, to aid in analyses like this.

6.3 Science Drivers – the Next 2-5 Years

6.3.1 Instruments and Facilities

Two significant upgrades are planned for the ALCF2 infrastructure beyond the next two years.

Anticipating extensive growth in data analytics, the ALCF has plans to more than double the size of the Tukey visualization cluster after two years. With a target design of 200 nodes, each capable of streaming real-time video, the need for significant external connectivity will continue to grow. The ALCF sees this machine supporting multiple projects simultaneously, potentially streaming to multiple target sites.

As data sets grow, so do storage requirements. To achieve a reasonable balance between excessive data sizes and excessive storage consumption, a High Performance Storage System (HPSS) data-mover cluster is planned to join the ALCF2 infrastructure after two years. These systems will manage data migration and caching for HPSS, providing higher throughput for users with large amounts of data to be studied over longer time periods.

6.3.2 Process of Science

The need will increase for in situ analysis, as the ability to dump raw data into a file system for later analysis will be prohibitively expensive in terms of time. It may be that this analysis will take place on the computational cluster itself, or that it will happen in real time with a companion visualization and analysis resource located on the WAN. Either way, the planned infrastructure upgrades should accommodate the change in use patterns of the compute cluster, the file system, and the analytics nodes.

6.4 Beyond 5 Years — Future Needs and Scientific Direction

The ability of HPC resources to produce data in the future time frame may be comparable to the ability of current large instruments (for instance, the Large Hadron Collider [LHC]) to produce data, and may require a similar set of collaborative analyses, where a small number of sites store mostly raw data, while progressively more distilled

data is passed along to other collaborators. This mix of collaborators will probably include both national laboratories on ESnet and universities on Internet2. It remains to be seen whether the reduction will happen in-place on the analytics machines, or whether the data replication will happen in real time to remote sites.

Additionally, there may be a greater call to share the large raw data from simulations so that research groups can perform verification and validation of the results from large simulations, rather than relying solely on the results of the original analysis.

6.5 Network and Data Architecture

As mentioned in Section 6.2.1's *Data Transfer Nodes*, the high-bandwidth public connectivity from the LIONs may facilitate novel methods of data transfer and analysis from an ongoing simulation, particularly after the planned upgrade of the Tukey system in two years. In terms of campus connectivity, the DWDM planned between our two data centers should enable high-speed transfer of data from ALCF1 to ALCF2.

6.6 Collaboration Tools

The use of videoconferencing tools is driven by the workflows of the individual science groups. The site has participated in webinars for INCITE proposals, and individuals participate in Skype and conference calls with their INCITE projects. We have recently begun recording conference audio and video for sharing with those unable to attend the meeting. Our building has just installed new videoconferencing facilities, which may increase the rate of remote meeting attendance.

6.7 Data, Workflow, Middleware Tools, and Services

A few projects at ALCF make note of their data workflow tools. SCEC and the CyberShake project, mentioned above, manage their data workflow using a tool called Pegasus, which uses the Globus Replica Catalog to locate replicas, and GridFTP to transfer the data. Other users, for instance HACC, have used Globus Online to manage the GridFTP transfer process with retries and partial file transfers. ALCF1's GridFTP endpoint has proved very stable, and is used by the Globus Online team for testing purposes.

6.8 Outstanding Issues

No outstanding issues at this time.

6.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
Near Term (0-2 years)				
ALCF1	INCITE and ALCC awards from national laboratories, universities, corporations, and international partners	Size varies, approx. 200 TB per award, and approximately 100,000 files per award	Largest simulations can move 84 GB/sec for the life of the job (up to 12 h)	- Largest users move up to 10 TB/day - Targets are globally diverse, but typically either ESnet or Internet2
2-5 years				
ALCF2	Possible increase in sharing of data sets in the Earth Science communities	Initial sizes expected to match ALCF1, system is sized for approximately 3X overall storage for same number of projects	Largest simulations will be able to achieve greater speeds, with 384 GB/sec peak theoretical across LAN	- Expectations are to reach 3x ALCF1 levels (up to 30 TB/day) - Targets will remain globally diverse
5+ years				
ALCF2 Upgrade – Tukey Visualization	See Section 1.3.1 – More than 2X increase in data analytics capacity	Less data is stored, and more data is streamed remotely	200 nodes with up to 800 GB/sec theoretical aggregate bandwidth on LAN	With live real-time visualization, up to 25 GB/sec (200 Gbps) will be possible

7 National Energy Research Scientific Computing Center (NERSC)

7.1 Background

The mission of the National Energy Research Scientific Computing Center (NERSC) is to accelerate the pace of scientific discovery by providing high-performance computing, information, data, and communications services for research sponsored by the DOE Office of Science (SC).

NERSC is the principal provider of high-performance computing services to SC programs — Fusion Energy Sciences, High Energy Physics, Nuclear Physics, Basic Energy Sciences, Biological and Environmental Research, and Advanced Scientific Computing Research. NERSC has more than 4,500 active users

NERSC is highly accessible and focuses on computing and data strategies that have high impact on the science productivity of our users. NERSC derives key benefits in that regard from its network connectivity through ESnet. Overall, NERSC is very satisfied with ESnet's directions and strategy.

7.2 Key Science Drivers

7.2.1 Instruments and Facilities

Hopper is NERSC's current main source of compute hours. The 150 K cores and 212 TB of memory are capable of producing simulation output at data rates that require substantial file system and networking capabilities. Simulation-driven data requirements have been detailed in the various program office requirements review findings and make their way into NERSC system and service plans.

On average, NERSC is an importer of data over the WAN. This trend has been true going back 10 years. Data from simulations and instruments comes to NERSC through XRootD, GridFTP, and BBP, to name a few transport mechanisms. Telescopes and sky surveys such as Palomar Transient Factory (PTF) use automated data mobility workflows to transfer data to NERSC every night for analysis, with data rates described in the diagram below.

The overall historical context of WAN data movement from NERSC computer systems is shown in the appendix, which summarizes the past 10 years of WAN traffic. The trend is unmistakably exponential and points directly to the need for terabit networking technology in the 2016 time frame.

2012 also saw a major transition in the data, computing, and networking strategy that the Joint Genome Institute (JGI) has taken. As the cost of sequencing has dropped, the data created has risen super-exponentially. As a consequence, data analysis and management issues have risen to the forefront of genetic bioscience.

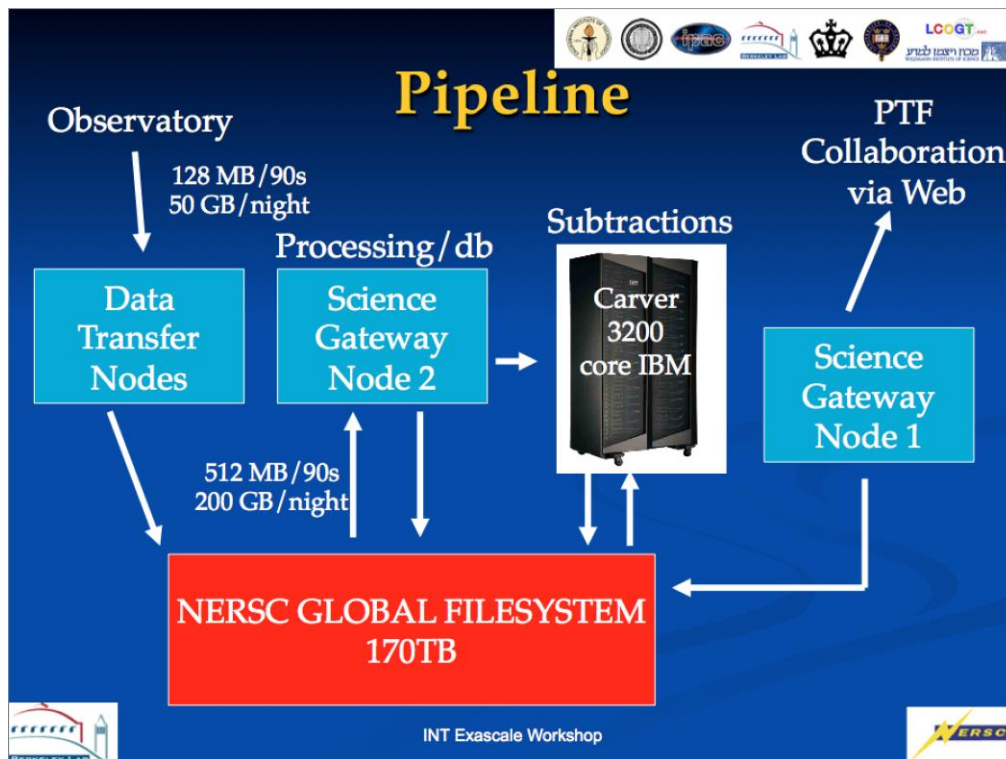


Figure 4. The Palomar Transient Factory (PTF) data pipeline. Nightly ingest of data followed by automated analysis guides the instrument's next steps.

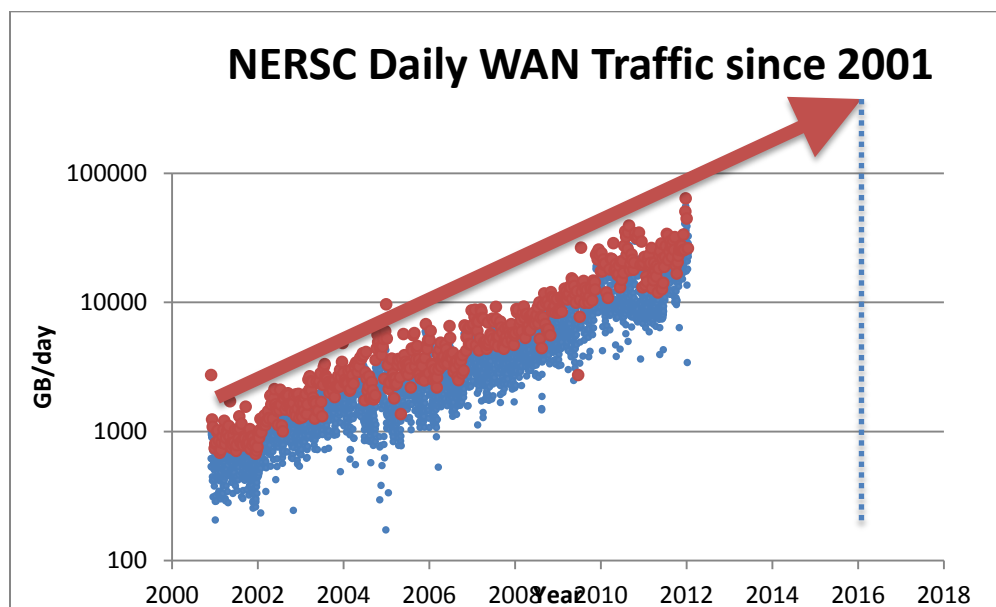


Figure 5. NERSC WAN data movement over the last ten years. Blue points are daily totals and red points are weekly maxima. The arrival of terabit networking in 2016 will match pace with the data needs based on these trends.

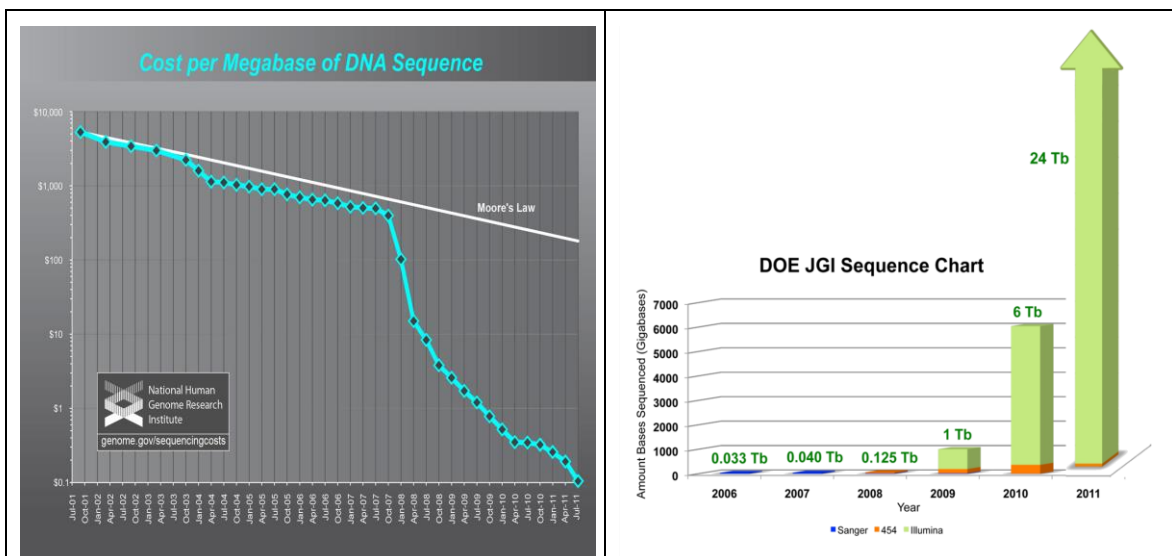


Figure 6. Decadal trends in biosequencing costs and rates. Advances in bioscience have driven super-exponential increases in data and networking. For DOE bioscience, it can be argued that research has transitioned from the wet chemistry lab to the data lab.

Astronomy and bioinformatics are well into a transition to data-centric science. As telescopes and sequencers move toward higher resolution, higher output, and greater automation, the science communities' attention, workflow, and breakthroughs are increasingly attached to data streams produced by these instruments. Methods to scalably manage data and make it easily available for collaborative discovery are thus among the most in need of R&D for these fields.

Future Priorities

Looking forward, NERSC identifies BES instruments as a likely source of the next wave of data-centric science projects. A similar wave of Biological and Environmental Research (BER) bioimaging data needs are predicted just behind those beginning in the photon sciences at BES-funded beamlines. As described in later sections, advances in detector resolution, repetition rate, and automated sample analysis are driving a similar move toward data being at the center of attention. Key strategic aspects to this new development include:

- A double exponential growth of detector resolution and experiment repetition rate points to a shift in photon science workflows
- Automated data pipelines → ESnet5
- Large-scale image processing (tomography, k-space, segmentation) → NERSC7,8
- Community access to data and analysis → NERSC science gateways

To the degree that beamline science is bracketed by data analysis bottlenecks, BES facilities and instruments will not deliver their full scientific value. Major DOE facilities investments hinge on big data performance. For some facilities, advances in detector/sequencer technology have disrupted plans for how these facilities operate. Without a targeted effort in solutions for extreme data challenges, DOE mission science output may be constrained.

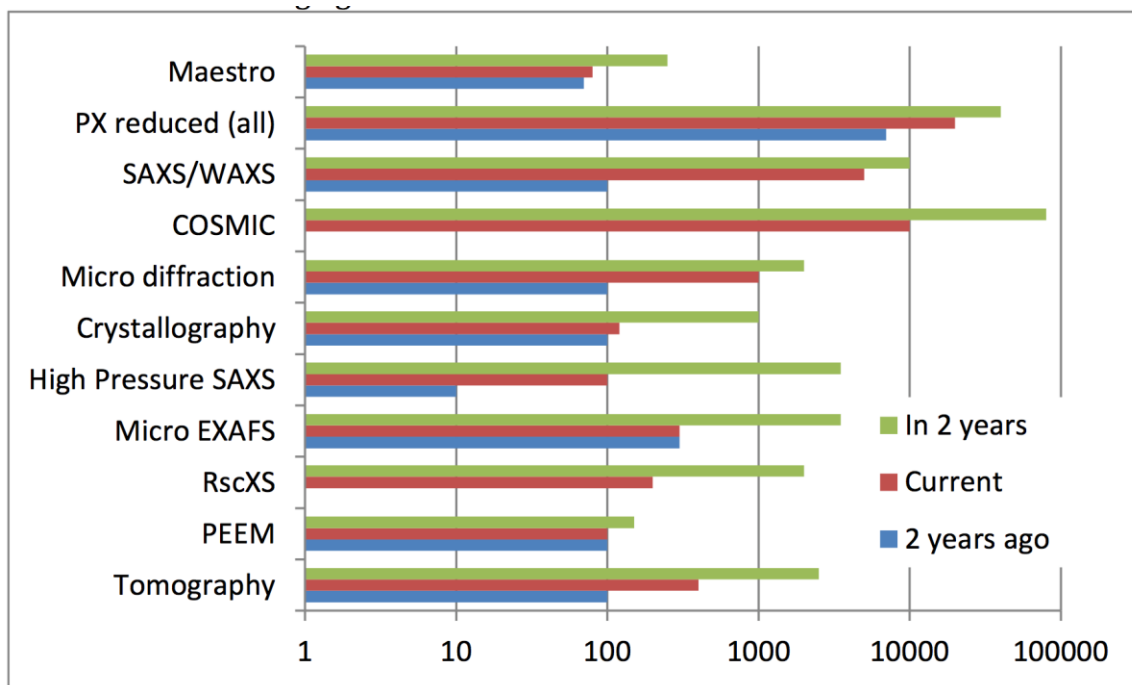


Figure 7. Results from a recent user survey at ALS that summarize the average data volume in GB/month for a variety of beamlines collected during a single month. The light blue bar highlights the average data volume two years ago, red the current volume, and yellow the projected volume due to beamline and detector updates in two years. Currently, the 18 beamlines in the survey generate about 300 TB of data per year. This is expected to rise to almost 2 petabytes.

The overarching risk of underperforming extreme-scale data solutions is an effective cap on the scientific ambition as to what can be attempted in the beamline, microscope, and other facilities. We already see restrictions on how much data can be taken, how many samples analyzed, or the resolution of the experiment due to data constraints. Given the present predictions of trends in extreme data, these minor constraints will, left unchecked, grow into fundamental limits on science output. Given the existing and likely trends toward data-centrism in the DOE SC science portfolio, it is imperative to examine and optimize the infrastructure and methods that underlie how scientists work with data. In large part these can be described as being the network through which data flows, the interfaces through which data analysis is made available, and new data-centric algorithms and methods.

NERSC Data and Science Gateways

NERSC maintains two data transfer nodes (DTNs) whose specification follows the best practices from ESnet on WAN data transfer (<http://fasterdata.es.net/science-dmz/>). The DTNs mount large parallel high-bandwidth file systems at NERSC and provide WAN data movement through Globus Online, GridFTP, BBP, and other data tools. Users either log in to these nodes directly, use a hosted service like Globus Online, or use other implicit and third-party data-movement methods. 2012 was the year that the DTNs made their way to the broad NERSC user population. These advances in the ease of use of high-performance data technologies along with burgeoning data volumes have led to a rush of users toward the DTNs. This recent development is detailed in the figure below that shows pickup in adoption of Globus Online at NERSC.

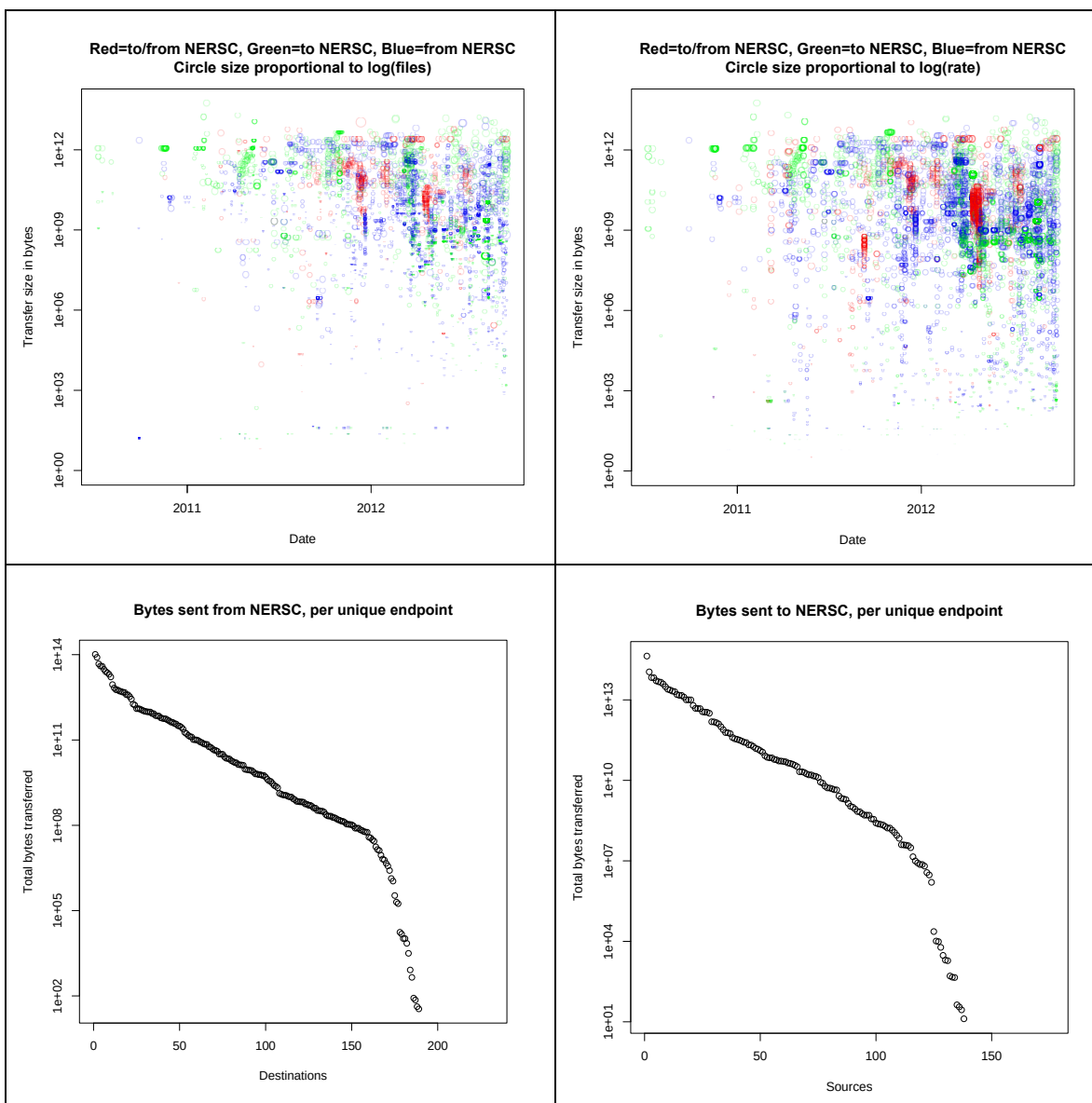


Figure 8. Recent trends in the adoption of high-performance data-movement methods through Globus Online show a substantial pickup in adoption in the past year, yielding distinct DTN endpoints in the hundred-plus range. NERSC has 5,000 users in total.

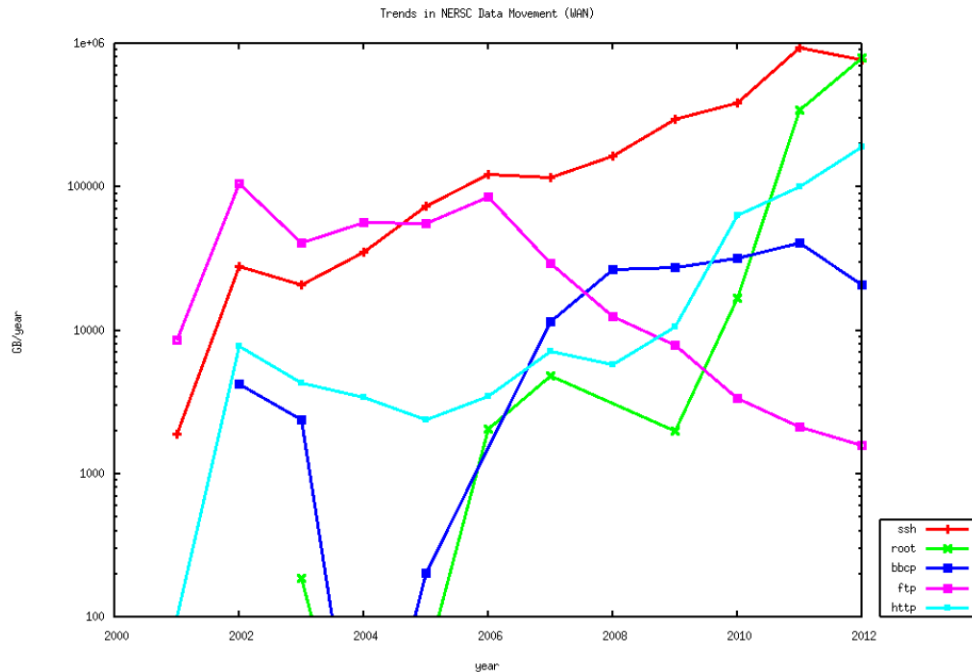


Figure 9. Decadal trends in network traffic related to data transfer. The y axis is a log scale showing the growth of HEP-focused XRootD (labeled root) transfers from the LHC and Web-based science gateway traffic. GridFTP is absent from this analysis but is shown in the previous figure.

NERSC is expanding the pool of DTNs from two to approximately 10. This work will be done with coordination from the DTWG (Data Transfer Working Group), with the goals of sharing best practices in infrastructure deployment and also optimizing DOE facilities' mutual benefit from the 100 Gbps ESnet5 network rollout. The imagined end state for this activity is a set of interoperable DTNs across DOE facilities that can fully drive the ESnet5 bandwidth.

NERSC assists science teams in building Web interfaces to access HPC computers and storage systems. These gateways allow scientists to access data, perform computations, and interact with NERSC resources using Web-based interfaces to run simulations and analyze data. The goal is to make it easier for scientists to use computing resources at NERSC while creating collaborative tools for sharing data with the rest of the scientific community. Greater access means greater science impact by increasing the net value of ASCR computing and data investments.

NERSC engages with science teams interested in using these new services, assists with deployment, accepts feedback, and tries to recycle successful approaches into methods

that other teams can use. Below is a list of current projects and methods. Building blocks are available to NERSC users interested in creating new science gateways.

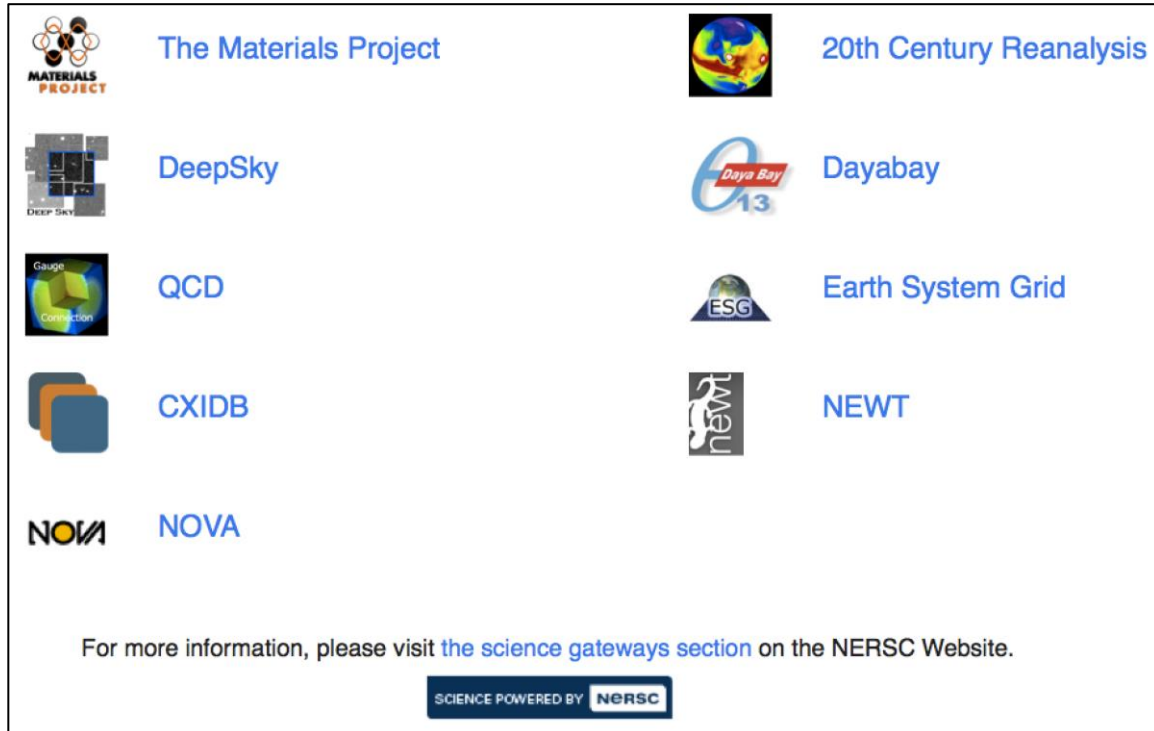


Figure 10. Science gateways in production at NERSC.

7.2.2 Process of Science

The following vignettes accurately describe a large portion of the processes currently in use by which scientific research happens at NERSC, with emphasis on the networking aspects of the processes involved.

- Downloading/publication of simulation data sets for secondary analysis. NERSC maintains interfaces to data on spinning disk and on tape that allow for user access
- Visualization involving remote users (NX is of growing demand)
- Ingest/export of data sets from other facilities (telescopes, light sources, etc.)
- DTNs are often used in the import process when new projects have time allocated at NERSC and their data inputs for simulation or analysis need to be moved to NERSC machines. The reverse process happens by the same accord as allocations for time shift from center to center.
- Long-term stable data repositories that are widely accessible, e.g., raw climate-simulation data curation

7.3 Science Drivers — the Next 2-5 Years

7.3.1 Instruments and Facilities

NERSC7 will be delivered in the next year and, along with new computing capabilities, will produce by the simplest estimates a factor of 2.5 increase in data and networking needs.

Advanced Light Source (ALS) and BES: New initiatives between the ALS and NERSC will drive data pipelines between the facilities to liberate data from the beamline in ways that make its analysis more scalable and the science more sharable. Initial work on these pipelines was demonstrated in 2012, and in 2013 a funded Laboratory Directed Research and Development (LDRD) effort will bring this networked data analysis to BES beamline users. In some cases, beamline scientists use HPC-backed analysis while still at the beamline to guide and shape their investigations. In other cases, data is collaboratively re-analyzed and curated in ways that allow communities of remote scientists to interact with and benefit from beamline data sets.

This workflow is detailed in the following figure.

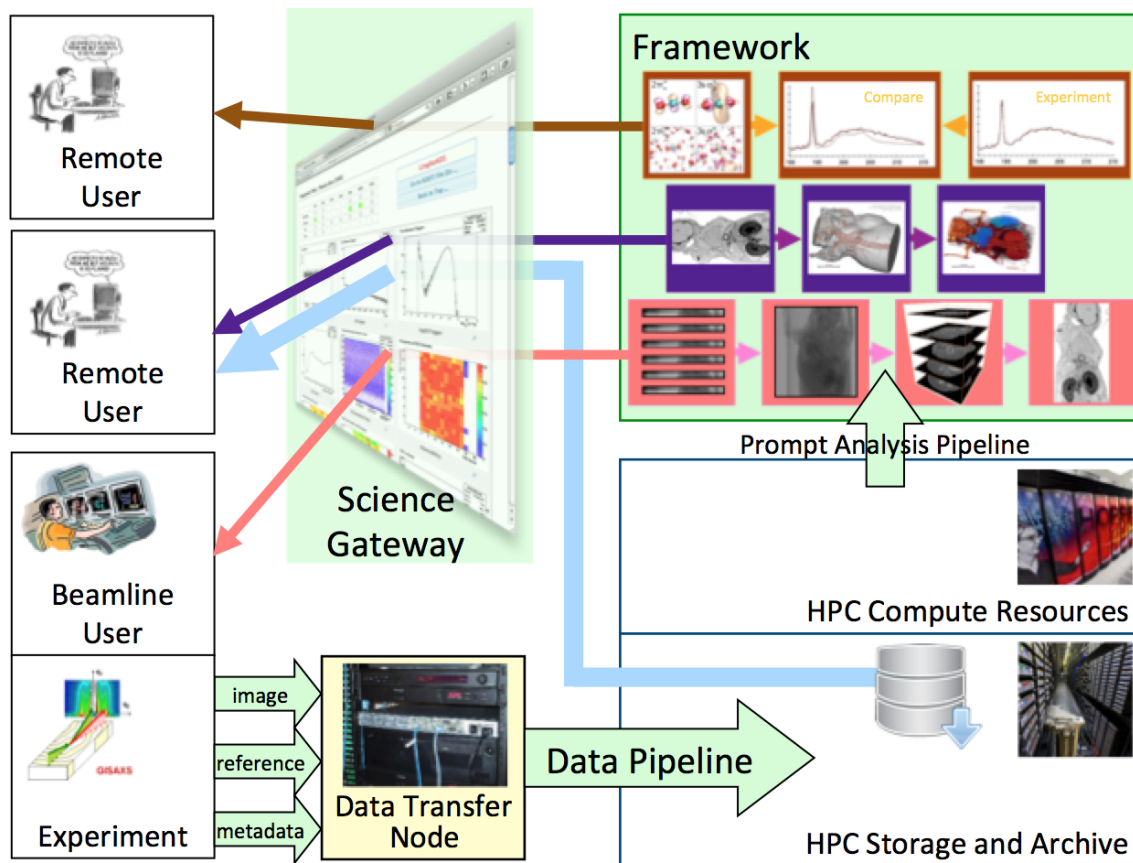


Figure 11.

Linac Coherent Light Source (LCLS): Recent work with researchers in X-ray science has demonstrated that once a high-repetition-rate high-resolution device such as LCLS is taking data, a variety of data challenges immediately arise. The main ones identified are: How to store the data, if only temporarily, at the facility; how to transfer it to a computing center for analysis; and how to archive the data such that it can be made useful multiple times. The LCLS/NERSC team recently reported transfers of 11 TB at 500 MB/sec, which comes out to about 22,000 seconds, or a bit over six hours.

Materials Genome Initiative: The materials project (www.materialsproject.org) is a science gateway residing at NERSC. Its database is expected to grow substantially from 100 GB toward 500 GB in the coming years. This data, while small compared with HEP data, is complex in its variety and variability and must be made highly searchable at low latency over a Web interface.

LHC/ALICE (A Large Ion Collider Experiment) and STAR (Solenoidal Tracker At RHIC): Each of these projects is budgeted for ~500 TB/year of data, which will move through XRootD (ALICE) and GridFTP (STAR) to NERSC for analysis and storage.

The Baryon Oscillation Spectroscopic Survey (BOSS) (current), eBOSS (2014), and BigBOSS (2017) all have similar and growing data trajectories that call out the need Globus Online-like data sharing capabilities.

7.3.2 Process of Science

The projects in the next few years, mentioned above, paint a reasonable sketch of resources and methods that must be in place in the coming years. We need data pipelines that are effective both for achieving network bandwidth and for accessibility. Increasingly, that means automated, hands-off, reliable delivery of data. NERSC will grow into areas that increasingly look like database services for science teams. Access to those databases over the network will drive their usefulness. Building software interfaces, as gateways or as APIs, to data and computing is the capstone of these directions, because no matter how much plumbing and wiring there is underneath, the software interface is what the researcher will ultimately see and use.

7.4 Beyond 5 Years – Future Needs and Scientific Direction

Beyond five years, the specifics of what is needed in computing and networking are harder to illuminate. Some of the best thinking about such issues has been done at this review and in the reports from community discussions on these topics. The Scientific Collaborations for Extreme Scale Science workshop in December 2011 produced the following high-level findings that are relevant to this current discussion:

Finding: Integrated data operations — unimpeded discovery and collaborative exploration of all relevant data — is a dream currently out of reach of most scientists engaged in extreme-scale research. The dream seems particularly distant for interdisciplinary collaborative science. Where the dream can be realized, the benefits to extreme-scale science will be revolutionary.

Recommendation: Solicit proposals in the area of Integrated Data Operations for research into groundbreaking new tools, and for the consolidation and generalization of promising existing approaches.

Finding: A rigorous treatment of provenance is pivotal in extreme-scale science. Provenance allows complex analyses to be validated, allows credit to be rigorously attributed without the latency of passing through journal publications, and brings many practical benefits such as avoiding the need to instantiate and preserve every derived data set. Current provenance-tracking technologies have had little impact on extreme-scale science.

Recommendation: Solicit proposals focused on moving existing provenance technologies into extreme-scale science, thus exposing the extent to which both fundamental research and generalization/support are required. Research and development to fill clearly identified gaps should also be encouraged.

7.5 Network and Data Architecture

NERSC currently uses ESnet as its sole provider of Internet connectivity for the Oakland Scientific Facility (OSF). NERSC uses several diverse fiber paths into OSF. Two paths support the legacy 10 Gbps connection between NERSC and ESnet. These 10 Gbps links are part of the legacy Bay Area Metropolitan Network (BAMAN) metro ring and are slated to be retired by the end of 2012. NERSC also has several new fiber paths that carry 100 Gbps wavelengths from ESnet into NERSC. These wavelengths terminate in two diverse locations within ESnet: Sunnyvale and Sacramento. This mitigates possible outages due to either natural or man-made (backhoe) disasters. NERSC does have the long-term risk of only using a single Internet service provider (ESnet); should ESnet have an outage, NERSC has no alternate path to the Internet. The 100 Gbps wavelengths will increase general Internet bandwidth to universities and DOE laboratories as well as support dedicated channels to other DOE facilities such as telescopes, light sources, and other large-scale data-acquisition instruments.

In CY 2015, NERSC will transition to the Computational Research and Theory (CRT) building at the main LBNL campus. NERSC will require several 100 Gbps wavelengths of dedicated bandwidth between the two facilities to make this transition. Current estimates are for co-occupation to last from CY2015 through CY2017.

7.6 Collaboration Tools

NERSC provides remote-desktop services through NX. These services are of increasing popularity and allow high-bandwidth, low-latency access to remote software for visualization. Persistent connection and terminal connections allow NERSC users to checkpoint and restart their working sessions at NERSC independent of their location.

We make increasing use of ReadyTalk for both staff and user presentations. Bringing H323 videoconferencing methods into a more accessible mode would be one idea to bring ESnet collaboration technology to more of the science community.

The Scientific Collaborations for Extreme Scale Science workshop report presented a clear set of guiding concepts for tools to enable collaboration in science:

- **Discovery:** All resources are easy to find and understand. (“I cannot use resources that I do not know exist!”)
- **Centrality:** Standardized services reduce costs and encourage commonality. (“Don’t make me install software or learn arcane details to collaborate!”)
- **Portability:** Resources are widely usable in a transparent fashion. (“If I can’t use your data or software, it isn’t science!”)
- **Connectivity:** Where information came from, and what other information it relates to, are easy to find. (“No information exists in isolation!”)

7.7 Data, Workflow, Middleware Tools, and Services

NERSC’s Science Gateways efforts are built on the foundational experience that if middleware is complicated, it will at best find adoption only by large-scale science teams that can marshal the considerable effort to integrate and deploy such systems. We seek a more agile approach through the following goals:

- Easily expose scientific data on NERSC Global File systems (NGFs) and HPSS archive to larger communities.
- Allow all team members to analyze large data sets and manage computational workflows and batch jobs remotely through the Web.
- Broaden the scientific impact of computational science through Web methods that are as easy as online banking.

A principle path to those goals is a framework suitable for science teams of all sizes to participate in making the gateways that best serve their research needs. NERSC, Lawrence Livermore National Laboratory (LLNL), and Texas Advanced Computing Center (TACC) are loosely collaborating to achieve this through Representational State Transfer (REST)-ful Web APIs. REST is easy to learn, can be portable, and solves many middleware complexities that have previously prevented all but the largest science teams from building scalable Web interfaces to their science. The API in use at NERSC is the NEWT API (NERSC Web Toolkit/API, <http://newt.nersc.gov>), which attempts to make common HPC needs easily addressable through the Web. In summary, NEWT provides:

- Building blocks for science on the Web that cover the same tasks most scientists currently approach through the command line
- A means for science teams for to write science gateways using simple HTML + Javascript
- Support for authentication, submission, file access, data analysis, accounting, start/stopping compute jobs, viewing queues, user-defined data tables

So far, more than 30 projects at NERSC use this Web-based gateway approach to organize, share, and interact with their data and computing.

PerfSONAR is quite useful to NERSC. One minor request: A clickable map with common names for institutional hardware as opposed to knowing router names would make it more accessible to the non-network engineer.

7.8 Outstanding Issues

There is a need for optical paths between NERSC's OSF location and the LBNL campus. One path forward would be to provision additional circuits on ESnet5 between NERSC's current Oakland location and its new location in the CRT building on the LBNL main campus.

The provisioning of network circuits remains somewhat opaque to user communities that may benefit from them. Given knowledge like "we'd like to move X TB from A to B on or around this date," the steps to evaluate whether a circuit makes sense and how to provision it remain somewhat unclear.

Services that communicate ESnet resource utilization to computing center staff are a much-needed means to foster interoperation of DOE facilities. Services like <https://my.es.net/> are very much appreciated. For NERSC, it is useful to cross-check our measurements and conclusions about resource utilization.

7.9 Charge: Is Terabit Networking a Requirement in the 2016 Time Frame?

Answer from NERSC:

1. Yes. The last decade of ESnet traffic at NERSC shows this. To meet trends in data demand in 2016, an order of magnitude increase is required.
2. Yes. The science case for terabit networking in 2016 is strong. Discoveries in photon and neutron science will come from data-centric capabilities expanding the scope of BES instruments.

Details

NERSC WAN interface monitoring. Based on daily reports gathered over the past 10 years, various trends in network utilization are apparent. The data in Figure 5 show a longitudinal view of data movement to and from NERSC. Measured in the same way across the same interfaces for the duration, the data is very systematic and provides trending detail as to NERSC data and network needs.

7.10 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
Near Term (0-2 years)				
BES synchrotron light source beamlines	Processing of data to resolve microstructure, diffraction imaging, tomography from 2-D to 3-D, reverse Monte Carlo simulation for small angle X-ray work	- Data set volume highly variable (LCLS = 11 TB) - Data set composition images	10 TB in 3-6 hours	2009 = 65 TB 2011 = 312 TB 2013 = 1.9 PB - APS, LCLS, LBL, NERSC
2-5 years				
Bioimaging sources such as protein surveys and confocal electron microscopy data	Data sets are too large to analyze near the microscope. Image segmentation and reduction from pixels to geometric models will require large-scale data and computing.	- Data volume 1 TB/day per microscope - Data set composition 32k x 32k x 20k image stacks	1 TB per hour, 4 times per day	300 TB/year
5+ years				
BigBOSS 2017	Image processing for highest resolution 3-D map of space yet made			

8 Oak Ridge Leadership Computing Facility (OLCF)

8.1 Background

Oak Ridge National Laboratory's (ORNL's) Leadership Computing Facility (OLCF) delivers the most powerful resources in the United States for open science. At 2.33 petaflops (PF) peak performance, the Cray XT Jaguar delivered more than 1.4 billion core hours in 2012 to researchers around the world for computational simulations relevant to national and energy security; advancing the frontiers of knowledge in physical sciences and areas of biological, medical, environmental, and computer sciences; and providing world-class research facilities for the nation's science enterprise. The OLCF continues to fulfill this mission through its fielding of a 20 PF Cray XK7 system named Titan. The OLCF leverages massive data storage, high-bandwidth network connectivity, and advanced visualization resources to deliver the world's leading science cyber-infrastructure.

8.2 Key Science Drivers

8.2.1 Instruments and Facilities

Leadership computing is listed as the highest domestic priority in the DOE-SC report *Facilities for the Future of Science: A Twenty-Year Outlook*. Upgrade of the leadership computing facilities to tens of petaflops within the 2011–2013 time frame is vital to the United States playing a leading role in several important international programs, including climate science (Intergovernmental Panel on Climate Change), fusion energy research (ITER), and the Nuclear Energy Advanced Modeling and Simulation program. Moreover, the United States faces serious economic, environmental, and national-security challenges based on its dependence on fossil fuels. To address the scientific grand challenges identified by DOE-SC programs alone would require a leadership-class computing capability of at least 100 PF by 2015. In the near term, based on the requirements for making incremental steps in application software and the projected availability of technology, DOE-SC has a mission need for a total leadership-class computing capability of 20–40 PF in the 2011–2013 time frame.

To allow substantial advances on near-term requirements in numerous mission-relevant science domains, the OLCF will deliver a 20 PF computing capability, Titan, as part of a DOE-SC strategy that requires architectural diversity of computer systems to minimize risk within the program.

ORNL is participating in the DOE/ESnet Advanced Networking Initiative (ANI) that provides a native 100 Gbps optical network connection among DOE SC sites, including ORNL, ANL, LBNL, and other facilities in the northeast. Additional connections into ORNL include the NSF XSEDE and the University of Tennessee. To meet the increasingly demanding needs of data transfers among major facilities, ORNL has provisioned extra capacity into the border and WAN infrastructures to accommodate substantial growth.

Cray XK7 “Titan”

Titan, a Cray XK7 system, is the third generation of major capability computing systems at the DOE SC OLCF. It is an upgrade of the existing Jaguar system first installed at the OLCF in 2008. The initial upgrade from Cray XT5 to Cray XK7 compute nodes was accepted in February 2012 and consists of 18,688 compute nodes for a total of 299,008 AMD Opteron 6274 Interlagos processor cores and 960 NVIDIA X2090 Fermi Graphical Processing Units (GPU). The peak performance of the Opteron cores is 2.63 PF and the peak performance of the Fermi GPUs is 638 teraflops (TF). In late 2012, the 960 NVIDIA X2090 processors will be removed and replaced with at least 14,592 of NVIDIA’s next-generation Kepler processors, with a total system peak performance of the GPUs in excess of 20 PF.

The OLCF worked with Cray to design Titan to be an exceptionally well-balanced system for modeling and simulation at the highest end of high-performance computing. The AMD Opteron processors double both the memory bandwidth and memory capacity per node as compared with the Jaguar Cray XT5 system it replaced. The system will be linked to its file system by twice the number of I/O nodes and will use InfiniBand (IB) host channel adaptors (HCAs) that provide at least twice the bandwidth of the IB HCAs in Jaguar. The file storage system is being acquired independently from the Titan system and will have at least twice the bandwidth and capacity of Jaguar’s file system. The key new component of Titan is that most of the Cray XK7 nodes have an NVIDIA GPU application accelerator. In the November 2011 Top500 list of the world’s most powerful computers, 39 of the 500 computers on the list used application accelerators, including three of the five fastest computers.

OLCF Storage

As ORNL transitions to the heterogeneous Titan platform, the storage system will be upgraded to support higher-aggregate I/O bandwidth as well as scalable metadata performance. This upgrade will include an aggregate performance increase to 500–1000 gigabytes per second (GB/sec) measured in terms of tens of petabytes as well as increased file system capacity. Whereas the current storage system is divided among four parallel file systems to improve aggregate metadata performance, horizontally scalable metadata performance through Distributed Namespace will allow higher aggregate metadata performance by distributing metadata workload based on namespace hierarchy. To support this approach, the OLCF plans to deploy multiple metadata servers (MDS) and distribute the namespace based on distinct users and project directories across these servers. Other changes in the Lustre software stack that improve vertical scalability of the MDS may allow higher-performance storage media technologies such as Flash-based metadata targets (MDTs) to improve single-server metadata performance.

Wide Area Network

The ORNL campus has access to every major research network at rates of 10 Gbps or greater. Layer 1 connectivity to these networks is provided via optical networking equipment owned and operated by UT-Battelle that runs over leased fiber-optic cable. This equipment has the capability of carrying multiple 10, 40, or 100 Gbps circuits. Twenty of the 10 Gbps circuits and two of the 40 Gbps circuits are committed to various purposes, providing virtually unlimited expansion of the networking capability.

Currently, ORNL connectivity to ESnet consists of a 10 Gbps IP connection to Nashville, a 10 Gbps Science Data Network (SDN) connection to Nashville, a shared commercial OC-48c connection to Atlanta, and the 100 Gbps ANI connection that will soon be transitioned to production. ORNL has a dark fiber infrastructure that provides last-mile fiber to Nashville, Chattanooga, Atlanta, and Chicago. This infrastructure carries the ESnet connectivity to Nashville. With ESnet5, the 100 Gbps connection is expected to replace the shared OC-48c to Atlanta, and the 10 Gbps connectivity to Nashville will be carried optically through Nashville to Chicago. The expected ORNL connectivity is depicted in the diagram below. PerfSONAR network monitoring equipment is deployed at the primary border router.

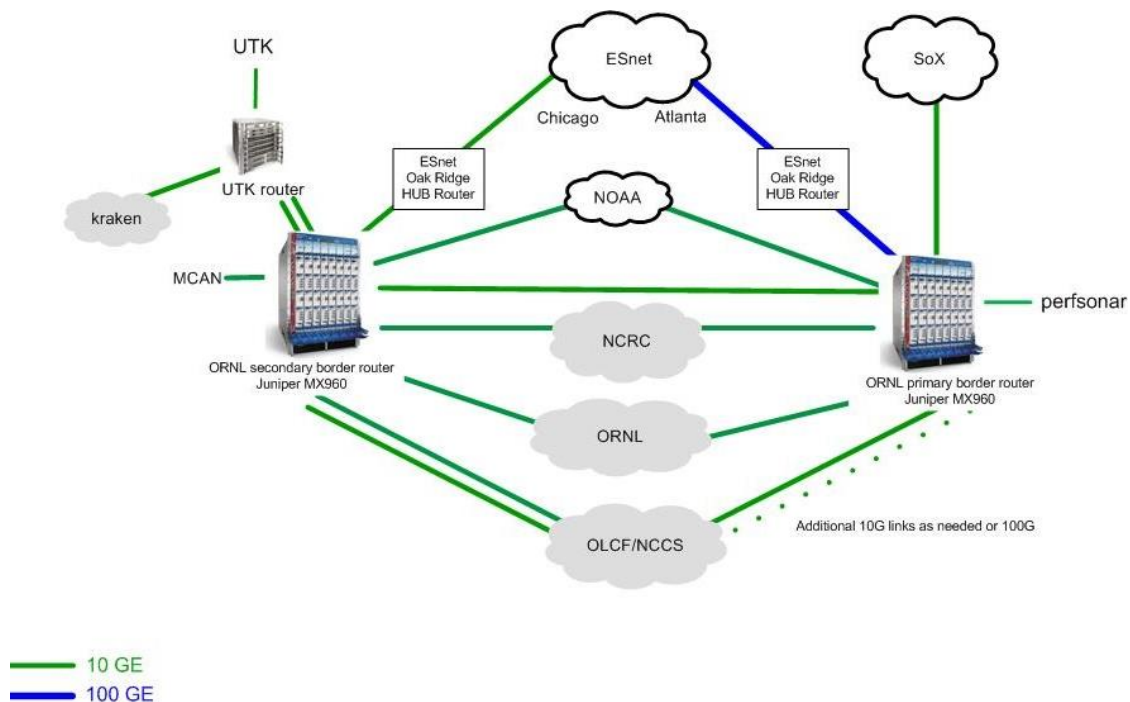


Figure 12. ORNL Network Overview

Local Area Networks

The local-area network is a common physical infrastructure that supports separate logical networks, each with varying levels of security and performance. Each of these networks is protected from the outside world and from one another with access control lists and network intrusion detection. Line rate connectivity is provided between the networks and to the outside world via redundant paths and switching fabrics. A tiered security structure is designed into the network to mitigate many attacks and to contain others.

OLCF Network Road Map

OLCF is testing and deploying stateful 10 Gbps line rate-capable firewalls and will move networks over to them in the near future. This is required in order to retire older hardware at the perimeter, and allow ORNL to move forward with the production connection to the 100-gigabit router. These firewalls are in a high availability configuration, ensuring greater reliability of the OLCF network.

Staff members have deployed a new router within the core of the OLCF network. This router gives the OLCF a path forward to 40 and 100 Gbps network connections, and potentially terabit-per-second connections a few years from now. This upgrade enables us to retire aging hardware, save funds on maintenance, and reduce power usage. It also frees up rack space for higher-density equipment. OLCF is reconfiguring the internal network to use more low-latency, high-speed, nonblocking switches. OLCF deployed this architecture for infrastructure services and the tape archive HPSS this year in order to provide a much more scalable upgrade path for the HPSS network.

OLCF currently has three 10 Gbps connections to ORNL. During the next three to six months, the migration to the line-rate firewalls will be complete. Production traffic will transit multiple 10 gigabit links to the 100 gigabit router at ORNL. The OLCF will deploy PerfSONAR internally at strategic points to monitor network performance. No fewer than eight new DTNs will be deployed within the Science DMZ. During the next six to 12 months, OLCF will purchase an additional 100-gigabit line card for the ORNL router, and a 100-gigabit line card at the OLCF demark. This will provide a dedicated native 100 Gbps path directly to the OLCF. In this same time frame, OLCF will purchase two 40 Gbps line cards for next-generation DTNs, and possibly 40 Gbps HPSS connections.

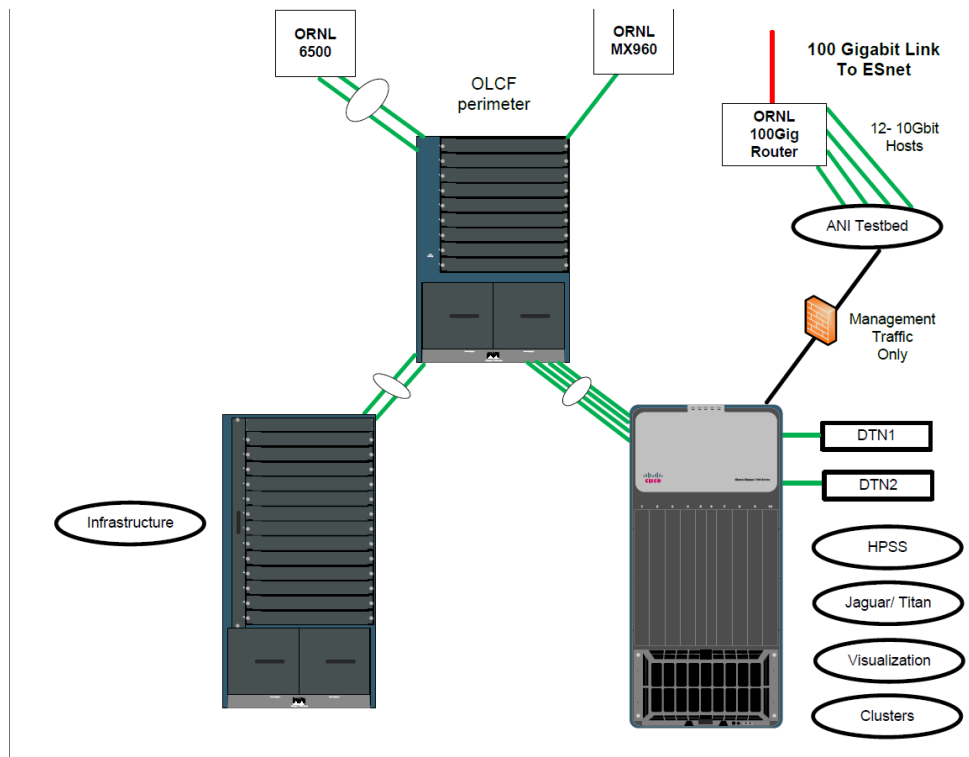


Figure 13. OLCF today.

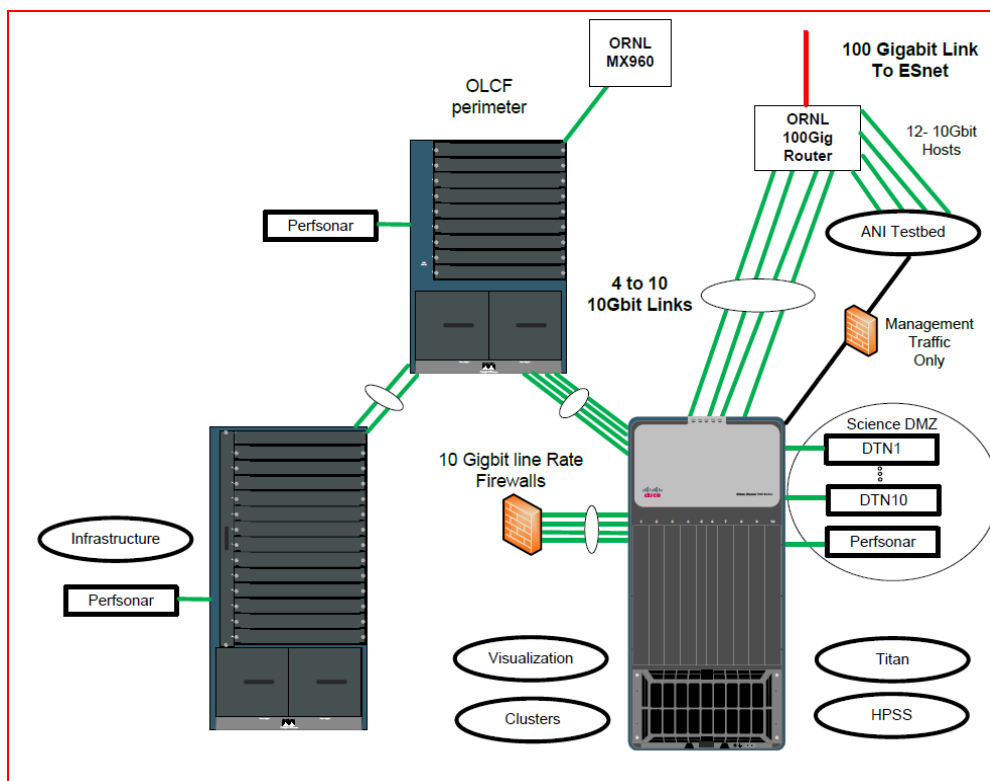


Figure 14. OLCF, three to six months.

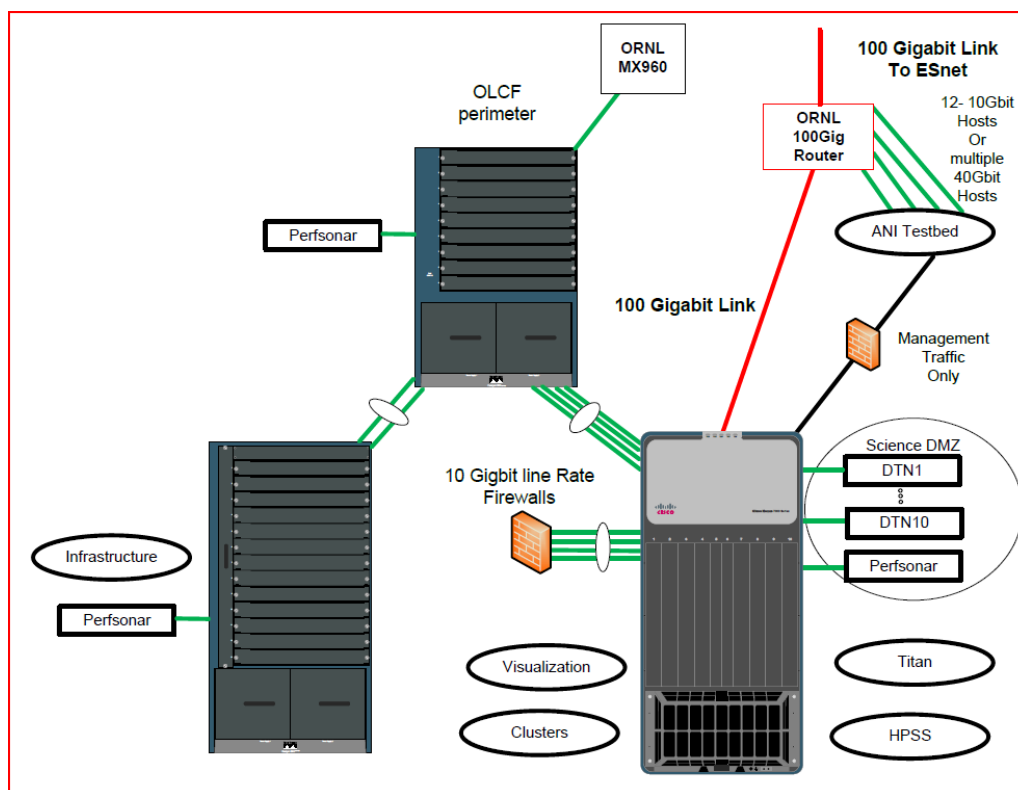


Figure 15. OLCF, six to 12 months.

Visualization Resources

Lens is a 77-node heterogeneous Linux cluster dedicated to data analysis and high-end visualization. Each of the 77 nodes provides 16 cores (four 2.3 GHz AMD quad-core Opteron processors) for a total of 1,232 cores. Forty-five of the nodes contain 128 GB of memory each. The other 32 nodes are configured with 64 GB of main memory and an NVIDIA Tesla GPGPU with 4 GB of memory and an NVIDIA 8800 GTX GPU with 768 MB of memory. As a whole, the cluster provides an aggregate of 7.8 TB of main memory and a high-speed IB interconnect. The primary purpose of Lens is to enable data analysis and visualization of simulation data generated on Titan so as to provide a conduit for large-scale scientific discovery. All Titan users are automatically granted access to Lens. Lens cross-mounts the Spider file system, obviating the need for data copy operations between Titan and Lens.

EVEREST (Exploratory Visualization Environment for REsearch in Science and Technology) is a large-scale venue for data exploration and visualization. The EVEREST room is undergoing renovation and will be completely reconfigured by January 2013. The EVEREST room contains two large-format displays. The primary display is a X30.5' x 8.5X' tiled wall containing 18 individual displays and an aggregate pixel count of X37 million pixels. It is capable of displaying interactive stereo 3-D imagery for an immersive

user experience. The secondary display is a 13.3X' x 7.5X' tiled display containing 16 individual panels, and an aggregate pixel count of 33 million pixels. Both displays may be operated independently, providing the ability to view two or more sources of information simultaneously. The EVEREST displays are controlled by both a dedicated Linux cluster and by "fat nodes," large-memory nodes allowing the display of information from commodity hardware and software. The diversity of display and control systems allows for a wide array of uses, from interactive and deep exploration of scientific data sets to engaging scientific communication to the public.

8.2.2 Process of Science

Numerous scientific activities and collaborations under way at ORNL require and leverage the computing and network resources as well as the collaborative tools at the laboratory. Substantial research and development efforts exist to promote and foster this environment; representative examples are described below.

In the age of large-scale, collaborative experimental facilities, scientists are overloaded with an abundance of data that must be quickly understood and acted upon in making decisions that affect multimillion-dollar experimental equipment. In large projects such as ITER and KSTAR, such collaborative decisions must be made by a diverse group of scientists distributed around the world. Such decisions have to be made quickly in order to adjust the ongoing experiment or to control the experiment's next run. To support such a large-scale distributed collaboration, we must develop infrastructure around extreme-size data streams integrated with dynamic workflows that can help support the real-time analyses and decisions, as well as in-depth post-experiment analyses and decisions.

The DOE SC listed ITER as the number-one priority of a large-scale facility in the 2003 publication *Facilities for the Future of Science: A 20-year Outlook* (and updated in 2007). ITER is aimed at producing 10 times more energy output from a fusion reaction than the input energy. It is designed to be a long-pulse (300–500 seconds) fusion-reacting research tokamak, with the anticipated raw data production rate at ~1 TB/sec. This rate corresponds to up to 0.5 PB per pulse (also known as a shot). Without a validated advanced data-management framework in place before its operation, ITER scientists will be forced to abandon most of the valuable experimental data after a predetermined filtering of the information. In order to go back to the unsaved data, they must repeat the multimillion-dollar shot.

A real-time data analysis accompanied with real-time experimental control could maximize the multimillion-dollar experimental pulse in a long-pulse experimental study. Furthermore, post-processing data analysis and physics research can be used for the next experimental pulse. For this purpose, large amounts of streaming data will have to be distributed quickly and analyzed efficiently by remote scientists.

In the full operation mode, the data production rate from KSTAR is anticipated to be well over 10 Gbps. Eventually, a 300-second shot will produce over 3 TB of data to be studied by scientists in the United States and other parts of the world. Existing data-

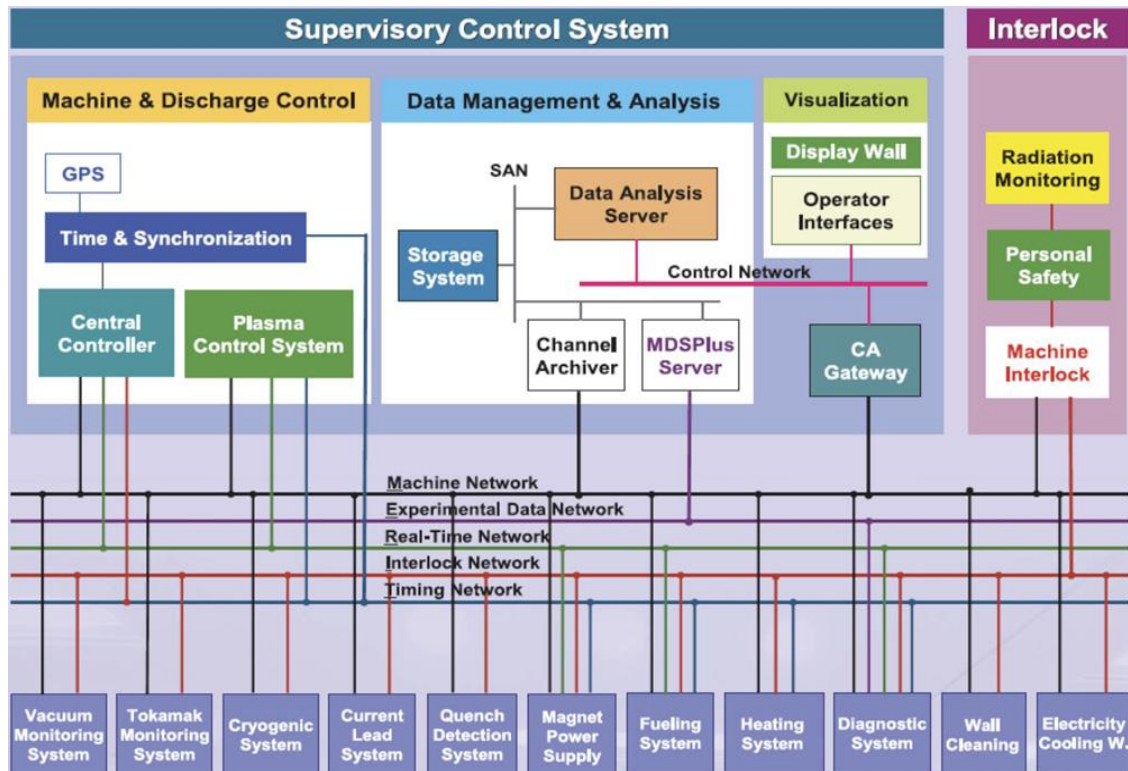


Figure 16. KSTAR control system schematics (Adapted from [OYK+2008]).

analysis frameworks are not expected to handle this data rate gracefully. We anticipate that as the length of experiments grows, the amount of collaboration increases, so that users can explore new real-time analysis in order to control the current and next experiment.

During the time we mine data from simulations, such as those in the Edge Physics Simulation (EPSI) SciDAC project, we will need to transfer subsets of data from these calculations to remote resources. Typical simulation data today is about 100 TB of output in a large EPSI simulation, and we anticipate this to grow an order of magnitude in the next two years. Thus, we assume that we will need to move roughly 10 TB over the lifetime of a KSTAR simulation (100 seconds). We will further reduce the burden of the data movement by applying query-based data movement, and will look at using “dynamic-workflow-automation” as part of the ASCR International Collaboration Framework for Extreme Scale Experiments (ICEE) collaboration project. We anticipate that this will reduce the data requirement by at least one more orders of magnitude. We also research new techniques for collaboration that can be integrated into an advanced framework through which the distributed scientists can efficiently study large volumes of scientific data.

Earth System Grid Federation

The Earth System Grid Federation (ESGF) is operating production and next-generation testbed nodes on the OLCF network¹. The OLCF ESGF resources consist of four servers and approximately 150 TB of disk storage for the production systems, plus several more servers and virtual machines (VMs) for testbeds and development. The production servers connect to the HPSS system through a combination of off-the-shelf and custom software.

The OLCF ESGF servers currently house two major data sets: a partial copy of the Community Climate System Model, Version 3 (CCSM3), data and output from the Ultra-High Resolution Global Climate Simulation project. The OLCF's portion of CCSM3 data is approximately 60 TB and there is currently approximately 100 TB of published data from the Ultra-High Res project. The total CCSM3 volume at the OLCF is expected to grow to something in excess of 250 TB. All this data is stored on the HPSS system. There are a few other smaller data sets stored on the disk array, but their combined size is negligible compared with the first two data sets.

Demand for the CCSM3 data is increasing. Scientists accessing this data are downloading 10-100 GB at a time.

Once the Ultra-High Res project is completed, the expectation is that scientists may request approximately 10 TB of data at a time and would like to be able to download that amount of data in 24 hours, or roughly 120 MB/sec. OLCF anticipates no more than one or two of these 10 TB requests per day.

Accelerating Data Acquisition, Reduction, and Analysis (ADARA) for Neutron Science

ADARA is a collaboration between the Laboratory's Computing and Computational Sciences Directorate (CCSD) and the Neutron Sciences Directorate (NScD) to enable instant feedback from experiments conducted at the Spallation Neutron Source (SNS) and to provide a data backplane to enable new capabilities in the integration of simulation and experiment. The project team has developed (1) the software to stream data from SNS beamlines over a high-speed network (Stream Management Service [SMS]), (2) the software that can subscribe to this data stream and create the event-NeXus files on the fly (Streaming Translation Service [STS]), and (3) the software within Mantid so that it can subscribe to the data stream and perform the reduction of the data live (Streaming Reduction Service [SRS]). Live data reduction provides users with instant access to Neutron data in energy/momentum space. With ADARA, data can be streamed both to the local data-analysis workstation on the beamline and to computational resources co-located with ORNL's high-performance computing

¹ As was mentioned in the 2009 ASCR Science Network Requirements report, the Earth System Grid servers have been moved off the TeraGrid and onto the OLCF network.

resources in the National Center for Computational Sciences (NCCS), the Titan platform at the OLCF, or even to remote facilities such as NERSC or ALCF. The ADARA project team is capturing and reducing data using this system on the HYSPEC beamline² and has developed plans for deployment across all SNS beamlines.

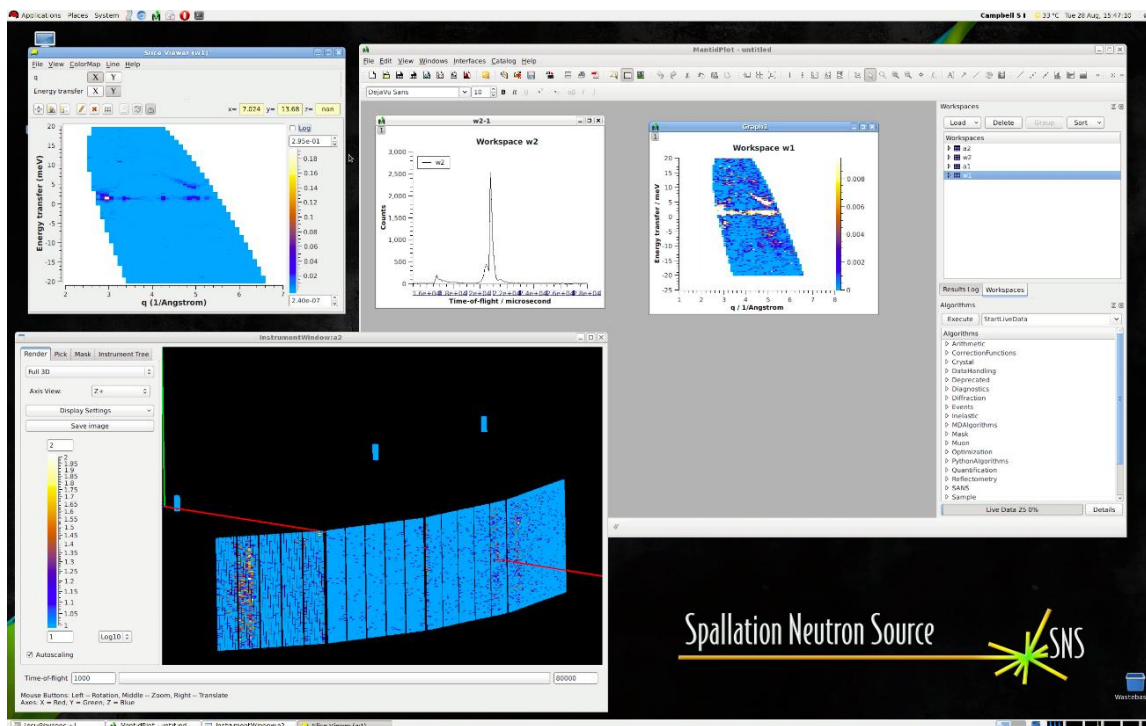


Figure 17. Live event processing with ADARA and Mantid on HYSPEC beamline at SNS with cuprite sample.

Integration of HPC

In addition to work on live streaming analysis, the ADARA team has begun integration of HPC into the data-analysis chain for SNS. The two primary components of this work are the use of parallel computation to enable high-performance data reduction (conversion from neutron-even time of flight and position to energy/momentum space) and the use of models and simulation to allow real-time fitting of experiment data to models. Integration of modeling and simulation increasingly relies on mid- to large-scale HPC resources due to the high resolution available at SNS.

Although the data file size from some SNS beamlines (e.g., Backscattering Silicon Spectrometer [BASIS] and Hybrid Spectrometer [HYSPEC]) is only a few hundreds of megabytes, those from other beamlines (e.g., Nanoscale-Ordered Materials Diffractometer [NOMAD]) are tens of gigabytes, up to 1–2 terabytes. Thus, to perform fast post-acquisition processing of the large data files, it is necessary to be able to read the data in parallel and process it in parallel using HPC systems. For data reduction we

² <http://neutrons.ornl.gov/hyspec/>

are engaged in the development of the software for Mantid to exploit HPC clusters. This software allows users to execute parallel reduction operators on remote HPC clusters and display the results of this within the local Mantid workspace. Leveraging these parallel reduction operators has allowed large-scale data sets such as 120 GB NOMAD data sets to be reduced in less than five minutes, as illustrated. In addition to these efforts, we have recently begun work on the integration of simulation and experiment at SNS through an FY 2013 BES Seed award. Furthermore, software to perform statistical analysis on experimental and model/simulation data needs to be constructed and put in place. Simulation codes need to be modified so that their output can be directly compared to neutron scattering data.

As our efforts in integration of simulation and experiment continue, we will require further advances in WAN capabilities to enable near-real-time feedback from experiments at the SNS coupled with data analysis that may be conducted at different computational facilities from the OLCF, ALCF, and NERSC. The ability to co-schedule large-scale compute resources in tandem with beam time at SNS, coupled with high-performance networking between these sites will be required to leverage these remote computational resources for near-real-time feedback from simulation to better inform experiments at the SNS. This usage pattern (interactive use of remote computational facilities) will necessitate significantly higher bandwidth capabilities during the experiment when compared with steady-state usage. During these experiments, users will require feedback in interactive time scales of seconds rather than minutes or hours. From a networking perspective, this will necessitate extremely high-performance network connectivity for “bursty” workloads and perhaps advances in quality of service guarantees to ensure that a significant percentage of the networking resource is available to support interactive requirements. For higher-resolution / high-event-rate beamlines, this may necessitate the ability to move multi-terabyte data sets in seconds.

Visualization Involving Remote Users

Interactive remote visualization is a very common method employed by OLCF users to explore their simulation data sets. Scientists involved in such diverse fields as astrophysics, climate, and molecular biology use this method to interactively explore large data sets generated on Jaguar and stored on the Spider file system. Remote visualization, though frequently used, does not place high demands on the WAN. As an example, a typical 1,000 x 1,000-pixel image, generated and sent over the network to a remote user at 5 frames per second, will only require 0.12 Gbps. If 10 users were simultaneously performing remote visualization using this method, the data rate would hit ~1.2 Gbps. This data rate can also be mitigated by deployment of multiresolution and video-compression technologies.

Analysis and visualization of simulation data at remote sites is a common use model for simulation scientists. Use cases could include the transfer of images and graphs of a running simulation (using in situ visualization, for example) to a dashboard to be shared among members of a research team. It could also include much more data-intensive

operations, such as the transfer of much larger sets of processed data for further analysis and visualization. Assuming that in situ processing of data results in a reduction of two orders of magnitude, a simulation using all of the memory on Jaguar would produce 6 TB of processed data per time step. If a simulation took 10 minutes to compute a time step, this would be equivalent to a data rate of 10 GB/sec or 80 Gbps.

8.3 Science Drivers — the Next 2-5 Years

8.3.1 Instruments and Facilities

8.3.2 Process of Science

8.4 Beyond 5 Years — Future Needs and Scientific Direction

One of the drivers for large data movement over the WAN will be in data generated from the OLCF. An example use case involves Jacqueline Chen and other research staff of Sandia National Laboratories. Eighty-three percent of the energy in the United States comes from combustion of fossil fuels; the design of more efficient engines has the potential to ensure a viable and secure energy future. Chen explores using rapidly evolving fuel streams such as ethanol, butanol, and biodiesel in a new generation of highly efficient, low-emission combustion systems. Chen's simulations provide detailed characterization of flow and combustion, which can be coupled with experimental analyses.

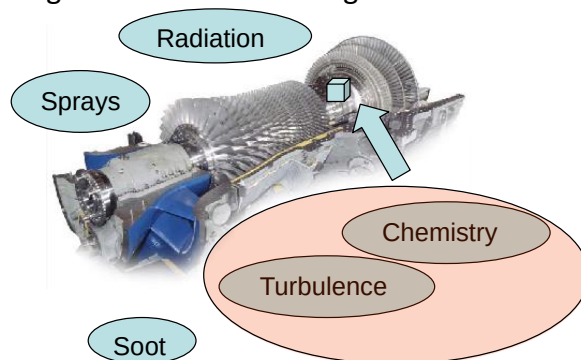


Figure 18.

In the next three years, the combustion simulations will work with 2 billion to 7 billion grid points on a uniform mesh, for a Reynolds number of about 300–1000. The simulation typically looks at 22 transported species (ethylene/air chemistry) along with five variables related to density, the momentum vector, and total energy. Each petascale simulation contains about 500,000 time steps, which compute over one week of wall clock time. Each file produced is about 1 TB per time step, and is output every 0.2 microseconds of the 1 millisecond (in real time) simulation.

In the exascale future, we are looking to work with pressure going from 1 atmosphere for the petascale simulation to 30-60 atmospheres for the exascale, with a Reynolds number going to 4,300. This means that our state size will contain 10^{12} grid points (1 PB of data), and run one wall clock week, generating about 400 PB of data if dumped on the file system.

We are transitioning the simulations to work with more “hybrid-staging,” where many of the analytics are placed inside the simulation workflow. This has the benefit of

allowing data operations to reduce the total amount of information that goes into the storage system. We are also investing in approaches that can allow much of this reduced data, from topological features, images from parallel volume visualization, etc., to be streamed over the WAN to investigators' desktops, clusters for future analysis. We envision that this will reduce the storage, and thus WAN requirements, by at least two orders of magnitude. Thus, we believe that in the next eight years, our requirements over the WAN will be about 400 TB of data to be moved in a week.

Another transition we are starting in our project is the ability to run exploratory analytics during simulations. There are many research ideas that try to reduce the overall amount of data to be moved during these short time periods (30 minutes between time dumps), but we believe that the minimum amount of data that can be used is $1/1000^{\text{th}}$ of the total data size, or 1 TB of data every 30 minutes.

8.5 Network and Data Architecture

ORNL maintains a wide area optical network carrying circuits into the Laboratory from Atlanta, Chattanooga, Nashville, Knoxville, and Chicago. The original parts of this infrastructure, the paths from ORNL to Nashville and Atlanta to Chicago, were constructed in 2004 using Ciena CoreStream optical equipment. Although continued support is expected for the Ciena CoreStream equipment for several more years, the maximum circuit capability is 10 Gbps, and it is now considered legacy equipment. The path from ORNL through Knoxville to Chattanooga was constructed in 2011 using the newer Ciena 6500 equipment capable of 100 Gbps circuits. The ESnet circuits from ORNL to Nashville are carried on the legacy infrastructure to the CenturyLink Metroplex site. These circuits are extended to the ESnet hub located in the Level3 facility at Sidco Drive through a trade agreement with the University of Alabama System (UAS). Circuits on the ORNL optical infrastructure are exchanged for circuits on the UAS infrastructure between the CenturyLink Metroplex and the Level3 Sidco Drive site.

ORNL is faced with two long-term risks. First, the UA agreement is a three-year agreement. Furthermore, this infrastructure is an older system supporting only 10 Gbps circuits. This presents a risk for the ORNL connectivity to ESnet if the UAS agreement is terminated or higher bandwidth circuits are needed. ORNL is considering options to establish a presence at Sidco Drive and acquire an ORNL-owned dark fiber cross-connect.

The second long-term risk is maintenance support for the legacy Ciena CoreStream equipment that provides ESnet connectivity from the Nashville hub to the Oak Ridge hub. Upgrade options are being considered.

8.6 Collaboration tools

The OLCF encourages greater adoption of telepresence tools for activities such as user council meetings, regular calls with the DOE program office, remote-review activities, etc. We currently have an instance of the Cisco telepresence, as well as H.323-compliant

videoconferencing capability. We have found these telepresence tools ample for scientific collaboration purposes.

However, gaps remain in our ability to support collaboration in other ways. We find near-universal needs concerning revision control for scientific codes maintained by geographically dispersed teams. We have instantiated local Subversion repositories to support this need, but authentication methods remain difficult to resolve for all users. In addition, scientific teams continue to need easy-to-use (and install) tools for collaborative and in situ visualization and analytics. The eSimon³ and Bellerophon⁴ projects are two attempts by OLCF staff to fulfill these requirements.

The OLCF is collaborating with major facilities, centers, and projects across all SC offices. These collaborations would substantially benefit from a more uniform approach to security and data policies. As an example, a shared trust environment in which user credentials from remote facilities could be honored by other facilities across SC would be of significant benefit. To realize this, these facilities would need to adopt compatible security policies wherever possible. Sharing of data that is federated across multiple DOE sites would also be of significant benefit to our researchers but would require more uniform and compatible data policies to be adopted. Technologies to enable sharing of credentials and of data is not the challenge — the real challenge is rooted in differing and conflicting policies that prevent more effective collaboration.

8.7 Data, Workflow, Middleware Tools, and Services

The trends in data analysis from 2012-2017 will continue into the exascale era, showing both a dramatic increase in the amount of data generated, but also a relative decrease in the amount of off-box I/O possible. This will necessitate a large movement toward in situ data analysis and visualization. Libraries for general data analysis, plugged into HPC simulations, will be developed in the next five years. These libraries will have to be expanded to include many domain-specific analysis capabilities to continue to be relevant in the exascale era. Since I/O will become a very constrained resource, it's expected that the adoption of I/O middleware libraries will significantly accelerate over the next five years, with the majority of HPC applications using some middleware system after this point. Scientific workflow tools are likely to rise somewhat in prominence over the next five years, especially for scientific analyses that involve parameter sweeps or manage ensemble runs. Workflows that extend beyond the

³ R. Tchoua, S. Klasky, N. Podhorszki, B. Grimm, A. Khan, E. Santos, C. Silva, P. Moullem, M. Vouk: "Collaborative Monitoring and Analysis for Simulation Scientists." 2010 International Symposium on Collaborative Technologies and Systems, (CTS 2010), Chicago, Illinois, USA, May 2010.

⁴ E.J. Lingerfelt, O.E.B. Messer, J.A. Osborne, R.D. Budiardja, A. Mezzacappa. *A Multitier System for the Verification, Visualization and Management of CHIMERA*. *Procedia Computer Science* 4, 2076–2085 (2011).

individual computer system are also likely to see greater adoption, including management of archival storage, off-site data transfer, exploitation of multiple computing systems, and distribution to Web-based portals. Finally, systems for data-analysis services may become more diverse, stretching beyond the traditional cluster-based MPI computing for analysis and visualization, and reaching into graph analytics and MapReduce types of computing architectures.

We also envision much of the I/O middleware, such as the Adaptable I/O System (ADIOS)⁵ becoming increasingly used in data streaming and file exchanges due to its flexibility in abstracting data movement. The KSTAR collaboration is a good example of this, where we are working with ESnet to create a new method inside ADIOS to move data streams via RDMA transport and establish a real-time feedback loop into ADIOS staging to allow for more interactive data exploration. We envision that this technology will also be used inside the Fusion SciDAC, EPSI, and simulations, as well as the S3D combustion code. These abstractions will make it transparent to a user of data transfer, which we believe will place much greater demands on the network.

8.8 Outstanding Issues

None.

⁵ J. Lofstead, Z. Fang, S. Klasky, K. Schwan. "Adaptable, metadata rich IO methods for portable high performance IO." Parallel & Distributed Processing, 2009. IPDPS 2009. IEEE International Symposium on , vol., no., pp.1-10, 23-29 May 2009.

8.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	Process of Science	Data Set Size	LAN Transfer Time Needed	WAN Transfer Time Needed
Near Term (0-2 years)				
• OLCF	• INCITE and ALCC awards from national laboratories, universities, corporations and international partners. In addition, partnerships with the DOE experimental facilities will increase.	• The largest simulations are generating about 200 TB of data/week. The data sets are typically broken into a series of time steps, and subfiles are commonly moved, so many files/time steps are produced.	• Typically data steps are 2 TB, and for interactive analysis; users need to digest the data in O(60) seconds.	• Users can move subsets of data generated. Typically this data reduction can reduce data 1/100 of the original data. This implies that 20 TB/10 days of data needs to be moved.
2-5 years				
• OLCF	• Possible increase in sharing of data sets in the Earth Science communities, along with many of the DOE experimental sciences	• New multi-coupled physics runs, from combustion, fusion, astrophysics, and others will potentially increase data 10X from today (20 PB/week)	• Largest simulations will need 20 PB/week of data to be interactively moved and analyzed	• Rough estimates on the order of 2 PB/10 days of data to be moved
5+ years				
• OLCF++	• With the common use of staged-I/O, we predict that exascale computing will reduce data requirements, but the velocity will increase. Validation will play a more important role, so data ingestion will become more critical.	• Less data is stored, and more data is streamed remotely	• For staged analytics and visualization, LAN will be stressed. We anticipate a need for networks capable of 1 TB/sec.	• Combining staged visualization and analytics will be used with auxiliary clusters. Data movement on the order of 20 PB/10 days.

9 Appendix A – The ESnet OSCARS Service

9.1 Background

Due to its increasingly collaborative nature, large-scale scientific research has become more and more dependent on advanced research and education (R&E) networks. This is because unique, geographically dispersed scientific instruments and facilities—like the Large Hadron Collider in Switzerland or NOAA’s Geophysical Fluid Dynamics Laboratory in Princeton, New Jersey—need to be accessed and used remotely by thousands of researchers worldwide. Furthermore, these facilities create and/or maintain massive data sets—which for some experiments can reach hundreds of terabytes to tens of petabytes—that have to be archived, catalogued, and analyzed. In many cases these experiments also depend on large-scale computing resources.

The ability to reliably manage the round-the-clock data flows from these supercomputers and experimental facilities are essential for scientists to conduct research. Traditional shared IP networks are unable to elegantly handle the large and sustained bursts of massive data that these experiments produce without disrupting other traffic on the network. These “best effort” IP networks cannot assure the scientists a consistently high quality network path from end to end, researcher to researcher, site to site. This unpredictability can obstruct new scientific discovery.

To ensure that scientists, no matter where they are located, can meet the time-critical needs of their research, ESnet designed and developed the On-demand Secure Circuits and Advance Reservation System (OSCARS). This open source software application allows scientists as well as network operators to create and reserve virtual circuits with guaranteed end-to-end performance. These circuits are tuned for moving and exchanging massive data sets between collaboration sites and can do so across multiple network domains—especially important in the R&E community, where multiple national, regional, and local networks are involved in connecting collaborators. In doing so, OSCARS gives users the ability to engineer, manage, and automate the network based on the specific needs of their work with scientific instruments, computation, and collaborations.

Through an easy-to-use web interface, the user can define the virtual circuit’s end points, specify the amount of bandwidth needed, and the time and date when the bandwidth is required. OSCARS also allows scientists to automate the process of setting up temporary virtual circuits by having their workflow management systems “talk” to it directly through an easily programmed API. This is particularly useful as the workflow of a project may involve dozens of stages and components, all of which must happen at a

certain time to achieve the right science output. OSCARS circuits enable effective process planning of workflow, ensuring that each activity happens at the right time to ensure the needed scientific result. Together, the user interface and API automatically reduce to a few minutes the transactions that previously took weeks or months or a multitude of phone calls and e-mails to accomplish manually.

Built using R&E community-developed protocols like the Interdomain Controller Protocol (IDC), OSCARS is interoperable with most other R&E virtual circuit services and the equipment used to provide those services. To ensure wide deployment, OSCARS had to leverage the capabilities of various vendor devices while remaining independent of them. To accomplish this, OSCARS builds a very flexible Path Setup System that allows easy development of capabilities to control multi-vendor boxes. This has allowed OSCARS-enabled networks to offer consistent network services across a wide range of equipment from vendors including Juniper, Cisco, Ciena, Brocade, Adva, and Force 10 Networks, among others.

OSCARS received a 2011 University of California Larry L. Sautter Award honorable mention and a 2011 Internet2 IDEA award.

9.2 OSCARS Innovations and Futures

During 2012 a lot of production improvements, research collaborations, and cutting-edge demonstrations were done leveraging the OSCARS open-source code base. The following three highlight some of the impactful activities:

1. Version 0.6 Production Software

ESnet is continually evolving OSCARS to meet the needs of the community. In spring 2011, ESnet launched OSCARS 0.6A, making it available on its website for downloading and testing by users. OSCARS 0.6B was released in mid-summer 2011, followed by OSCARS 0.6RC1 (Release Candidate 1) in Oct 2011 in time for SC11. Since 2011, version 0.6 now has been adopted and deployed by **18 networks** in production including ESnet, Internet2, KREONET and Universities part of the DYNES project, etc. Version 0.6 of OSCARS offers enhanced inter-domain error reporting as well as a more flexible software architecture, including increased modularity and exposed internal interfaces so that the community can standardize IDC components, code development, and collaborations. This includes a flexible path computation engine (PCE) framework, which allows atomic PCEs to be executed in any configurable arbitrary sequence, as well as the flexibility to streamline the path computation process. OSCARS' PCE framework has been available for download as a software development kit (SDK), since November 2010, supporting researchers creating complex

algorithms for path computation. Multiple DOE-funded research projects have downloaded this SDK and used it for research and development.

2. Network Services Interface demonstration

Given the success of Inter-domain Control Protocol (IDCP) developed by a few Research and Education Networks (ESnet, Internet2, GEANT and Canarie), Open Grid Forum took on the standardization of a generalized inter-domain network services negotiation called Network Services Interface (NSI) working group. OSCARS developers actively participated in development of the NSI protocol standard, and participated in an interoperability demonstration of connection services protocol as part of the GLIF Automated GOLE testbed in SuperComputing 2012. This demonstration showcased the interoperability of NSI-CS v2.0 standard among various implementations including G-Lambda, OpenNSA, and OpenDRAC. The NSI Connection Services standard is targeted to support multi-domain advance reservations and connection setup.

The OSCARS project welcomes wider community participation in developing a reference implementation of the NSI standard. As part of this open initiative, ESnet started a pragmatic collaboration with SURFnet (The R&E network in Netherlands) to implement a NSI Aggregation function that will enable ESnet's implementation of NSI to create tree-style⁶ reservations. This collaboration will allow ESnet and SURFnet to pool their software-development resources to implement a production quality NSI Aggregator for the broader R&E community.

3. Worlds-first OpenFlow Control of Layer 1 switches

In late November, OSCARS was extended to become an OpenFlow controller and enhanced to demonstrate a prototype SDN Open Transport Switch (OTS), implemented by Infinera on their optical gear, capable of dynamically controlling bandwidth services at the optical layer. The proof-of-concept demo was conducted on ESnet's testbed network ring in New York, connecting Brookhaven National Laboratory with a network hub in Manhattan.

The development team is also working to extend OSCARS' capabilities to provide substantially more complex services such as overlay networks, and multi-layer circuits for even greater operational and cost efficiency.

⁶ Tree-Style reservations consist of a centralized-style of creating end-to-end circuits, with one domain contacting other intermediate domains to reserve portions of the circuit and then, stitching the various connections into a usable end-to-end circuit. So far OSCARS only supports chain-style connection setup, which is very similar to how MPLS signaling protocols like RSVP work.

9.3 OSCARS Deployments

As of December 2012, OSCARS is currently deployed in 45⁷ networks worldwide, ranging from wide-area backbones (such as ESnet, Internet2, USLHCNet, NORDUnet, and RNP) to regional networks (such as GpENI, MAX, and LONI), exchange points (such as SoX, and AMPATH), local-area networks (such as University of Delaware, Vanderbilt, CALTECH, Rutgers, and UMich), and research testbeds (such as JGN2 and KOREN). Through these deployments, OSCARS circuits on ESnet and in other domains are supporting a wide variety of science disciplines including Genomics, LHC, Astrophysics and Climate research.

In addition, the DOE currently funds three projects that have actively developed OSCARS v0.6 path computation engine (PCE) modules:

- University of Southern California/Information Sciences Institute
- University of New Mexico, and ESnet: multi-layer, multi-technology PCE
- UC Davis: protection/resiliency PCE
- University of Massachusetts and Dartmouth College: anycast/manycast/multicast PCE.

⁷ *OSCARS Deployments for production, research or testbed:* ESnet, AMPATH, Boston University, California Institute of Technology, Corporation for Education Network Initiatives in California (CENIC), Florida Institute of Technology, Florida International University, Front Range GigaPop, Great Plains Environment for Network Innovation (GpENI), Harvard University, Indiana University, Internet2, Japan Gigabit Network II (JGN-II), John Hopkins University, Korea Advance Research Network (KOREN), Korea Research Environment Open Network (KREONET), Lonestar Education and Research Network (LEARN), Louisiana Optical Network Initiative (LONI), Mid-Atlantic Gigapop in Philadelphia for Internet2 (MAGPI), Mid-Atlantic Crossroads, NORDUnet, Northrop Grumman, Oklahoma University, Renaissance Computing Institute (RENCI), Rice University, Rede Nacional de Ensino e Pesquisa (RNP), Rutgers University, Southern Crossroads, TransPAC3, Tufts University, UC San Diego, UC, Santa Cruz, University of Colorado Boulder, University of Delaware, University of Florida, University of Houston, University of Iowa, University of Massachusetts Dartmouth, University of Michigan, University of Texas at Dallas, University of Nebraska Lincoln, University of Texas Arlington, University of Wisconsin Madison, US LHCNet, Vanderbilt University

10 Appendix B – The ESnet 100G Testbed

The ESnet 100G testbed provides network researchers with a realistic environment for 100G application and middleware experiments. The testbed consists of a dedicated 100G wave from Oakland, CA to Chicago, IL, and a number of high-speed hosts at each end, the combination of which can easily saturate a 100G link.

The ESnet 100G testbed provides a rapidly reconfigurable high-performance network research environment that enables researchers to accelerate the development and deployment of 100G networking through prototyping, testing, and validation of advanced networking concepts. For example, the testbed has been used to validate the use of RDMA techniques over a wide area network, and quantify the advantages of RDMA over TCP. The testbed also provides support for prototyping middleware and software stacks to enable the development and testing of 100G science applications. The ESnet 100G testbed eliminates the need for network researchers to obtain funding to build their own network testbeds or use an artificial laboratory environment to prove the viability of innovative ideas.

Besides being useful to network researchers, the testbed is also extremely valuable to ESnet itself, and allows ESnet to stay on the cutting edge of network technology. ESnet uses this testbed to experiment with various router configuration settings, including quality of service (QoS), NetFlow, SNMP, and so on. ESnet also uses the testbed to experiment with various "Data Transfer Node" hardware configurations for a Science DMZ. From this testing we derive a reference implementation of a DTN, which DOE sites have found to be very useful. In addition, ESnet also uses the testbed to test new versions of OSCARS.

11 Glossary

ADARA	Accelerating Data Acquisition, Reduction, and Analysis
ADIOS	Adaptable I/O System
ALCC	ASCR Leadership Computing Challenge
ALCF	Argonne Leadership Computing Facility
ALS	Advanced Light Source
ANI	Advanced Networking Initiative
ANL	Argonne National Laboratory
API	Application programming interface
ASCR	Advanced Scientific Computing Research
BAMAN	Bay Area Metropolitan Area Network
BASIS	Backscattering Silicon Spectrometer
BBCP	BaBar copy
BER	Biological and Environmental Research
BES	Basic Energy Sciences
CCSD	Computing and Computational Sciences Directorate
CCSM3	Community Climate System Model, Version 3
CRT	Computational Research and Theory
DDN	Data Direct Networks
DOE	Department of Energy
DTN	Data transfer node
DWDM	Dense wavelength division multiplexing
DMZ	Demilitarized Zone
EPSI	Edge Physics Simulation
ESGF	Earth System Grid Federation
ESnet	Energy Sciences Network
EVEREST	Exploratory Visualization Environment for REsearch in Science and Technology
GB/sec	Gigabytes per second
Gbps	Gigabits per second
GPFS	General Parallel File System
GPGPU	General-purpose graphics processing unit
GPU	Graphics processing unit
HACC	Hardware/Hybrid Accelerated Cosmology Code
HCA	Host channel adaptor
HPC	High-performance computing
HPSS	High Performance Storage System
HYSPEC	Hybrid Spectrometer
IB	InfiniBand
ICEE	International Collaboration Framework for Extreme Scale Experiments
INCITE	Innovative and Novel Computational Impact on Theory and Experiment
I/O	Input/output

ION	Input/output node
JGI	Joint Genome Institute
KSTAR	Korea Superconducting Tokamak Advanced Research
LAN	Local area network
LANL	Los Alamos National Laboratory
LBNL	Lawrence Berkeley National Laboratory
LCF	Leadership Computing Facility
LCLS	Linac Coherent Light Source
LDRD	Laboratory Directed Research and Development
LHC	Large Hadron Collider
LION	Log-in input/output node
LLNL	Lawrence Livermore National Laboratory
LTO	Linear Tape-Open
MB/sec	Megabytes per second
Mbps	Megabits per second
MPI	Message Passing Interface
MDS	Metadata server
MDT	Microsoft Deployment Toolkit
NASA	National Aeronautics and Space Administration
NCCS	National Center for Computational Sciences
NERSC	National Energy Research Scientific Computing Center
NGF	NERSC Global File System
NOMAD	Nanoscale-Ordered Materials Diffractometer
NScD	Neutron Sciences Directorate
NSF	National Science Foundation
OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
OSCARS	On-Demand Secure Circuits and Advance Reservation System
OSF	Oakland Scientific Facility
PB/sec	Petabytes per second
Pbps	Petabits per second
perfSONAR	PERformance Service Oriented Network monitoring Architecture
PF	petaflop
PPPL	Princeton Plasma Physics Laboratory
PTF	Palomar Transient Factory
PVFS	Parallel Virtual File System
QDR	Quad data rate
RDMA	Remote direct memory access
REST	Representational State Transfer
RPI	Rensselaer Polytechnic Institute
SAN	Storage area network
SC	DOE Office of Science
SCEC	Southern California Earthquake Center

SDN	Science Data Network
SMS	Stream Management Service
SNS	Spallation Neutron Source
STS	Streaming Translation Service
TACC	Texas Advanced Computing Center
TB/sec	Terabytes per second
Tbps	Terabits per second
TF	Teraflop
UAS	University of Alabama System
VM	Virtual machine
WAN	Wide area network

12 Acknowledgements

This work would not have been possible without the contributions and participation of those who provided information and attended the review. ESnet would also like to thank the ASCR program office for its help in organizing the review and providing insight into the facilities supported by the ASCR program.

Mira cover image courtesy of the Argonne Leadership Computing Facility (ALCF), Argonne National Laboratory.

Hopper image courtesy Oakland Scientific Facility (OSF), Lawrence Berkeley National Laboratory.

Titan cover image courtesy of the Oak Ridge Leadership Computing Facility (OLCF), Oak Ridge National Laboratory.

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR). Vince Dattoria is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract

DE-AC02-05CH11231.

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

This is LBNL report LBNL-XXXX