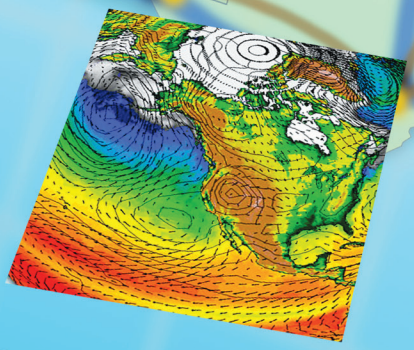
ASCR Science Network Requirements

ESnet
Energy Sciences Network

Office of Advanced Scientific Computing Research
Network Requirements Workshop
Conducted April 15 and 16, 2009

Final Report



Berkeley Lab

U.S. Department of Energy
Office of Science

ASCR Network Requirements Workshop

Office of Advanced Scientific Computing Research, DOE Office of Science
Energy Sciences Network
Gaithersburg, MD — April 15 and 16, 2009

ESnet is funded by the US 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 US 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.

This is LBNL report LBNL-2495E

Participants and Contributors

Bill Allcock, ANL (ALCF, GridFTP)

Rich Carlson, Internet2 (Networking)

Steve Cotter, ESnet (Networking)

Eli Dart, ESnet (Networking)

Vince Dattoria, DOE/SC/ASCR (ASCR Program Office)

Brent Draney, NERSC (Networking and Security)

Richard Gerber, NERSC (User Services)

Mike Helm, ESnet (DOEGrids/PKI)

Jason Hick, NERSC (Storage)

Susan Hicks, ORNL (Networking)

Scott Klasky, ORNL (OLCF Applications)

Miron Livny, University of Wisconsin Madison (OSG)

Barney Maccabe, ORNL (Computer Science)

Colin Morgan, NOAA (Networking)

Sue Morss, DOE/SC/ASCR (ASCR Program Office)

Lucy Nowell, DOE/SC/ASCR (SciDAC)

Don Petravick, FNAL (HEP Program Office)

Jim Rogers, ORNL (OLCF)

Yukiko Sekine, DOE/SC/ASCR (NERSC Program Manager)

Alex Sim, LBNL (Storage Middleware)

Brian Tierney, ESnet (Networking)

Susan Turnbull, DOE/SC/ASCR (Collaboratories/Middleware)

Dean Williams, LLNL (ESG/Climate)

Linda Winkler, ANL (Networking)

Frank Wuerthwein, UC San Diego (OSG)

Editors

Eli Dart, ESnet — dart@es.net
Brian Tierney, ESnet — bltierney@es.net

Contents

1 Executive Summary

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the US Department of Energy Office of Science, the single largest supporter of basic research in the physical sciences in the United States. In support of the Office of Science 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 20 years.

In April 2009 ESnet and the Office of Advanced Scientific Computing Research (ASCR), of the DOE Office of Science, organized a workshop to characterize the networking requirements of the programs funded by ASCR.

The ASCR facilities anticipate significant increases in wide area bandwidth utilization, driven largely by the increased capabilities of computational resources and the wide scope of collaboration that is a hallmark of modern science. Many scientists move data sets between facilities for analysis, and in some cases (for example the Earth System Grid and the Open Science Grid), data distribution is an essential component of the use of ASCR facilities by scientists. Due to the projected growth in wide area data transfer needs, the ASCR supercomputer centers all expect to deploy and use 100 Gigabit per second networking technology for wide area connectivity as soon as that deployment is financially feasible.

In addition to the network connectivity that ESnet provides, the ESnet Collaboration Services (ECS) are critical to several science communities. ESnet identity and trust services, such as the DOEGrids certificate authority, are widely used both by the supercomputer centers and by collaborations such as Open Science Grid (OSG) and the Earth System Grid (ESG).

Ease of use is a key determinant of the scientific utility of network-based services. Therefore, a key enabling aspect for scientists' beneficial use of high performance networks is a consistent, widely deployed, well-maintained toolset that is optimized for wide area, high-speed data transfer (e.g. GridFTP) that allows scientists to easily utilize the services and capabilities that the network provides. Network test and measurement is an important part of ensuring that these tools and network services are functioning correctly. One example of a tool in this area is the recently developed perfSONAR, which has already shown its usefulness in fault diagnosis during the recent deployment of high-performance data movers at NERSC and ORNL. On the other hand, it is clear that there is significant work to be done in the area of authentication and access control — there are currently compatibility problems and differing requirements between the authentication systems in use at different facilities, and the policies and mechanisms in use at different facilities are sometimes in conflict.

Finally, long-term software maintenance was of concern for many attendees. Scientists rely heavily on a large deployed base of software that does not have secure programmatic funding. Software packages for which this is true include data transfer tools such as GridFTP as well as identity management and other software infrastructure that forms a critical part of the Open Science Grid and the Earth System Grid.

2 Workshop Background and Structure

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

1. Work with the SC community to identify the networking implication of the instruments, supercomputers, and the evolving process of how science is done
2. Develop an approach to building a network environment that will enable the distributed aspects of SC science and then continuously reassess and update the approach as new requirements become clear
3. Keep anticipating future network capabilities that will meet future science requirements with an active program of R&D and Advanced Development

Addressing point (1), the requirements of the Office of Science science programs are determined by

A) Exploring the plans and processes of the major stakeholders:

- Data characteristics of scientific instruments and facilities — what data will be generated by instruments and supercomputers coming on-line over the next 5-10 years?
- Examining the future process of science — how and where will the new data be analyzed and used — that is, how will the process of doing science change over 5-10 years?

B) Observing current and historical network traffic patterns

- What do the trends in network patterns predict for future network needs?

The primary mechanism of accomplishing (A) is the Office of Science (SC) Network Requirements Workshops, which are sponsored by ASCR and organized by the SC Program Offices. SC conducts two requirements workshops per year, in a cycle that will repeat starting in 2010:

- Basic Energy Sciences (materials sciences, chemistry, geosciences) (workshop in 2007, report published)
- Biological and Environmental Research (2007 — published)
- Fusion Energy Science (2008 — published)
- Nuclear Physics (2008 — published)
- IPCC (Intergovernmental Panel on Climate Change) special requirements (BER) (August 2008)
- Advanced Scientific Computing Research (Spring 2009)
- High Energy Physics (Summer 2009)

The workshop reports are published at http://www.es.net/hypertext/requirements.html.

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

In April 2009 ESnet and the Office of Advanced Scientific Computing Research (ASCR), of the DOE Office of Science, organized a workshop to characterize the networking requirements of the science programs funded by ASCR. The most network demanding ASCR facilities or programs include the Argonne Leadership Computing Facility (ALCF), the Oak Ridge Leadership Computing Facility (OLCF), the National Energy Research Scientific Computing Center (NERSC), the Earth System Grid (ESG), and the Open Science Grid (OSG).

Workshop participants were asked to codify their requirements in a case study format that included a network-centric narrative describing the science, the instruments and facilities currently used or anticipated for future programs, the network services needed, and the way in which the network is used. Participants were asked to consider three time scales in their case studies — the near term (immediately and up to 12 months 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 time scale and columns for network bandwidth and services requirements. The case study documents are included in this report.

3 Office of Advanced Scientific Computing Research (ASCR)

3.1 Introduction

The mission of the ASCR is to discover, develop, and deploy the computational and networking capabilities that enable researchers to analyze, model, simulate, and predict complex phenomena important to the Department of Energy. In the past two decades, leadership in scientific computation has become a cornerstone of the Department's strategy to ensure the security of the nation and succeed in its science, energy, environmental quality, and national security missions. A particular challenge of this program is fulfilling the science potential of emerging multi-core computing systems and other novel "extreme-scale" computing architectures which will require significant modifications to today's tools and techniques. ASCR supports DOE's mission with world-class research capacity:

- To develop mathematical descriptions, models, methods and algorithms to enables scientists to accurately describe and understand the behavior of complex systems involving processes that span vastly different time and/or length scales.
- To develop the underlying understanding and software to make effective use of computers at extreme scales
- To transform extreme scale data from experiments and simulations into scientific insight.
- To advance key areas of computational science and discovery that advance the missions of the Office of Science through mutually beneficial partnerships.
- To deliver the forefront computational and networking capabilities, enabling world-class researchers to extend the frontiers of science.
- To develop networking and collaboration tools and facilities that enable scientists worldwide to work together.

These challenges require teams of scientists distributed across the country, as well as the full national portfolio of experimental and computational tools. ASCR has a leading role in the development of the networks needed to remove geography as a barrier, including advancing US participation in international collaborations. ASCR-supported high-performance networks and networking research enables scientists to move millions of gigabytes generated by large-scale scientific instruments and supercomputers.

ASCR's Research Division supports world-class research in applied mathematics, computational and computer science, and next generation networking for science. ASCR's Facilities Division manages three of the world's top supercomputer centers and the Energy Sciences Network (ESnet). It also supports an inter-agency research program investigating next generation, high-performance computing architectures. Of particular interest to ESnet are the supercomputing facilities and the Next Generation Research for Science Program in the Research Division.

3.2 High Performance Computing and Network Facilities

3.2.1 Supercomputer Facilities

There are two types of ASCR supercomputer facilities: High Performance Computing (HPC) Facility and Leadership-Class Computing Facility (LCF).

HPC — As a national resource to enable scientific advances to support the missions of the Department of Energy's Office of Science, the National Energy Research Scientific Computing Center (NERSC), operated by the Lawrence Berkeley National Laboratory, annually serves approximately 3,000 scientists from DOE laboratories, universities, industrial laboratories, and other Federal agencies throughout the United States. Computational science conducted at NERSC covers the entire range of scientific disciplines, but is focused on research that supports DOE's missions and scientific goals. NERSC currently supports 400 projects, and provides 220M processor hours for Allocation Year 2009. NERSC's main computer is a Cray XT4 named Franklin with about 40,000 cores (processors). NERSC also provides large archival data storage to a number of large national and international scientific collaborations.

LCF — The LCFs provide maximum computational capability to a small number of select users who can take advantage of some of the largest computational capabilities available in the world. ASCR has LCFs at Argonne National Laboratory and at Oak Ridge National Laboratory. LCF resources are awarded to a small number of projects that address grand challenges in sciences and engineering through a peer-reviewed program, called the Innovative and Novel Computational Impact on Theory and Experiment (INCITE). INCITE encourages proposals from universities, other research institutions, and industry, U.S. and international.

The Argonne Leadership Computing Facility (ALCF) provides resources that make computationally intensive projects of the largest scales possible. The IBM Blue Gene/P system, nicknamed Intrepid, possesses a peak speed of 557 teraflops and a LINPACK speed of 450 teraflops, one of the fastest supercomputers in the world. Intrepid's configuration features 40,960 nodes, each with four cores, for a total of 163,840 cores and 80 terabytes of memory. ALCF currently supports 28 INCITE projects and about 100 small projects for over 400 users, and provides a total of 500M processor hours (400M hours for INCITE).

The Oak Ridge National Laboratory Leadership Computing Facility (OLCF) supercomputer, nicknamed Jaguar, reached a theoretical peak of 1.64 petaflops, and became the world's first petaflop system dedicated to open research in November 2008. The Cray XT5 system utilizes over 45,000 of the latest quad-core Opteron processors (a total of 180,000 cores) from AMD and features 362 terabytes of memory and a 10-petabyte filesystem. The system has 578 terabytes per second of memory bandwidth and an unprecedented input/output (I/O) bandwidth of 284 gigabytes per second to tackle the biggest bottleneck in leading-edge systems—moving data into and out of processors. OLCF currently supports 39 INCITE projects and about 100 small projects for over 500 users, and provides a total of over 1000M processor hours (552M hours for INCITE).

3.2.2 Energy Sciences Network (ESnet)

The science research supported by SC takes place at universities and at the 15 major national laboratories and facilities. Many of the laboratories have a mission to build and operate scientific instruments that are too large, too expensive, or too long-term to be reasonably situated on a university campus. ESnet serves all of these laboratories and facilities. The primary examples of such large projects are the big particle accelerators (Tevatron at Fermilab, Relativistic Heavy Ion Collider at Brookhaven, and the synchrotron light sources at Lawrence Berkeley and Argonne); three major supercomputer centers (at Lawrence Berkeley (NERSC), Oak Ridge, and Argonne; the magnetic fusion reactors at General Atomics; Princeton Plasma Physics Lab; and MIT; etc. A large part of the data from these devices is delivered (via ESnet) to university campuses for analysis. Furthermore, Fermilab is the US LHC/CMS Tier 1 data center, and Brookhaven is the US LHC/ATLAS Tier 1 data center. These are the largest of the LHC Tier 1 data centers.

The ESnet mission is to transport all of the data to and from the DOE Labs and the US and international R&E community resulting in (1) a very close connection between ESnet and the US and international R&E institutions and networks, and (2) the need to transport massive amounts of data.

Architecturally, ESnet has a high-speed core network with sites attached mostly by metropolitan area fiber networks. The current generation ESnet network consists of multiple rings and a dual core with a national footprint. One core (10 Gbps) is primarily oriented toward commodity IP traffic; and the other core (20 Gbps), growing to 50 Gbps in 2010, is a virtual circuit-oriented network designed to handle the massive data flows of large-scale instrument-based sciences, such as the LHC.

ESnet provides a full suite of network services to its user community: IPv4, IPv6, address space management, DNS, guaranteed bandwidth data transport, etc; user network testing services, performance monitoring, trouble tickets, engineering support, trouble-shooting; PKI services and video conferencing.

3.3 Next Generation Networking Research for Science Program (the "Next-Gen Program")

The Next-Gen Program addresses end-to-end high-performance, high-capacity middleware technologies and advanced networking needed to provide secure access to distributed science facilities, high-performance computing resources, and large-scale scientific collaborations.

The Next-Gen Program builds on results from ASCR's Computer Science and Applied Mathematics to develop integrated software tools and advanced network services to enable large-scale scientific collaboration and to utilize the new capabilities of ESnet to advance DOE missions. The research falls into two general categories described below.

- Distributed systems software including scalable and secure tools and services to facilitate large-scale national and international scientific collaboration and high-performance software stacks to enable the discovery, management, and distribution of extremely large data sets generated by simulations or by science

experiments such the Large Hadron Collider, the Intergovernmental Panel on Climate Change, and ITER.

- Advanced network technologies including dynamic optical network services, scalable cyber security technologies, and multi-domain, multi-architecture performance protocols to seamlessly interconnect and provide access to distributed computing resources and science facilities.

Coordinated efforts across ESnet, the LCFs, NERSC, and the Next-Gen program support both an R&D portfolio and ongoing deployment in a manner that advances the end-to-end performance needed for large-scale collaboration and data-intensive science. The goal is to advance the networking, collaboration tools, and facilities that enable scientists worldwide to work together and share extreme scale scientific resources to address the most challenging scientific questions. Re-use and shared allocation synergies must be discovered, while also accommodating differences among scientific communities and needs for autonomy. Appropriate governance mechanisms are also needed to achieve high quality environments for scientific discovery.

3.4 Two Examples Requiring Integrating Networks and Data Management with Science Environments

Two frontier distributed science communities, the Earth System Grid and the Open Science Grid (which supports the US component of the LHC experiments), demonstrate today how distributed science communities in the near future will be augmented by persistent but evolving high performance infrastructures (including advanced networking, computational resources, and storage) that address common and unique requirements for security, openness, and extensibility through providing their services in ways that will integrate with service-oriented architectures (SOA).

The Earth System Grid Center for Enabling Technologies (ESG-CET) is a collaboration of seven national laboratories and a university that has become recognized as a world-class source for climate modeling data and a global leader in developing a federated "built-to-share" scientific discovery and sharing infrastructure for climate modeling communities world-wide. Designed to address the distributed access, management, and use of petabytes of observation and climate-simulations in a collaborative environment, ESG-CET is currently supporting more than 9,000 registered global users and managing more than 200 terabytes of data in a collaborative environment that supports shared use of federated data, information, models, analysis, visualization tools, and computational resources. Based on the Internet and the need to access and analyze large-scale data anywhere in the world, ESG handles between 400 and 600 gigabytes of data downloads each day.

The Open Science Grid (OSG) is a virtual facility jointly funded by DOE and NSF. Its primary goal is to provide the organizational framework, middleware stack, and operational support for a national distributed computing infrastructure for high throughput computing in support of the broadest possible scientific community within DOE and NSF. OSG accomplishes this via a consortium of scientific and engineering communities at universities and national labs, in addition to the funded OSG project.

The Open Science Grid (OSG) provides a production environment for distributed data-intensive science through a consortium (30 communities and 60 sites) that consists of domain scientists, software developers and providers of computing resources using distributed computing tools and computing resources. In addition, OSG has significant US responsibilities as the sole US contribution to the Worldwide Large Hadron Collider (WLHC) Computing Grid. The scale and complexity of the LHC global collaboration is unprecedented. The OSG has effectively mobilized the community's engagement and expertise to meet worldwide LHC goals in FY09, in preparation for the anticipated 15 petabytes of data per year, which has to be stored, backed up, and made available to more than 7,000 scientists around the globe. OSG is the US trusted leader and global representative to the WLHC Computing Grid in supporting the multi-institutional tiered method for data movement and stewardship that is the only viable means to address the scale of this shared endeavor and optimize the distributed resources and needs of all institutional partners.

4 Argonne Leadership Computing Facility (ALCF)

4.1 Background

For centuries, scientists have built instruments to aid our understanding of our complex universe. From Galileo's telescope to today's particle accelerators, these devices, combined with our indomitable curiosity to unravel how everything works, have been critical to scientific discovery.

Sponsored by the U.S. Department of Energy's Office of Science, the Argonne Leadership Computing Facility (ALCF) works hand in hand with the world's best computational scientists to support research in over a dozen different scientific domains, ranging from chemistry, astrophysics, and climate research to computational proteomics and life sciences. This broad span of disciplines also covers a mind-boggling range of physical scales. At one end of the scale, scientists using Intrepid are seeking to understand the interactions of the smallest components in the universe — quarks and gluons — that account for most of the visible matter in the universe. At the other end of the scale, computational scientists are using Intrepid to understand the brightest and most powerful exploding stars.

However, scientific discovery with Intrepid is not limited to the hard-to-imagine realms of sub-atomic particles or galaxies light years away. Researchers are using the ALCF to study and explore key scientific problems that underlie important challenges that face our society. For example, a team of researchers funded by the National Institutes of Health is investigating one of the leading causes of death in the United States: catastrophic rhythm disorders of the heart. Their discoveries could lead to safer and more effective treatment of patients. Likewise, scientists from Pratt & Whitney are using Intrepid to understand the complex processes within aircraft engines. Expanding our understanding of jet engine combustors is the secret to improved fuel efficiency and reduced emissions. Lessons learned from the scientific simulations of jet engine combustors have already led Pratt & Whitney to newer designs with unprecedented reductions in emissions, noise, and cost of ownership.

In addition to working with scientists running experiments on Intrepid, we have become a nexus for the broader global community. In partnership with the Mathematics and Computer Science Division at Argonne National Laboratory, we have created an environment where the world's most challenging computational science problems can be addressed. Our expertise in high-end scientific computing enables us to offer guidance for applications that are transitioning to petascale as well as to produce software that facilitates their development. Our software advances include the MPICH library, which provides a portable and efficient implementation of the MPI standard — the prevalent programming model for large-scale scientific applications — and the PETSc toolkit, which provides a programming paradigm that eases the development of many scientific applications on high-end computers.

4.2 Key Local Science Drivers

4.2.1 Instruments and Facilities

The ALCF is a relatively new facility. The first eight racks of Blue Gene/P went into production about a year ago, and we brought the facility to full production on Feb. 2nd, 2009. Unfortunately, this means we do not have historical data to draw upon for trend analysis. The ALCF currently fields four user-accessible resources: two compute resources and two visualization resources (Figure 1). For porting, testing, debugging, and early scaling work, we have a compute resource named Surveyor, which is comprised of a single rack of the IBM Blue Gene/P. It has 4,096 cores based on the PowerPC 450 chip operating at 850 MHz for a total peak performance of 13.9 teraflops (TFLOPS). The associated test visualization resource is Gadzooks. It consists of four COTS 1U servers with dual quad-core 2.0 GHz Xeon processors and 32 gigabytes (GB) of RAM. Pairs of these servers "sandwich" an NVidia S4, which contains four NVidia Quadro FX5600 graphics cards. The servers are connected to the S4 via a PCIe V2.0 x16 card. Logically, each server contains two of the graphics cards. Each of these cards is capable of producing 500 single-precision gigaflops (GFLOPS).

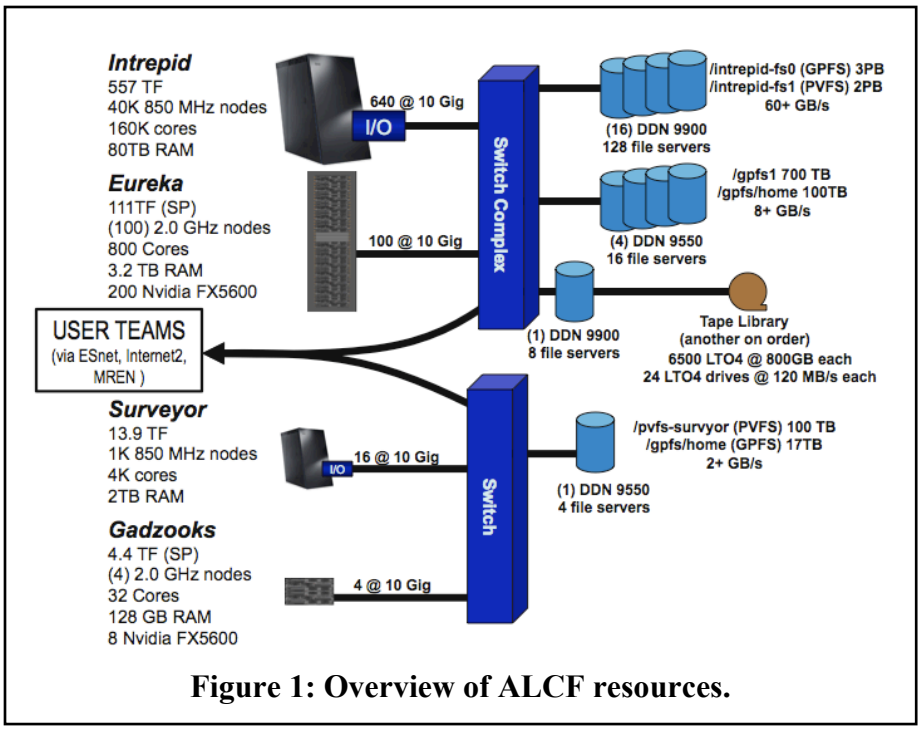


Figure 1: Overview of ALCF resources.

Our production resources use the same base building blocks as our test and development resources. Intrepid is our major computing resource. It is a 40-rack, 160K-core, Blue Gene/P with a peak of 557 TFLOPS. It currently ranks #5 in the TOP500 list with a LINPACK score of 450 TFLOPS. Paired with Intrepid, Eureka handles data analytics and visualization. It consists of 100 nodes identical to those described above for Gadzooks. Eureka provides 111 mostly single-precision TFLOPS in four very densely packed racks. At the time it was installed, Eureka was the largest installation of NVidia S4s in the world.

The Blue Gene/P contains five networks. Three of them (torus, tree, and barrier) are internal and used primarily by MPI for node-to-node communication. The fourth is a gigabit Ethernet-based RAS (Reliability, Availability, Serviceability) network used for administration, monitoring, booting, etc. The fifth and final network is the I/O network. As is typical of supercomputers of this scale, compute node I/O is aggregated at I/O

nodes before going to the storage system. On Intrepid we have a ratio of 64 nodes/256 cores per I/O node for a total of 640 I/O nodes. The Blue Gene/P I/O nodes have 10 Gigabit Ethernet on a chip. Given that we needed at least 640 ports of 10 Gigabit Ethernet, we chose to run a single network fabric based on Myricom 10 G/Myrinet. We were the first users of Myricom's new 2Z technology, which accepts a 10 Gigabit Ethernet frame from an I/O node and converts it to a Myricom frame and vice-versa. We have 10 of Myricom's 512 port switches. We currently have just over 1,000 ports active, and the core can scale to 2,048 ports by adding only line cards (no additional switches needed).

We brought our disk storage up in two phases. The first phase was based on four Data Direct Network (DDN) 9550 SANs. This provided 160 terabytes or TB (raw) for home directories and 640 TB (raw) for fast parallel file systems. Each DDN 9550 is capable of 2.2 GB/s for an aggregate maximum theoretical storage bandwidth of 8.8 GB/s. The second phase was based on the brand-new DDN 9900. (We got the first 17.) Sixteen of them are used together for our fast parallel file system. We support both GPFS, an IBM commercial parallel file system, and PVFS, an open source parallel file system, for which Argonne National Laboratory is a major contributor, specifically optimized for MPI parallel I/O. Raw storage is 7.6 petabytes (PB). However, DDN runs RAID 3 with dual parity, meaning 20% is lost to parity, leaving about 6 PB. Our usable disk space is 3 PB for GPFS, 2 PB for PVFS, with a small fraction left over for testing. Each DDN 9900 is capable of 5.5 GB/s for an aggregate theoretical bandwidth of 88 GB/s. We have achieved 60+ GB/s with the IOR parallel file system benchmark and have seen applications achieve 30 GB/s.

For tape storage, we currently have one Spectralogic T-950 8 frame, a 10,000-slot library with 6,500 slots in use. We have 24 LTO4 tape drives for an aggregate theoretical bandwidth of 2.8 GB/s to tape. The 17th DDN 9900 is used as a disk cache in front of the tape. It has 480 TB of raw capacity and a peak bandwidth of 5.5 GB/s. We use Amanda, an open source backup tool, for doing system critical backups of our hosts, and user-initiated HPSS transfers for archiving users' scientific data. Currently, all offline storage accesses are staged through the disk cache, though we do have the capability to go directly to tape. Also note that our HPSS installation is not visible to the external world. To get data into or out of HPSS, the data must pass through our parallel file system.

The ALCF supports a broad range of scientific disciplines. Figure 2 shows the distribution of core hours across those disciplines for the 2009 INCITE year. In terms of projects and users, we currently have 28 INCITE projects, 104 director's discretionary projects, and 1 Office of

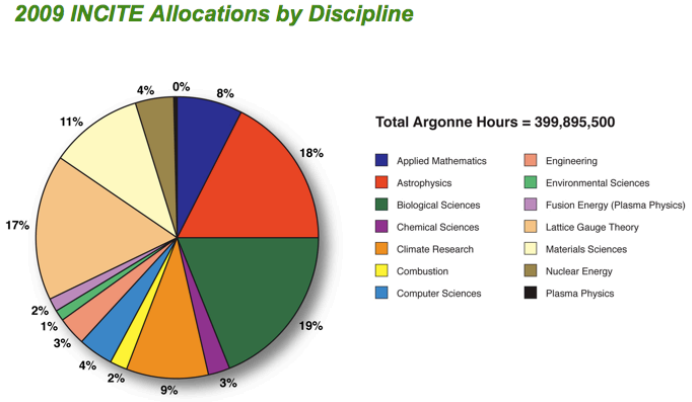


Figure 2: ALCF INCITE allocations.

Science discretionary project. This translates to 221 INCITE users and 437 active Intrepid users in total. The latter number includes overlap with the INCITE users who also have discretionary projects, vendors who have accounts but no allocations, and so on. Surveyor has 407 users, and there is significant overlap with the Intrepid users. We believe that the number of projects and users will stay relatively constant, with the science demands of the projects growing and driving future expansion of resources.

4.2.2 Process of Science

Most agree today that science is built on three "legs": theory, experimentation, and computation. Within the computation realm, we tend to follow the sequence of "develop code, validate code, run code, analyze results, publish findings, repeat." There are, of course, a variety of ways to go about accomplishing this sequence. As noted above, we provide substantial resources in the four primary areas of interest: compute, fast parallel storage, analytics/visualization, and archival storage. Many of our users are content to operate entirely, or almost entirely, within the confines of the ALCF. For instance, consider this recent quote from one of our larger users, computational scientist Paul Fischer of Argonne National Laboratory:

> Eureka provides a vital link between simulation and analysis by allowing scientists to probe and interrogate their data in an interactive manner. Since Eureka and Intrepid share a disk, there is no need to move data between machines. Eureka dramatically reduces the amount of time needed to create these hugely complex visualizations, while greatly boosting their quality.

In contrast, there are other users who operate as part of a large consortium or have a central community data repository. In those cases, the ALCF is just a compute farm, and data is only cached locally for a limited time until it can be moved offsite. The Lattice QCD community fits this model. They use Grid tools, such as GridFTP, SRM, etc., to move their data offsite. CHiMES is another project in a similar situation. They produce on the order of 5 TB per wall-clock day. They move it all back to a central repository for storage and post-processing, but this is not considered time critical.

For many projects, I/O and storage requirements are minimal relative to their compute requirements, so either model is completely viable and has no significant impact on our network planning.

4.3 Key Remote Science Drivers

4.3.1 Instruments and Facilities

From a facilities perspective, the primary remote science drivers are our campus networking and the dedicated data movement (GridFTP) servers in our facility. Figure 3 shows a network diagram of our WAN infrastructure. In the ALCF, we have a Force10 E600 switch with 10 Gigabit Ethernet that serves the ALCF public IP space. We currently have a single 10 Gigabit uplink lit but have facilities in place for two more and have 12 pairs of fibers available. The ALCF has two dedicated GridFTP servers. They are four core Opterons with 12 GB of RAM and a Myricom 10 Gigabit NIC installed. They have essentially no local data storage, so the 10 Gigabit NIC is doing double duty. It brings the data in from the network-mounted parallel file system, and then sends it over the WAN or vice versa. Since we just went fully into production in early 2009, we do not have a good sense for what our WAN transfer workload will look like over the long haul.

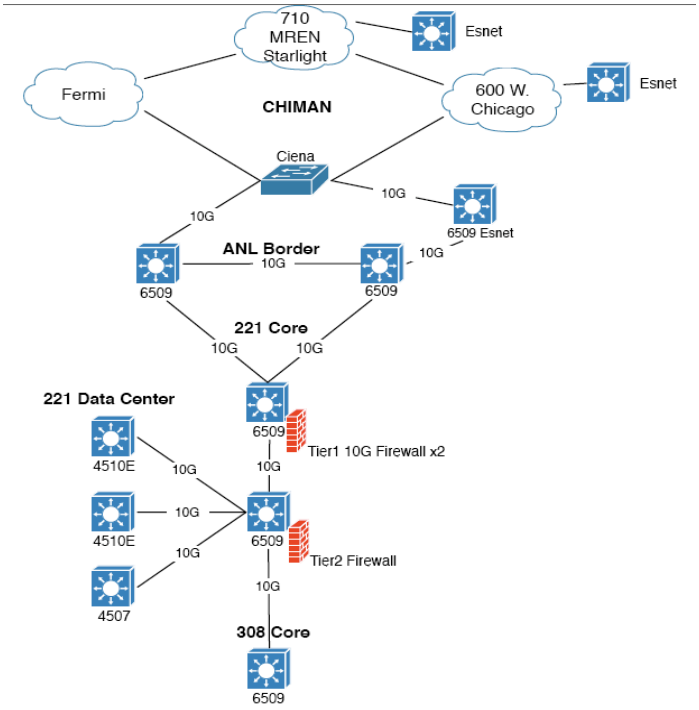


Figure 3: Argonne WAN infrastructure.

4.3.2 Process of Science

The primary difference in the science process for remote science drivers is data movement. One could envision the local science process a sub-case of remote science, where the data movement stages are trivial or skipped entirely. Most typically, this represents movement of file data. This could be staging in input files, staging out results files for archiving, or further analysis at another site. Workflows become much more complicated, particularly since a data movement task of significant size represents a co-scheduling problem. Significant resources must be available at both the source and destination simultaneously for a data transfer to take place. However, support for other services such as remote computational steering, telepresence for experiments, and real-time control systems over the WAN are potential issues as well. One wonders if a chicken and egg problem exists: Do people not do these types of things more often because the tools don't exist, or do the tools not exist because people don't really need them?

4.4 Local Science Drivers — the Next Two to Five Years

4.4.1 Instruments and Facilities

The ALCF is already in the planning phase for our next-generation system. It is too early to provide solid details, but the following are characteristics that a leadership class machine will likely have in a two to five year timeframe:

- Compute. In the tens of petaflops, likely between 10 and 30. A core count of more than one million cores is probable.
- RAM. Even with the RAM/core ratios dropping, at one million cores, there will be a petabyte or more.
- *Internal networks.* Bandwidth will continue to go up. The challenge will be to maintain low latency as systems scale. Torus networks will likely continue to be the architecture of choice to control costs for high-speed connections to hundreds of thousands of nodes.
- I/O. Aggregation levels will continue to grow. A single I/O node may be servicing thousands of compute cores. Even with that level of aggregation, the number of I/O nodes will climb to over a thousand as the number of cores increases into the millions.
- Storage. This is an interesting one. One school of thought says the FLOP/BYTE ratio must get larger because of the difference in rates for CPU FLOP increases and disk storage/bandwidth increases. On the other hand, many supercomputing sites are already I/O starved in their designs, and I/O will increasingly be the bottleneck. The solution will likely be a combination of technological innovation to ameliorate the differences in growth rate and a realization that a larger fraction of the total budget will need to go to storage. Likely capacity ranges are high tens to low hundreds of PB. Bandwidth ranges are high hundreds of GB/sec to a few TB/sec.
- Archival Storage. The problem is even worse than with the disk. The LTO consortium roadmap shows tape densities doubling every two years, but bandwidth is only increasing by 150% every two years. With core counts in the millions, a file per process simply is not feasible. Thus, very large files, on the order of tens and hundreds of terabytes will be the order of the day. This means striping across wider and wider drive sets will be required to get any kind of reasonable access times. With files striped across multiple tapes and stored for many years, loss of data to a failed tape becomes highly probable. This will drive RAID-like parity protection schemes into tape. The potentially unacceptable access times will also mean that MAID (Massive Array of Idle Disks) technologies will continue to erode the domain of tapes.
- Software. Just as important as the hardware drivers for the above are the associated systems software pieces. How will MPI scale to millions of processes? How will internal networks deal with a state where a failed node is a daily occurrence? How will parallel file systems deal with thousands of I/O nodes and a trend toward all writes appearing random? How will user-level I/O libraries maintain performance?

4.4.2 Process of Science

The basic process of science will not change radically in the next two to five years. The steps taken will be basically the same, though computation will likely continue to grow in importance across all fields of science, and the challenges we face as we execute the steps will change. The challenges we expect to arise involve the following:

- Scale. As we move to many cores in the processor, multiple hardware threads per core, and millions of total cores, we will need to address the software stack and its ability to cope with this scale. Approaches that will be investigated include MPI everywhere, hybrid MPI-OpenMP, Partitioned Global Address Space (PGAS) languages, etc.
- I/O. One school of thought believes that I/O will be the limiting factor in the multi-peta or exascale. Processor speeds and computational power are outrunning I/O, particularly storage performance at an exponential rate. I/O time will become an increasingly large fraction of wall-clock time. As we scale, all I/O is tending toward random I/O, and keeping block sizes up to maintain storage performance becomes increasingly difficult.
- Resiliency in the Face of Failures. Despite the fact that the Blue Gene (and other) machines have very reliable parts, when there are millions of them, failures are going to be a daily event. Software, both systems and applications, must learn to adequately cope with this increase in failure rate.
- Power Consumption. Power consumption could become the overriding design consideration for computing. Besides being buzzword-compliant and offering environmental benefits, there is a sound business reason for this consideration. The cost to run and cool data centers is becoming a larger and larger fraction of operating costs and could easily exceed the cost of the hardware over its useful life.

4.5 Remote Science Drivers — the Next Two to Five Years

4.5.1 Instruments and Facilities

At the facility level, the remote science drivers fall into a few basic categories:

- Upgrade Our Networking. We expect the backbone will be at 100 Gbps to the border routers. It is unlikely that single hosts will be able to sustain even a large fraction of 100 Gbps in this time frame; so multiple 10 Gigabit paths will be required.
- InfiniBand. InfiniBand is a highly likely candidate for the storage/external connectivity needs in this timeframe. InfiniBand gateways into Ethernet networks will likely be of some importance.
- Data Movement Hosts. We will achieve massive bandwidth by aggregating more flows and that will require more servers. This will become a non-trivial infrastructure element.

4.5.2 Process of Science

There are a few key science drivers that will drive remote science in the next two to five years:

- Science today is an exercise in collaboration; widespread teams are the norm. This trend will only grow as we tackle more difficult problems.
- Data sizes will grow exponentially and may become impractical for any single facility to maintain.
- Real uses of computational steering could become commonplace.
- Real-time or near real-time processing of data could become much more important: simulation-assisted surgery, disaster response scenarios, etc.
- Hybrid simulations in the earthquake engineering community could require a low-jitter, highly reliable connection to control simultaneous experiments at different sites (for instance, two shake tables), as well as a coordinated simulation.
- The possibility of hosting large community data sets is being explored.

The reality is: Software, not hardware, will determine whether there is a need for more network capacity. The science opportunity is there. In the business world, it is all about the bottom line. This drives all decisions. Researchers are not that much different. They will do whatever they have to do to accomplish the best science. If that means running everything at a single center, they will run everything at a single center. If it means working in a far-flung collaboration, they will do that. Even today, network capacity is not the problem. Being able to utilize the network resources already available is the problem. If we make utilizing the network the way to accomplish better science, then scientists will consume everything we put in front of them and then some. We need software tools and network protocols that will provide the following:

- Easy often equates to more science and less frustration, so use of the network needs to be simple and reliable (from the user experience perspective), as well as fast.
- Applications utilizing the network need to negotiate their requirements dynamically, and then the network needs to meet those requirements without packet loss.
- The applications must have network tools and protocols available that let them achieve their negotiated levels of service.
- Failures have to be addressed gracefully. If users are moving 10,000 files and the transfer fails, we can't ask them to figure out what got there and what didn't.
- Real-time and control systems over the WAN will have extremely tight jitter specifications and will need high reliability. This will likely require protocol development to support these requirements.

4.6 Beyond Five Years — Future Needs and Scientific Direction

The discussion for needs beyond five years is not so different than that for two to five years. Scientists can already show problems that require exascale computation to solve. The astrophysics community is projecting the production of hundreds of petabytes of data per day. The question is: Will a disruptive technology be discovered that resolves a major problem of today (at least for a little while) and moves the bottleneck somewhere else?

A type of usage that might potentially become more widespread in the long run is computational steering. A key reason for the relatively low use of computational steering

is that large simulations take a long time for each time step, which means watching the evolution of a simulation in real time is rather tedious. As systems approach exaflops in computing speed and I/O speeds grow proportionately, some applications will run fast enough on those systems that real-time visualization and computational steering will be practical and desirable. In some cases, this type of use will require higher network bandwidth and lower and more predictable latency.

As archives of scientific data grow over time and the scientific community embraces more fully the sharing of data, there will be more applications that involve retrieval of large subsets of data from several huge, distributed archives for comparative analysis and data mining. Examples include digitized mammal brain collections and seismic data for geophysics.

4.7 Outstanding Issues

One particular issue we are concerned with is the impact of computer security on data transfers. The leadership computing facilities are high-visibility targets for all manner of computer crime. As a result, One Time Password (OTP)/cryptocards are desired for all access to production computing resources in the ALCF. However, this is not feasible out of the box right now. The default standard for GridFTP is GSI. GSI is cryptographically strong, but a site can't easily guard against weak or non-existent certificate pass phrases, very long-term proxies, etc.

There is the new sshftp functionality of GridFTP, but many users have existing Grid credentials. OTP is also a problem for long-running transfers due to re-authentication in the face of failures, multiple authentications if running high levels of concurrency, etc. What we would like to see is a standard, out-of-the-box solution that would interface with site PAM authentication and add a signed assertion to an X.509 credential proving the user successfully authenticated using our site authentication methods. The site should also be able to set policy, such as: "refuse proxies with a remaining time of greater than X hours." Standard GSI can then be used to get the benefits of a single sign-on, but a site can also be sure that its authentication processes have been followed.

Pieces of this standard exist in tools like VOMS, GridShib, etc., but they have not been packaged in a way that provides exactly the functionality we desire.

4.8 Summary

Table 1 summarizes the key science drivers and anticipated network requirements for the ALCF in three time frames.

Table 1: ALCF requirements summary

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• ALCF production resources (intrepid)	• Large file transfers. Other labs and computing centers are common targets, but it can be any institution based on INCITE users needs. • Some real-time video, computational steering, real-time control apps possible.	• Node to node is handled by proprietary vendor interconnect. 425 MB/s per link. • Node to storage is approx. 1,000 ports of 10 gigabit. • Other local needs are primarily admin-related and are trivial.	• 10s of TB/day • 10-30 Gbps
2-5 years	• Next major machine upgrade	• Large file transfers. Other labs and computing centers are common targets, but it can be any institution based on INCITE users needs. • Real-time video, computational steering, real-time control apps more common, but still relatively small in comparison to file transfers.	• Node to node is handled by proprietary vendor interconnect. 1-5 GB/s per link. • Node to storage is likely InfiniBand-based and on the order of 3K-5K ports • Other local needs are primarily admin-related and are trivial.	• 100s of TB/day • 100-300 Gbps
5+ years	• Push towards exascale computing	• Massive data sets are common. File transfers still dominate, but WAN file systems, distributed databases use grows. • Machines are sufficiently powerful that computational steering, real time simulations are used regularly • Use of collaboration tools continues to grow	• Node to node is probably still handled by proprietary vendor interconnect, but could be standards based, such as InfiniBand. • Node to storage is likely InfiniBand or other standards-based interconnect. • Other local needs are primarily admin-related and are trivial.	• Petabytes per day • Terabit networks

5 National Energy Research Scientific Computing Center (NERSC)

5.1 Background

NERSC is a DOE High Performance Computing Center that supports a broad range of science and a wide user base. In 2009, NERSC supported over 400 projects and close to 3,000 scientists in their computational and storage needs. In Allocation Year 2009, NERSC committed to 40% of its delivered cycles going to capability jobs that use at least 1/8th of the available processors (i.e., 2048) on NERSC's largest system (dual-core Franklin) under ASCR's Program Assessment Rating Tool (PART) metric. With the upgrade of Franklin to quad-core processors, 1/8th of the available processors will jump to 4096 processors for Allocation Year 2010 with the target of 30% of its delivered cycles going to capability jobs.

NERSC delivers these production resources to scientists in the broadest range of disciplines, inclusive of all six programs under the Office of Science. As of the beginning of August, 87 projects have run capability jobs at NERSC totaling over 59 million CPU hours (see Table 2).

Table 2: Capability CPU hours by program office

Capability Projects at NERSC by Office and Discipline				
DOE office	**Science Area**	**Number of Projects**	**Capability CPU Hours**	
ASCR	Applied Math	3	875,938	1.47%
	Combustion	2	4,560,530	7.66%
	Computer Sciences	11	804,720	1.35%
ASCR Total		**16**	**6,241,188**	**10.48%**
BER	Applied Math	1	238,091	0.40%
	Climate Research	5	7,101,707	11.92%
	Environmental Sciences	4	1,316,731	2.21%
	Life Sciences	5	2,945,236	4.94%
BER Total		**15**	**11,601,765**	**19.48%**
BES	Accelerator Physics	2	1,662,563	2.79%
	Chemistry	8	4,542,343	7.63%
	Geosciences	2	1,139,828	1.91%
	Materials Sciences	6	3,885,989	6.52%
BES Total		**18**	**11,230,723**	**18.85%**
FES	Fusion Energy	19	13,968,251	23.45%
FES Total		**19**	**13,968,251**	**23.45%**
HEP	Accelerator Physics	6	5,613,200	9.42%
	Astrophysics	5	1,342,753	2.25%
	High Energy Physics	1	339,084	0.57%
HEP Total		**12**	**7,295,037**	**12.25%**
NP	Accelerator Physics	1	144,485	0.24%
	Lattice Gauge Theory	2	6,957,326	11.68%
	Nuclear Physics	4	2,128,623	3.57%
NP Total		**7**	**9,230,434**	**15.50%**
Grand Total		**87**	**59,567,399**	**100.00%**

The single largest consumer of network bandwidth at NERSC is bulk data movement, either through computational systems or directly to the Centers High Performance Storage System (HPSS). The data transferred through computational systems will subsequently be stored in HPSS. Thus, an analysis of HPSS usage gives a very accurate picture of WAN data requirements.

Table 3: HPSS usage statistics

	Total TB I/O	*Total TB Read*	*Total TB Retained*
2007	3,156	682 (21%)	1,327
2008	3,647	631 (17%)	2,618

Since ESnet is the only way NERSC resources can be accessed by NERSC users, it is critical that ESnet provide reliable, high performance connections to NERSC that are state of the art. By high performance, we mean that the actual end-to-end bandwidth (EEB) a scientist experiences from the host at their site to the host at NERSC is sufficient to accomplish their science. The EEB is dependent on many factors, including the ESnet backbone speed, the network capabilities at NERSC, the network capabilities at the remote site, end-host issues, software design issues, and the reliability and stability (amount of packet loss) from end to end.

5.2 Key Local Science Drivers

5.2.1 Instruments and Facilities

Currently, the largest system at NERSC is Franklin (NERSC-5), a Cray XT4. NERSC also has Bassi, an IBM p575 Power5 system; Jacquard, a commodity Opteron cluster; and DaVinci, an SGI Altix 350 system used for analytics. Lastly, NERSC is home to the PDSF computational cluster, which is used for high energy and nuclear physics research. NERSC has announced that the contract for its next system (NERSC-6) has been awarded to Cray. This upgrade of NERSC capabilities will be greater than one petaflop and have an aggregate filesystem bandwidth three times that of Franklin.

> As NERSC is the primary supercomputing center for DOE's Office of Science, making Cray's latest technology available to our users will accelerate innovation across a wide range of scientific disciplines, helping scientists tackle problems of vital importance to our nation's future.
>
> — Dr. Michael Strayer, ASCR Associate Director

NERSC also has two High Performance Storage System (HPSS) instances, which currently store approximately 7.1 PB of data, containing scientific data of national interest. For some details on the computational systems at NERSC, see Table 4.

Table 4: NERSC computational systems (2009)

	System Performance (Theoretical Peak TF/sec)	*Local File System Capacity (TB)*	*System NERSC Network Connectivity*
Franklin	356	436	4ea 10Gbps
Bassi	6.7	100	4ea 10Gbps
Jacquard	3.1	30	2ea 10Gbps
PDSF	5.0	471	2ea 10Gbps
DaVinci	0.2	24	2ea 10Gbps

The NERSC Center is currently working to connect the NERSC Global File System (NGF) directly to the computational systems listed above, thus removing a significant amount of network traffic between NGF and the computational system's local disk. The network traffic from these computational systems should be dominated by SSH and the various data transfer clients (GridFTP, FTP, HSI, HTAR, and BBCP).

For 2008 HPSS usage at NERSC, 71% of transfers originated from systems at NERSC. HSI is the most commonly used data transfer software, and is involved with 82% of transfers to HPSS. HTAR is a distant second, and used in 11% of the transfers to HPSS.

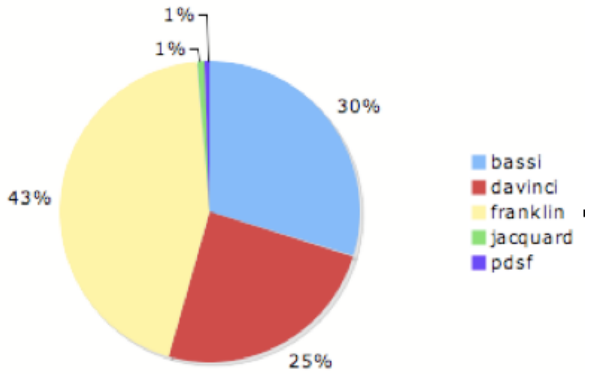


Figure 4: HPSS data received by system (March 2009)

Figure 4 shows the data volume distribution of data transfer to HPSS by NERSC systems for a typical month. Since Franklin went into production, HPSS daily I/O has increased by 50%. In the past year PDSF usage of HPSS has decreased markedly from being the dominant system using HPSS (about 50% in 2007) to just below Bassi and DaVinci on a regular basis. This is one reason NERSC places so much emphasis on optimizing I/O and HPSS bandwidth on Franklin.

5.2.2 Process of Science

As the flagship computational facility within DOE's Office of Science that serves a broad science base, users come to NERSC for efficient and reliable computation and storage resources. It is with rare special arrangement that NERSC provides a storage or computational resource dedicated to a small set of users. Figure 5 shows the number of distinct NERSC storage allocations for 2009 within each DOE Program and demonstrates the diversity of work performed at NERSC.

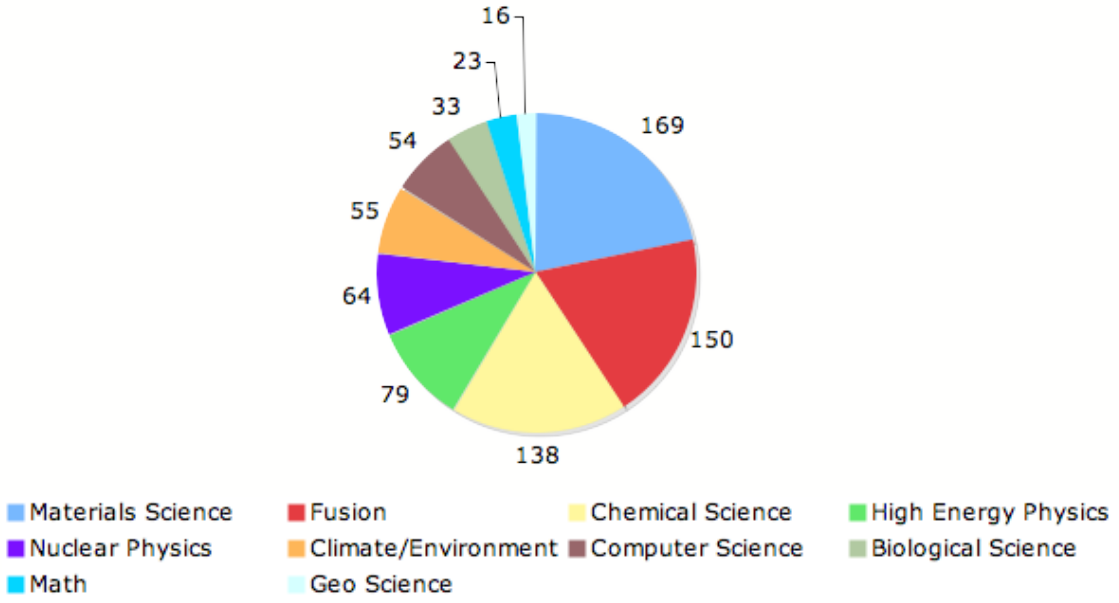


Figure 5: Number of HPSS allocations by DOE program (2009)

There is no one process of science or single method at NERSC that can be described here as meeting the broad NERSC user base, but most are complex and some have extensive pipelining. An example is the Berkeley Drosophila Transcription Network Project (BDTNP), which studies mechanisms controlling cell specialization during embryogenesis (embryo development). See Figure 6.

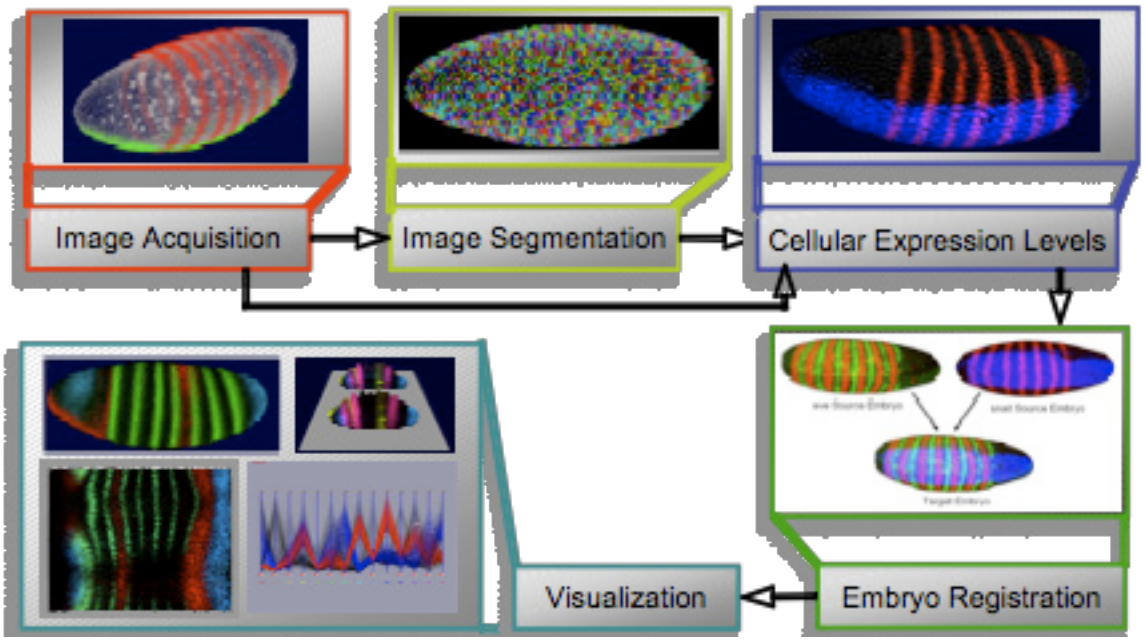


Figure 6: Computational workflow for study of embryogenesis

5.3 Key Remote Science Drivers

5.3.1 Instruments and Facilities

The two main drivers of WAN network bandwidth at NERSC are incoming data from remote sites that are stored on the NERSC HPSS system and then used on NERSC computational systems, and data generated on the computational systems that are usually stored on HPSS and on occasion transferred to a remote site. Since NERSC has many projects and users running simultaneously, we use the actual growth of scientific data archived on the NERSC HPSS system, the measured computational capabilities of NERSC systems, and the recorded traffic on the NERSC ↔ ESnet link to predict the NERSC to ESnet link bandwidth requirements.

NERSC developed software to categorize the border traffic into broad categories. Tables 5 and 6 summarize three months of ESnet network border traffic for NERSC.

Table 5: NERSC↔ESnet border traffic in bytes received by type (inbound)

	Jan 2009	*Feb 2009*	*Mar 2009*
HPSS	46%	53%	65%
Other	39%	29%	26%
SSH	6%	5%	6%
HTTP	1%	2%	1%
FTP	1%	2%	1%
TOTAL DATA (TB)	28.2	22.0	27.7

Table 6 shows that HPSS is the largest consumer of ESnet network bandwidth considering bytes sent to NERSC across the ESnet border. NERSC conducted an extensive study of the *Other* category in 2006 and recent spot checks show that the components that make up this category still apply to current network usage. *Other* contains transfer protocols that are not reliably singled out in our categorization software such as GridFTP and BBCP.

Table 6: NERSC/ESnet border traffic in bytes sent by type (outbound)

	Jan 2009	*Feb 2009*	*Mar 2009*
HPSS	22%	14%	22%
Other	46%	32%	21%
SSH	30%	52%	53%
HTTP	1%	1%	2%
FTP	1%	1%	2%
TOTAL DATA (TB)	10.9	16.2	18.4

Table 6 shows that the dominant protocol for moving data offsite from NERSC is SSH (i.e., SCP). This is the least efficient and least performing method of data movement of all the data transfer clients the Center provides. The Center recognized this fact and established the Data Transfer Working Group (DTWG) between ESnet, ANL, ORNL, and LBNL/NERSC to provide a dedicated, easy-to-use, high performing resource at each Center for data movement between the Centers. The DTWG provides data transfer nodes that became available in May 2009, achieving 200 MB/sec or greater on transfers between the Centers.

Figure 7 shows the network traffic for March 2009. This figure shows that NERSC is a net importer of data. Historical graphs show, that on average, NERSC ingests two or three times as much data as it exports. This is also consistent with the information on TOTAL TB provided by month in Tables 5 and 6 above. Figure 7 also shows a few impressive peak performance transfers (2 Gbps or greater) but otherwise reflects the typical SSH (i.e., SCP) rates of 3-30 Mbps that users can expect.

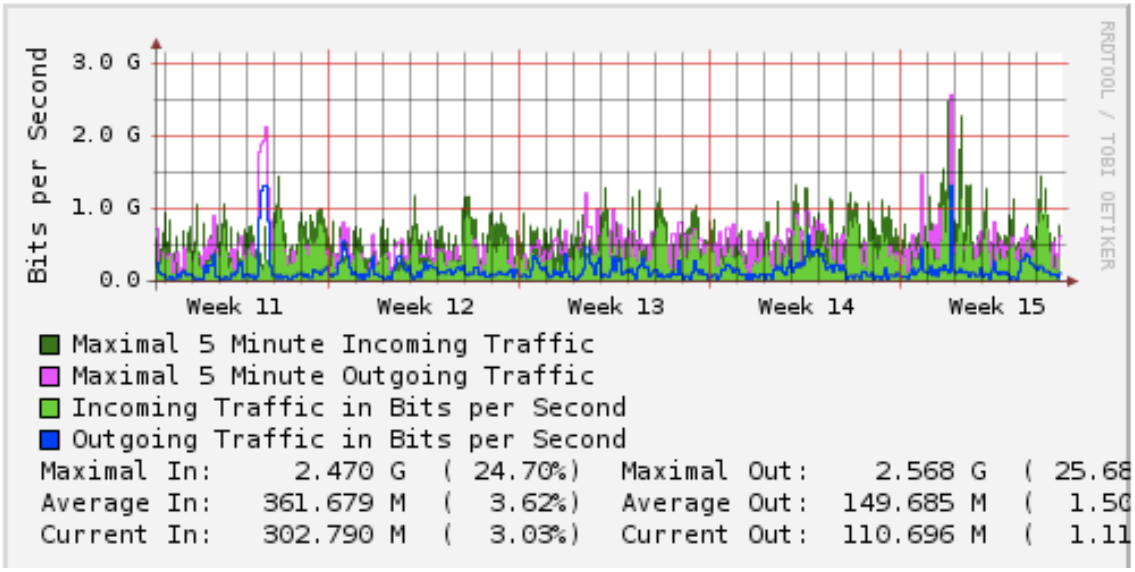


Figure 7: NERSC border traffic (March 2009)

HPSS is the primary network user, and its capacity helps determine the network capacity required to support the Center. HPSS has been growing at a consistent 40% per year for the last three years (70% for a number of years prior to 2007). Figures 8 and 9 show actual capacity of HPSS over its history at NERSC (Figure 9 is plotted on a logarithmic scale).

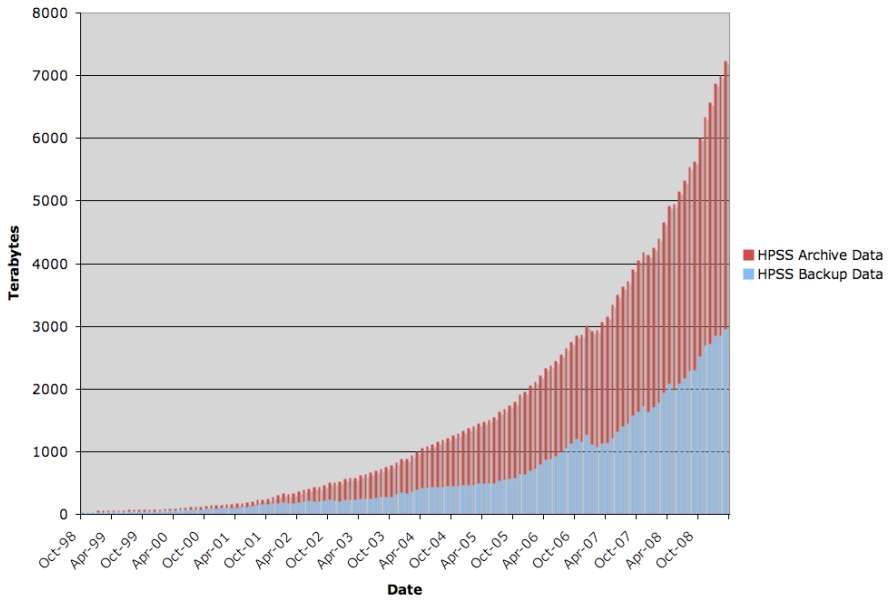


Figure 8: HPSS total data stored

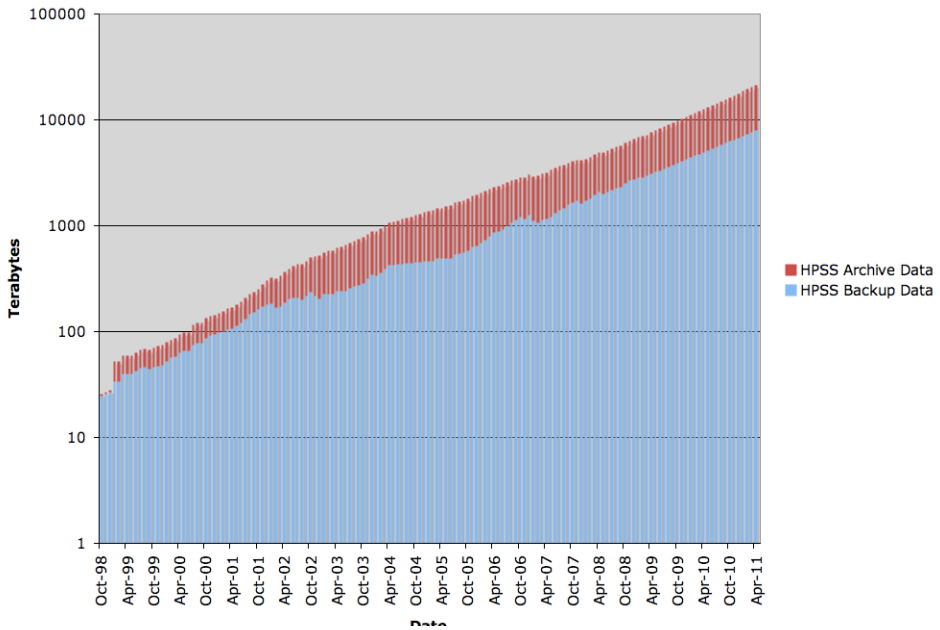


Figure 9: HPSS total data stored (log scale)

5.3.2 Process of Science

NERSC has identified the growing scientific need of our researchers to collaborate with their scientific fields and become "computational ambassadors" for their respective communities. To enable this goal, NERSC has developed a computational resource to support Scientific Gateways, where scientists can enable their communities to further discovery.

An example of the potential of Scientific Gateways is with the Deep Sky project (Figure 10). Its goal was to create a gateway for selecting and manipulating telescope images in a large dataset (60 TB and growing). Its impact was to discover 36 supernovae in six nights of data during the commissioning of the Palomar Transient Factory (PTF) survey. The scientific gateways allowed 15 collaborators from around the world to work non-stop for the first 24 hours during this discovery phase.

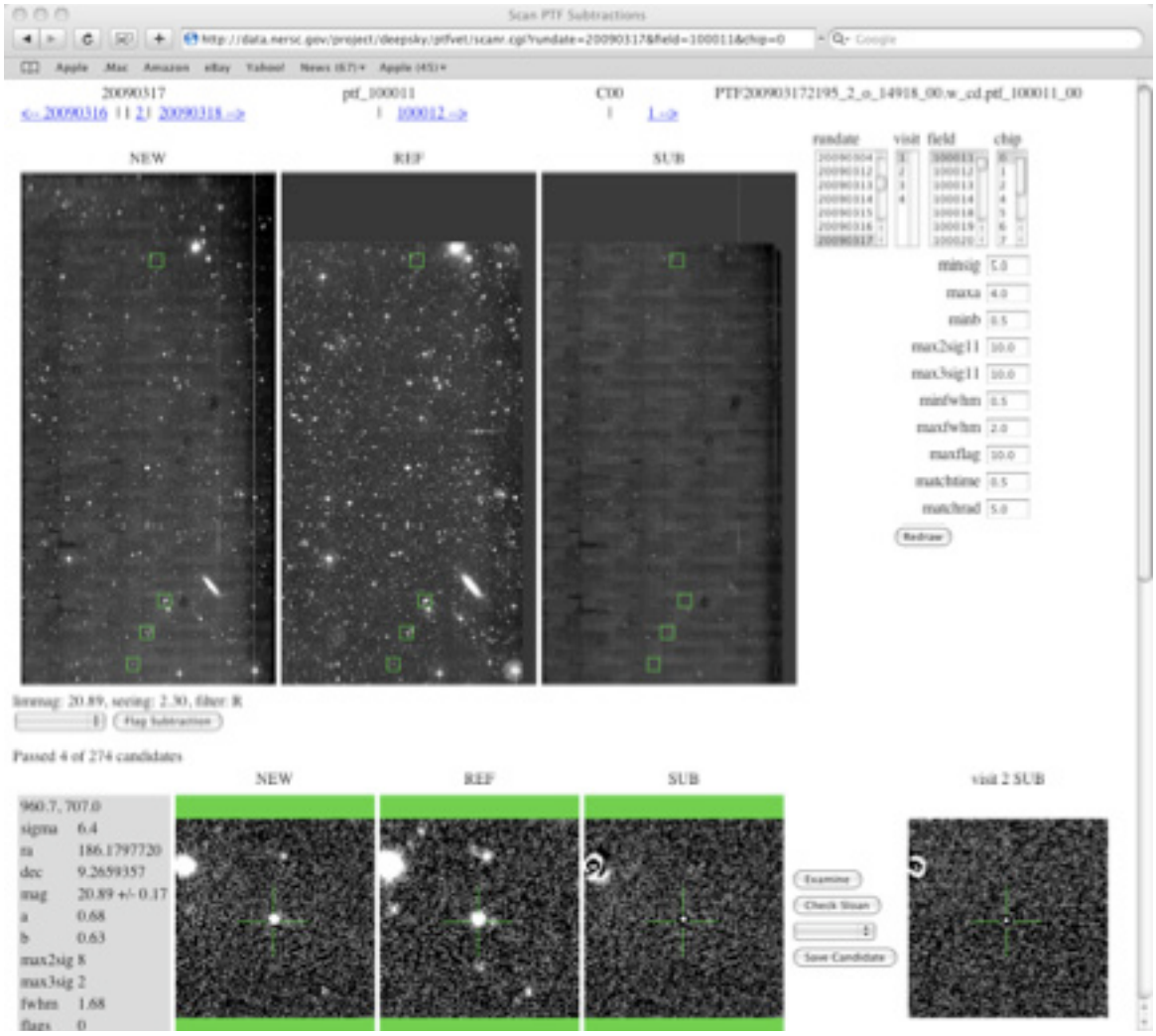


Figure 10: Deep Sky scientific gateway interface

5.4 Local Science Drivers — the Next Two to Five Years

5.4.1 Instruments and Facilities

NERSC expects that the major focus of the next two to five years in high performance computing will be in data-centric computing and responding to the challenges presented by the massively multicore nodes required to create exascale systems. Data-centric computing will likely result in an increase in massive datasets accessed through transactional queries as well as data presentation to collaborating scientists in diverse locations. Multicore nodes will require a change in programming models comparable in scope to the migration from vector to parallel computing.

With funding from the American Recovery and Reinvestment Act of 2009, NERSC has many opportunities to expand its computing portfolio in areas such as cloud computing and hosting computational or storage resources for specific projects. These potentials are

substantial and could affect the current storage and network statistics used for planning and forecasting. NERSC is pursuing several of these opportunities. However, it should be noted that despite numerous significant changes over time, NERSC continues to observe a steady increase in storage and network demand.

For the last five years, the NERSC facility has been a net data importer covering the entire Office of Science. However, Figure 11 shows that the gap between amounts of data imported versus exported is closing. NERSC identified this and established the Data Transfer Working Group to improve the data transfer performance between the DOE supercomputer centers and other key sites.

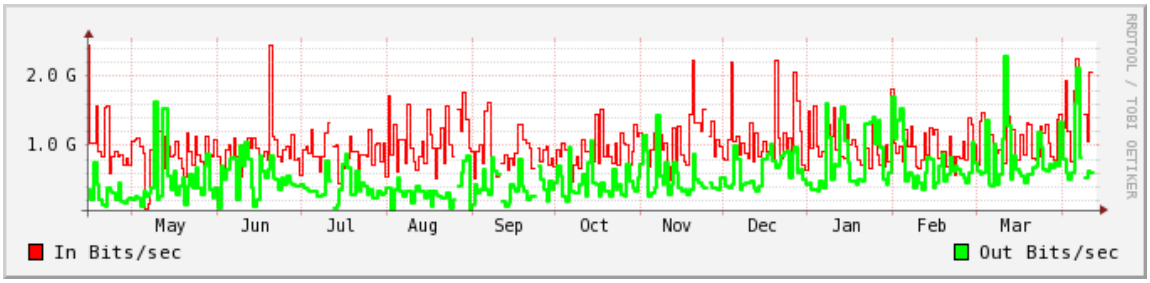


Figure 11: NERSC border inbound/outbound traffic (2008-2009)

5.4.2 Process of Science

NERSC will be concentrating on developing the programming models required to effectively use the next generation of HPC systems. NERSC expects that these massively multicore systems will demand communication capabilities that cannot be delivered by a flat MPI computing model and that a heterogeneous approach (e.g., threaded + MPI) will be required.

As well as a major overhaul in computation to meet the challenges of the future, the data transport mechanisms currently used will also have to evolve. It is unclear if single socket TCP streams will keep pace with the growing data demand or if new parallel data transports based on new models like Hadoop will become pervasive. NERSC will be evaluating the current data transfer technologies and determining how they will best scale in the future. Some of the local protocols or transfer clients used in bulk data movement among the various NERSC resources are discussed below.

- **HSI** — A data transfer client provided to all NERSC users for use on all the NERSC clusters as well as from their desktop systems external to NERSC. This client is used approximately 40% of the time in bulk data movement at NERSC. Intellectual property rights prevent us from distributing the source code and thus we provide a binary distribution preconfigured for a limited number of the most common operating systems. The current protocol in HSI is called MOVER protocol and is highly synchronous in that handshakes are required between the client and HPSS mover systems involved in the transfer each MB or so during the transfer. Networks with packet loss or high latency communications quickly reduce a client capable of GB/sec transfer rates to single MB/sec (e.g., as seen

with normal WAN transfers using this client). In 2009, NERSC is funding Gleicher Enterprises to offer two new protocols that work very well over a WAN. This client is also capable of machine-to-machine transfers (just like BBCP) and is the only client besides GridFTP capable of HPSS to HPSS transfers. The client has some tunable settings for improved transfer performance.

- **HTAR** — A data transfer client provided to all NERSC users for use in HPSS transfers on all NERSC clusters as well as from their desktop systems external to NERSC. This client is used approximately 6% of the time for bulk data movement at NERSC. It is especially good at optimizing the transfer rate for large numbers of small files. It tars files directly into HPSS during the transfer, saving user time and client disk that would be necessary if the user tarred the files up themselves and then used HSI to transfer the files to HPSS. The client has some tuning options for improving performance of individual transfers.
- **globus-url-copy** — A client installed on all NERSC clusters and available as open source software to nearly any other operating system. Statistics show that this client is used less than 10% of the time during bulk data movement at NERSC. The major benefits of this client are striped transfers (e.g., using multiple TCP streams); its potential for authentication credentials that would work at multiple DOE sites; and its non-interactive ease of use (e.g., good for scripting). The transfer protocol is optimized for LAN or WAN transfers capable of utilizing high amounts of available network bandwidth. The client is highly tunable for optimizing transfer performance.
- **BBCP** — a client available on most NERSC clusters. This client is the simplest to use and is most like SCP. The protocol is optimized for LAN or WAN transfers and is capable of parallel TCP streams. The client is highly tunable for individual transfers in optimizing performance. It has a non-interactive ease of use like globus-url-copy.
- **SCP** — a client available on any platform and the easiest to use. This client is used approximately 30% of the time for bulk data movement at NERSC. However, it is the least efficient protocol for LAN and WAN transfers, as its specialty is security. This is the most common transfer protocol used for current WAN transfers at NERSC. The Data Transfer Working Group is addressing this issue and making other transfer clients similarly easy to use and better optimized for performance. This client is not tunable for optimized transfer performance. NERSC is starting to deploy the new HPNSSH SCP software that does dramatically improve transfer performance by using larger internal buffers.
- **FTP** — there are many FTP clients available on NERSC platforms and on the various external systems used to access NERSC. These are somewhat easy to use if the user has familiarity with the protocol and client features; however, the protocol is limited in functionality and not optimized for performance bulk data movement. A smaller percentage of users (10%) use this for bulk data movement at NERSC. These clients are not tunable for optimized transfer performance.

5.5 Remote Science Drivers — the Next Two to Five Years

5.5.1 Instruments and Facilities

Data warehousing and data distribution for the major scientific collaborations will be the key remote science drivers for the next two to five years. It is expected that spinning disk and flash storage will increase the capacity and bandwidth of filesystems 10 times above today's rates. Local performance will far exceed 100 Gigabits per second, and geographically distributed collaborations will want to achieve a large percentage of this bandwidth over the wide area.

Figure 12 projects HPSS data storage using well-established growth expectations for five years into the future.

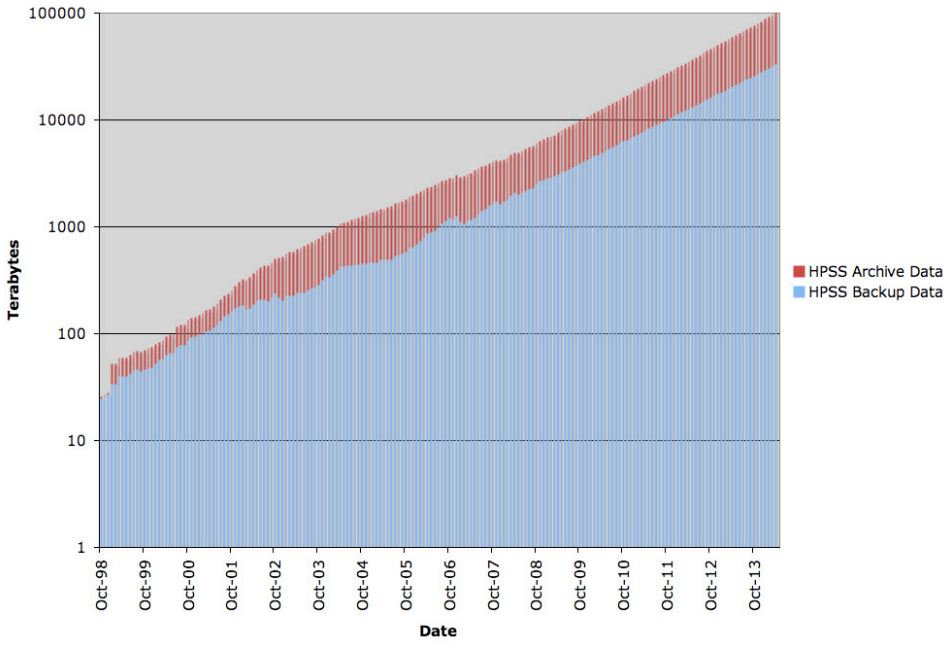


Figure 12: 2013 projection of total HPSS data stored (log scale)

5.5.2 Process of Science

The process of science for remote drivers appears to be the continuation of massive collaborations in the areas of high energy physics, climate and computational biology. These collaborations will require greater data sharing than has been observed in the past and will be a challenge to meet.

5.6 Beyond Five Years — Future Needs and Scientific Direction

NERSC has historical trending data that goes back ten years. This data has aided in reasonably accurate projections of storage and networking capacity and performance requirements. Figure 13 projects these requirements out ten years.

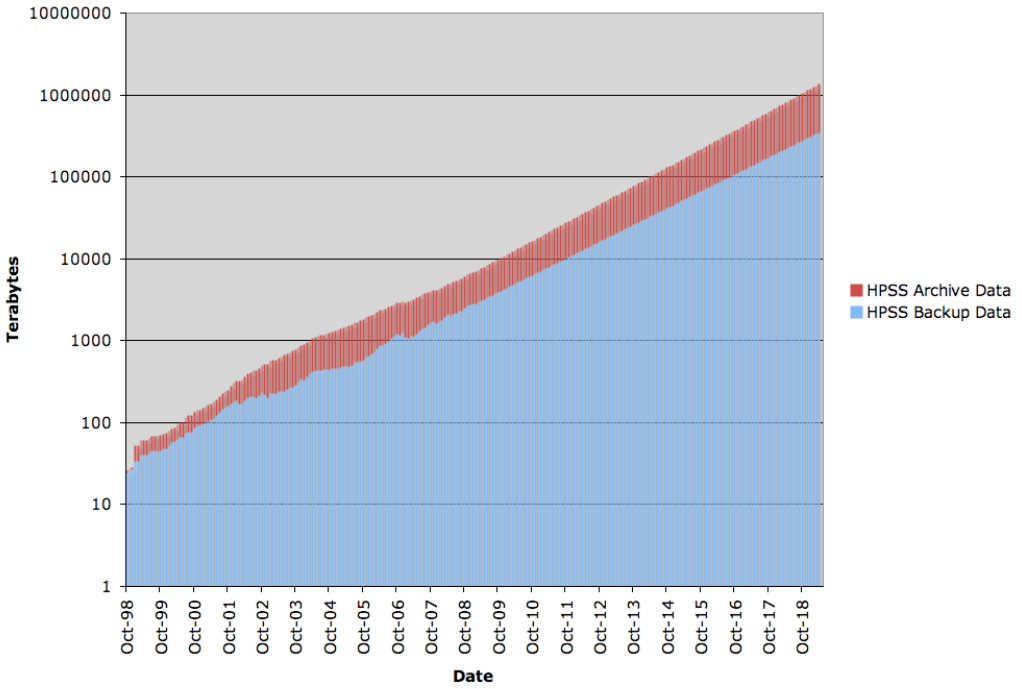


Figure 13: 2019 projection of total HPSS data stored (log scale)

5.7 Outstanding Issues and Recommendations

Bulk data transfer performance: NERSC would like to note that many of the bulk data transfer protocols used in transferring between our computation systems and storage resources are very sensitive to packet loss. It would be highly useful to be able to quickly observe expected network performance between two DOE sites connected by ESnet without having to run specialized tests to sample rates. Graphs or graphical tools would be ideal for trying to observe potential network saturation.

DOE-wide transport fabric for authentication: NERSC would like ESnet to provide a RADIUS based authentication fabric, which would enable the individual one time password (OTP) hardware token of a remote NERSC user to authenticate on NERSC systems. Many DOE sites are adding OTP authentication to their major computing systems. NERSC estimates that at least 50% of NERSC users will have a hardware authentication token provided by their home site. Unless these tokens can be used at NERSC, we may be forced to provide NERSC authentication tokens to every NERSC user. This approach is expensive as well as being a burden on the users, who must then carry multiple tokens, each with its own PIN.[1]

[1] "Secure, Extensible, Token Authentication for Department of Energy High Performance Computing," Matthew Andrews, Stephen Chan, Stephen Lau; email communication to Dave Goodwin, DOE.

Network advance reservation and co-scheduling: Fusion experiments would be enhanced if the data from one experiment could be transported to NERSC, analyzed, and the results returned in time (~10 min) to plan the next experiment. An advance reservation capability that would guarantee a minimum end-to-end bandwidth (EEB) as well as service separation/non-competition between the experiment's data flow and other network traffic such as bulk data and video is required. This capability could include label switched/lambda switched paths along with light path peerings with other networks/sites such as CERN and ITER.

Grid Certificate/PKI support: ESnet should continue to operate the NERSC DOEGrids Certificate Authority server, which permits the NERSC users who cannot obtain other DOE grid certificates to use NERSC grid resources. NERSC would also like the DOEGrids root certificate to be a root certificate trusted by major browsers (Internet Explorer, Mozilla and Firefox) so that these browsers will automatically trust DOEGrids certificates.

5.8 Summary

Table 7 summarizes the key science drivers and anticipated network requirements for NERSC in three time frames.

Table 7: NERSC network requirements

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• Large supercomputer center • Broad user base • Large HPSS storage system • NERSC6 system (~1PF) in 2009-2010	• Large data transfers requiring low packet loss • Non-TCP (e.g. UDP) data transport protocols	• Additional 10 Gbps link for special projects / dedicated bandwidth • Jumbo Frames • DoS mitigation • Native transport for non-IP traffic (e.g. Fibre Channel or InfiniBand)	• 20-40 Gbps capacity • Grid/PKI infrastructure • Network and computational co-scheduling • Dedicated WAN data transfer nodes • Distributed infrastructure for one-time passwords
2-5 years	• NERSC7 (Exaflop)	•	• 100 Gbps capacity	• 100 Gbps capacity • Consider HPSS-to-HPSS transfers/mirroring or redundancy with other DOE sites for improved Disaster protection • WAN global file system
5+ years	•	•	•	•

6 Oak Ridge Leadership Computing Facility (OLCF)

6.1 Background

Oak Ridge National Laboratory (ORNL) is a multi-program science and technology laboratory managed for the Department of Energy (DOE) by UT-Battelle, LLC. Scientists and engineers at ORNL conduct basic and applied research and development to create scientific knowledge and technological solutions that strengthen the nation's leadership in key areas of science; increase the availability of clean, abundant energy; restore and protect the environment; and contribute to national security. ORNL pioneers the development of new energy sources, technologies, and materials and the advancement of knowledge in the biological, chemical, computational, engineering, environmental, physical, and social sciences. The Oak Ridge Leadership Computing Facility (OLCF) at ORNL provides the most powerful computing services in the world for open scientific research.

6.2 Key Local Science Drivers

6.2.1 Instruments and Facilities

There are multiple compute systems at the OLCF. The largest is the 1.3 PF Cray XT5, with more than 150,000 cores. This system is based on the quad-core AMD Opteron, and more than 300 TB aggregate of physical memory. It is linked together using a proprietary high-speed interconnect. This system is in pre-production. Twenty-four "early science" projects have allocations of more than 500M hours aggregate. These will conclude in Summer 2009 as the system moves to a production role.

The second system is a 250 TF Cray XT4, also based on the quad-core AMD Opteron. This system was upgraded from a dual-core system to the quad-core socket in the first quarter of 2008. This system is currently the OLCF capability resource, pending release of the XT5.

There is a modest (80 node, 1280 core) quad-socket, quad-core AMD Opteron development system that provides a platform for application development as users move from single-core to multicore hardware platforms.

The OLCF is completing the deployment of a 10 PB Lustre file system, accessible from all major platforms within the Center. The system configuration allows for direct-connect and routed connections to the Cray XT5, and routed connections to all other systems. The storage subsystem is based on DDN 9900 controllers and 1 TB SATA drives. Benchmarks have pushed more than 170 GB/s to/from the filesystem.

Archive data moves across a Fibre Channel infrastructure. There are multiple Sun/STK SL8500 silos in separate facilities to support multi-copy and disaster recovery requirements. File systems such as the Lustre file system traverse DDR InfiniBand networks. The remainder of the traffic traverses 10 Gigabit Ethernet networks (internal and external).

The archive is based on HPSS. The total volume of data stored in the archive is greater than 5 PB. Growth rates in the total volume of data are severe. New data writes frequently exceed 30 TB/day. The incremental increase from 4 to 5 PB was seen in the three-month period ending in mid-April 2009.

Lens is a 32-node (quad-socket, quad-core, 512 cores) Linux cluster dedicated to data analysis and high-end visualization. Each node contains four quad-core 2.3 GHz AMD Opteron processors with 64 GB of memory, and 2 NVIDIA 8800 GTX GPUs. The primary purpose of Lens is to enable data analysis and visualization of simulation data generated on Jaguar (XT4/XT5) so as to provide a conduit for large-scale scientific discovery. Members of allocated Jaguar projects are automatically given accounts on Lens.

EVEREST (Exploratory Visualization Environment for REsearch in Science and Technology) is a large-scale venue for data exploration and analysis. EVEREST is 30 feet long by 8 feet tall. Its main feature is a 27-projector PowerWall with an aggregate pixel count of 35 million pixels. The projectors are arranged in a 9×3 array, each providing 3,500 lumens. Displaying 11,520 by 3,072 pixels, or a total of 35 million pixels, the wall offers a tremendous amount of visual detail. The wall is integrated with the rest of the computing center, creating a high-bandwidth data path between large-scale high-performance computing and large-scale data visualization. EVEREST provides a premier data analysis and visualization capability and facility in the Department of Energy's Office of Science.

EVEREST is controlled by a 14-node cluster. Each node contains four dual-core AMD Opteron processors. These 14 nodes have nVidia QuadroFX 3000G graphics cards connected to the projectors, providing a very-high-throughput visualization capability.

The OLCF network consists of a 2x10 GE bonded backbone with redundant uplinks to each of the ORNL border routers. A basic diagram of that network is shown in Figure 14.

GridFTP servers (Data Transfer Nodes, or DTN) are online. This is the method by which external users should move data to/from the facility. BBCP has been tested, and provides modest performance. GridFTP performance eclipses BBCP performance. SCP performance is very poor, as shown in Figure 15.

PerfSONAR network monitoring equipment is scheduled for deployment this quarter (Q3FY09).

ORNL has a dark fiber infrastructure that provides last mile fiber to Nashville (Qwest Metroplex), Atlanta, and Chicago. The current ESnet POP is in Nashville, TN (Sidco Drive).

There are network connections to TeraGrid, ESnet, and Internet2. There is a 10 Gbps ESnet connection to Nashville, and an OC-48 connection to ATL. There are near term plans (2Q09) to light one additional 10 Gbps connection to Nashville to support an SDN connection. This is scheduled for initial light in May 2009. ORNL anticipates implementing an SDN connection to initially support the ORNL-NERSC and ORNL-NOAA data ingest requirement. Other facilities have also expressed an interest in leveraging an SDN connection to ORNL.

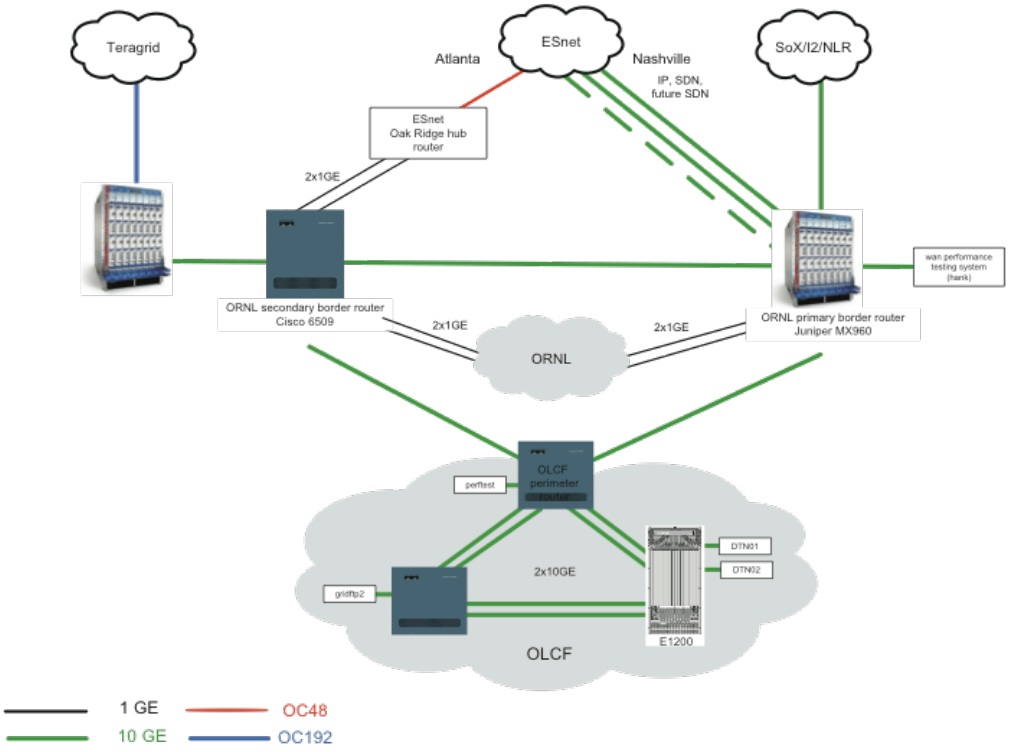


Figure 14: ORNL network overview

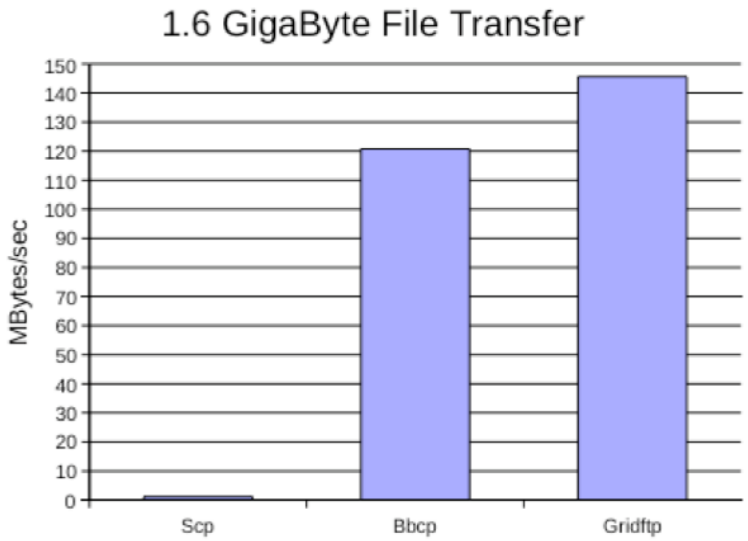


Figure 15: Performance comparison of SCP vs. BBCP vs. GridFTP

For the 2009 INCITE allocations, the user base is approximately 40 projects and 400 users.

The OLCF requires two-factor authentication. We allow users to use DOEGrids for our GridFTP servers. They are the only CA we trust at this time. Users must get proxy certificates from our server using two-factor authentication before they can be authorized for access to OLCF resources.

6.2.2 Process of Science

Leadership class computing at OLCF is delivering computation to a wide range of users in many communities, including DOE and NSF. One of the most difficult problems for users on our system is achieving performance in all areas of the application scientist workflow. Simulations can run from days to months and generate unprecedented amounts of data. The scientists typically work in small teams and need to collaborate by sharing the data. Not all of the data will be analyzed on the OLCF resources, and must be moved to other centers.

A sustained transfer rate of 100 Gbps translates to approximately 7.5 PB/week. For planning purposes, an operational requirement from the user community can be assumed to consume roughly 10% of this capacity (about 750 TB/week).

A few examples of the data that users need to move include:

1. The GTC (Gyrokinetic Toroidal Code — Fusion) code runs at OLCF and generates about 30 TB of analysis data on the 250 TF computer at OLCF (Jaguar). Estimates for data production in the next five years increase the total amount of data needed for analysis to increase to about 500 TB of analysis data for one week of runtime. This data needs to be moved to and from NERSC and ANL, where users can choose which of these facilities they will run on for analysis. In order to move this data in one week, GTC researchers need 8 Gbps sustained.
2. The GTS (Gyrokinetic Tokamak Simulation — Fusion) code generates data from all leadership class facilities, and generally writes about the same amount of data as the GTS code. Typically, a smaller portion of this data (roughly 1/10) needs to be moved to PPPL, the home institution of the GTS researchers. One-tenth is an estimate based on getting a smaller amount of the particle data, with a 1/sqrt(n) error in the PIC methods.
3. The S3D (Combustion/Chemistry) code currently generates about 50 TB of analysis data in one week on the petascale computers. This data is commonly moved to other leadership class facilities. Their team has been developing more complex analysis, and will see a bump in their data output in the near future (to about 300 TB/week). We anticipate that this data will move over to SNL in California, along with other LCFs.
4. The XGC-1 (Fusion) code has similar requirements to the S3D code currently, but since this is a full-F gyrokinetic code, we anticipate that the number of particles that will be generated and then analyzed from these codes will increase by 10x over the 50 TB/week to 500 TB/week. XGC-1 simulations use the OLCF/SDM workflow/monitoring system, and move all of the data from simulations from

NERSC to ORNL for later analysis. It is common for users to then move subsets of the data over to PPPL for smaller data analysis, along with data over to NYU.

5. Many codes on OLCF computers see the need for real-time analysis, and data needs to move in real-time for *in situ* analysis. We currently do this for several codes (GTC, GTS, XGC-1, Chimera, S3D). If we assume that the next generation computers can potentially have ten large-scale codes running simultaneously, then the data will be 500 TB (GTC), 500 TB (GTS), 500 TB (XGC-1), 300 TB (S3D). If all users want to move their analysis, this means they will require us to move 1.8 PB/week.
6. NOAA has an existing data transfer requirement to move 6-8 TB of data per day from OLCF to GFDL in Princeton. This is a current requirement. It is currently constrained by the disk I/O performance of the receiving systems at GFDL to 400-500 Mb/s. This constraint is being eliminated. NOAA uses an existing 1 Gbps connection to ESnet; this connection is being upgraded to 10 GE in the next 6-9 months. The data ingest requirement will grow substantially to the order of 40-50 TB/day in the next 12-18 months. They anticipate saturating the pending 10 GE connection very quickly, and having a strong need to move to 40 or 100 Gbps infrastructure as it becomes commercially viable/ available.
7. ESG is operating our current production and next-gen testbed nodes on the TeraGrid network. We expect the production next-gen to be integrated with OLCF. At a simple level, there are two phases to the next-gen ESG production. First, while sites around the world are generating data, we will be assembling a mirror of everyone's core data contributions. It looks like that will probably be ~600 TB, and will be built up over the course of months. It is not clear if we will draw directly from each international contributor or pull from the master site at LLNL. There will be perhaps 20 sites around the world contributing. We do not know all of them yet, but by this summer we will have an international testbed operating with sites in UK, Germany, and Japan as well as several US sites. The testbed will not be particularly taxing on the network. The second phase is after the data is generated and the collection is opened up to users to access. Our experience with the previous Inter-governmental Panel on Climate Change (IPCC) round leads us to expect that downloads will ramp up over time. A naive scaling of the current download rates to the expected data volumes for the new effort suggests the whole system will be delivering ~14 TB/day. This is across all of the gateways, probably more at LLNL than anywhere else if LLNL is the master site. If history is a guide, it will take a couple of years for download rates to build to this level.

There are other new/emerging programs that will create additional network requirements. Details related to the network requirements associated with these programs will emerge in the second half of FY09.

6.3 Key Remote Science Drivers

6.3.1 Instruments and Facilities

ORNL has redundant connections to ESnet. The primary connection is a dedicated 10 GE circuit from the primary ORNL border router to the ESnet backbone router in Nashville. The secondary connection is a 2 GE bonded channel from the ORNL secondary border router to the ESnet Oak Ridge hub router located at ORNL. The ESnet hub router connects to the ESnet backbone router in Atlanta via an OC48 circuit. The OLCF has redundant 10 GE connections to each of the ORNL border routers.

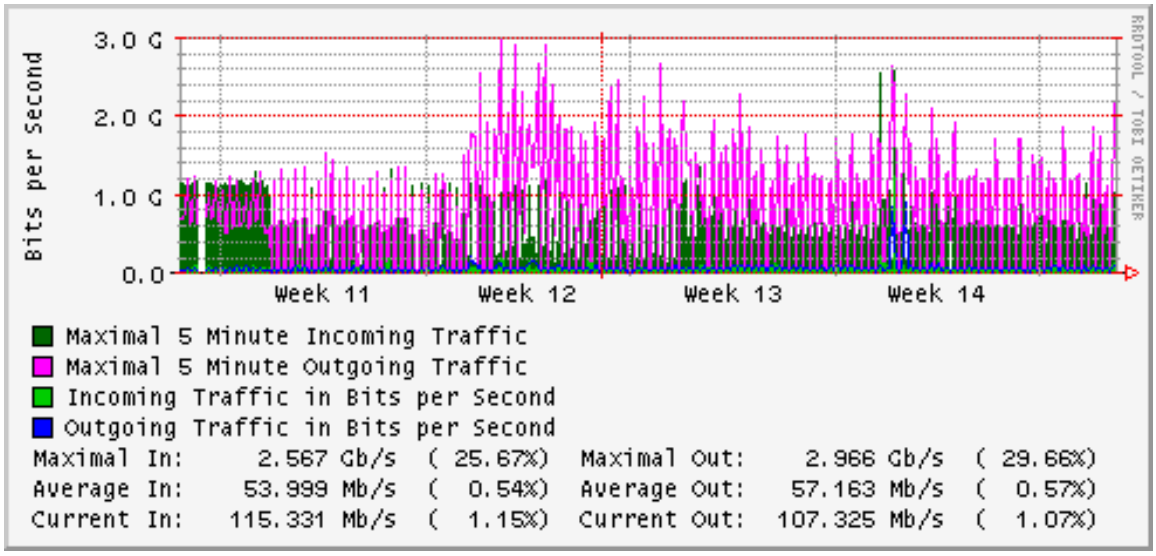


Figure 16: ESnet utilization since installation of Juniper MX960

6.4 Local Science Drivers — the Next Two to Five Years

6.4.1 Instruments and Facilities

The Cray XT5 will enter production in the late summer of 2009. It is anticipated to remain in production for approximately four years, subject to upgrade. The subsequent system, available in this next timeframe (two to five years), will be on the order of 20 PF, with commensurate file systems (O(50 PB)) and network infrastructure.

There are plans for an additional 100,000 square feet of computing space, with anticipated availability in the first half of CY2011.

6.5 Remote Science Drivers — the Next Two to Five Years

The NOAA data ingest requirement will grow substantially to the order of 40-50 TB/day within this period of time. They anticipate saturating the pending 10 GE connection very quickly, and having a strong need to move to 40 or 100 Gbps infrastructure as it becomes commercially viable/ available. In addition, there is a desire to move InfiniBand over Ethernet at DDR and QDR.

6.6 Beyond Five Years — Future Needs and Scientific Direction

(*No data provided*)

6.7 Outstanding Issues

The fiber optic footprint between the ORNL Nashville POP at the Qwest Metroplex and the ESnet Nashville POP at Sidco Drive is limited. Obtaining physical access to additional fiber between these locations remains an important part of any long-term solution.

6.8 Summary

Table 8 summarizes the key science drivers and anticipated network requirements for the OLCF in three time frames.

Table 8: OLCF network requirements

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**

Near-term (0-2 years)	• Cray XT5 (Jaguar)
	• Cray XT4
	• Other smaller computational systems
	• Visualization resources
	• 10PB Lustre filesystem
	• 10 Gbps Data Transfer Nodes with GridFTP
		• Fusion codes (GTC, XGC, GTS, S3D, etc)
		• NOAA data transfers
		• ESG/Climate data
			• Large global Lustre filesystem (170GB/sec aggregate)
			• 30TB/day HPSS writes
				• 10 Gbps total WAN data transfer (750TB/week)
				  • 30TB/week GTC data
				  • 3TB/week GTS data
				  • 50TB/week S3D data
				  • 50TB/week XGC-1 data
				  • 8TB/day NOAA data
				• PerfSONAR (Q3 2009)
				• DOEGrids CA for GridFTP auth, 2-factor auth for proxy certificates

2-5 years	• 20 Petaflop computer with 50PB filesystem
	• Additional space to house additional resources (CY 2011)
		• ESG next-gen in full production
		• NOAA data transfers
			•	• 100 Gbps WAN link
				  • 50TB/day NOAA data
				  • 14TB/day ESG data
				• WAN InfiniBand
				• Remote visualization and computational steering

5+ years	•	•	•	•

7 Open Science Grid — A Virtual Facility

7.1 Background

The Open Science Grid (OSG) is a virtual facility jointly funded by DOE and NSF. Its primary goal is to provide the organizational framework, middleware stack, and operational support for a national distributed computing infrastructure for high throughput computing in support of the broadest possible scientific community within DOE and NSF. OSG accomplishes this via a consortium of scientific and engineering communities and IT departments at universities and national labs, in addition to the funded OSG project.

Some of our science stakeholders are large international scientific collaborations that operate scientific instruments both within the US as well as abroad. These scientific communities depend on OSG to maintain global interoperability and federated trust across independent cyberinfrastructures, and to negotiate and manage the changing landscape of deployed middleware within this global context for years and even decades to come.

7.2 Key Science Drivers

7.2.1 Instruments and Facilities

The OSG infrastructure today spans roughly 70 active sites within the US and a few in South America. A site refers to a compute and storage cluster accessible via a set of OSG APIs for advertising the state and capabilities of the site, moving jobs and data in and out, and accounting for the consumed resources.

The sites on OSG include many of the DOE national labs, some of the NSF-funded supercomputing centers, and many university infrastructures. Resources available at these sites range from a few CPUs connected via a single Gbps WAN link, to more than ten thousand CPUs accessing more than ten Petabytes of storage, connected via multiple 10 Gbps WAN links. Few of the sites are capable of having sustained WAN I/O in excess of 10 Gbps; a couple tens of sites routinely sustain 2-5 Gbps of WAN I/O; and peak LAN I/O is typically a factor ten larger than peak WAN I/O.

The major scientific instruments connected to OSG are the Tevatron experiments at FNAL, the RHIC experiments at BNL, the LHC experiments at CERN, and the LIGO interferometers in Hanford, Washington and Livingston, Louisiana.

7.2.2 Process of Science

Science on OSG ranges from small, individual investigators to large scientific communities with thousands of members. The large communities bring their own compute and storage resources into the OSG and make them available for opportunistic use to a wide range of science, big as well as small. Large communities furthermore operate their own community specific middleware stack on top of the OSG stack. OSG thus has a science domain specific look and feel via this community specific software layer. Large communities include experiments in high energy, nuclear, and astrophysics

but also campus and regional grid infrastructures. What defines a community can thus be a common organizational context instead of or in addition to a common science goal.

As smaller communities have neither the budget nor the expertise to develop, operate, and support their own middleware stack, the so called Engagement Virtual Organization (VO), a project funded by NSF, operates a middleware stack for these communities, and supports porting of legacy applications onto the OSG. At present, the Engagement VO is the single largest user of OSG resources apart from the Tevatron and LHC experiments.

In addition, OSG provides advice and support to scientific communities that want to develop and operate their own domain specific middleware stack. Such support may entail anything from training, help in debugging deployment and operations problems in networking, storage, workflow design, etc. In this context, OSG brings together computer science researchers, middleware developers, and scientific communities, helping them to turn ideas into production quality middleware that is widely deployed and used.

The science supported by OSG includes astrophysics, biology, bioinformatics, chemistry, computer science, economics, geophysical sciences, genetics, high energy physics, industrial engineering, library sciences, mathematics, mechanical engineering, meteorology, nanotechnology, neurobiology, nuclear physics, psychology, renewable energy, and more. In 2008, the OSG had 2000 users, and scheduled typically 300,000 to 500,000 individual jobs per day. Those jobs are submitted mostly via pre-WS Globus GRAM, and to a lesser extent via WS-GRAM. The primary storage API is SRM on top of GridFTP. We support multiple implementations for both protocols. Some of our largest sites require in addition FTS on top of SRM. The LAN transfer landscape is more heterogeneous, including dcap, Xrootd, Fuse-hadoop, Lustre, GPFS, NFS, and more.

A typical scientific workflow might include the following:

- Installation of application libraries and related software
- Stage-in of data to process
- Job submission, data processing, result stage-out to local storage
- Result post-processing
- Result retrieval to archival storage at home institution.

7.3 Key Remote Science Drivers

At present, the key remote science drivers are the LHC physics program and LIGO, followed by the Tevatron experiments and RHIC. The LHC physics program includes both high energy and nuclear physics. In high energy physics, the DOE supports the ATLAS and CMS experiments. In nuclear physics, the DOE supports heavy ion physics in the ALICE, Atlas, and CMS experiments. DOE High Energy Physics supports the Tevatron physics program, while DOE Nuclear Physics supports the RHIC physics program.

7.4 Instruments and Facilities

The dominant consumers of network bandwidth on OSG today are undoubtedly the two LHC experiments, ATLAS and CMS. Both operate archival storage, so called Tier-1

sites, at the DOE national laboratories BNL and FNAL respectively. About one-third of all the sites on OSG are affiliated with ATLAS or CMS.

DOE national labs play a very significant role in exploiting the physics of the LHC. Among the ATLAS and CMS member institutions are LBNL, LLNL, SLAC, FNAL, ANL, and BNL. At present, LBNL (i.e., NERSC), SLAC, FNAL, BNL, and ORNL have established sites on OSG. LLNL is in discussions with OSG in the context of establishing an ALICE presence in the US. ALICE is a heavy ion physics experiment at the LHC. ANL is collaborating with OSG via the Globus Consortium. ANL and LLNL are member institutions of ATLAS and CMS respectively, and heavily involved in software development for the two large experiments but do not provide computing resources for the experiments at this point in time.

7.4.1 Process of Science

Data from the LHC is archived at Tier-1 centers worldwide. ATLAS and CMS have established such centers in the US. ALICE is in the process of doing so. Each Tier-1 is responsible for the processing of its datasets. Physics analysis of these datasets is then performed at the Tier-2 and Tier-3 centers. Today, the latter have disk caches of up to 0.5 Petabytes each. In addition, Monte Carlo simulations are produced at the Tier-2 and Tier-3 centers, and archived at the Tier-1. In general, the largest data flows are from Tier-1 to Tier-2. At present, all Tier-2 have at least one 10 Gbps WAN link, and roughly half of them have demonstrated that they can sink data at rates that fill at least 80% of this link.

We believe that the basic process of science for the LHC is likely to be emulated by other scientific domains in the future. The LHC experiments are presently a few years ahead of others on OSG in their ability to manage large volumes of data, move them around the globe, and analyze them across a widely distributed set of computing resources. We believe that the crucial missing technology for others to follow is an easy to operate and use data management and movement middleware stack. The LHC experiments have invested a significant amount of effort developing expertise in this area over the last five to ten years. It is one of the goals of the OSG to help spread this expertise to a wider audience. Initial steps in the right direction have been made in the context of leased storage as part of the SRM v2 protocol deployed on OSG. However, no end-to-end solution exists outside the LHC experiments.

7.5 ESnet Software and Services

OSG depends crucially on the CA services from ESnet. Most of our sites deploy host certificates from DOEGrids, and our documentation directs scientists to DOEGrids as the CA for obtaining personal certificates. We thus crucially depend on DOEGrids.

In addition, we make heavy use of ESnet audio and video conferencing. All our phone meetings use ESnet MeetingPlace, and our videoconferences are generally done via the ESnet Ad-Hoc H.323 service. We are very happy with both of these services from ESnet. Life without them seems hard to imagine.

We have started to deploy perfSONAR, and expect this to become a standard tool across the OSG infrastructure.

7.6 Beyond Five Years — Future Needs and Scientific Direction

As part of the ASCR Exascale workshops, the High Energy Physics community has written a detailed planning document looking towards the future up to 2020. It predicts data volumes at the Exabyte scale, and corresponding needs in data movement capabilities. While we expect other scientific communities with increasing data movement needs, it is likely that High Energy Physics continues to be the largest user of network bandwidth for the foreseeable future. We thus refer to that document for further detail.

7.7 Outstanding Issues

We are concerned about two major challenges. Short term, we are lacking easy to operate and use end-to-end data management and movement infrastructure for scientific domains other than High Energy Physics. Longer term, our biggest concern is the fact that the large experimental facilities have planning horizons measured in decades, while typical software lifecycles are measured in years. How to manage the middleware lifecycle in a globally federated environment is an unsolved problem that poses significant risk to large scientific undertakings.

In addition, we are generally concerned about the ability of ESnet and Internet2 to stay ahead of the networking needs of our stakeholders. This was spelled out in some detail for High Energy Physics in the ASCR Exascale workshop document, and we include an excerpt below.

> *HEP has undergone a revolutionary paradigm shift within the last 5 years which depends completely on highly capable, scalable, high speed, highly interconnected, and very reliable networks. Within the US, two DOE Office of Science Labs -- Fermilab and Brookhaven -- are the "Tier 1" repositories of the data from the LHC. The amount of data flowing first from CERN via USLHCNet, then via ESnet to the tier 1 centers, and then from there to the mostly university-based data analysis ("Tier 2" and "Tier 3") sites, produces several orders of magnitude more network traffic than any past science use of WAN networking. Even during the testing phase of the LHC data handling systems, network traffic was generated at the rate of 4.5-9 Gbps, sustained 24 hours a day for several months, with peaks in the 15-20 Gbps range. Networks supporting the distribution of the datasets produced to the Tier2s have been of comparable size, often reaching full use of a 10 Gbps link in the case of several of the US Tier2 sites.*
>
> *To support this sort of next generation large-scale science, the Office of Advanced Scientific Computing Research in the Office of Science funded ESnet to design and build a completely new network with a new architecture specifically tailored for science like that of HEP. The new network -- ESnet4 -- was based on use-cases and requirements that were identified in 2003-2005. The network took about 18 months to build and now provides about 20 Gbps throughout the US, connecting the Labs to other US and international Research and Education networks. The current network is designed and funded to grow to about 50 Gbps in late 2010.*

The currently scoped networks will handle the load probably for the next several years (~2010). Beyond this, the network capacity will have to increase considerably over the original plan. This expansion cannot happen by just adding many more 10 Gbps optical circuits for two reasons. First, the cost would be prohibitive. Second, the optical network infrastructure that ESnet is built on is shared on a dedicated optical fiber with Internet2 (US R&E network). By 2010 this optical infrastructure will be approaching its capacity and it is not practical to obtain a second complete set of fibers around the country.

In order to increase the capacity of the network until the next generation network is built (in the 2015-2017 timeframe) several new approaches are needed and these will require research, development, and deployment. The current most promising approaches are 1) the dynamic management of optical circuits thus allowing their integration with the user transport layers of the network; 2) increasing the current 10 Gbps per optical circuit to 100 Gbps per circuit capacity, and; 3) the transparent, selective, and dynamic re-routing of in-transit data flows from one part of the network to another. All of these technologies are designed to maximize the use of the entire current optical fiber infrastructure. Another topic that is important for the effective utilization of the network by the science community is highly capable, "universally" deployed, end-to-end (user application to user application) network monitoring across all of the intervening network domains (e.g. ESnet, Internet2, GÉANT, etc.).

7.8 Summary

Table 9 summarizes the key science drivers and anticipated network requirements for the OSG.

Table 9: OSG network requirements

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• Several, depending on OSG user	• Varies, based on OSG user community	• Up to 4 x 10 Gbps	• Up to 4 x 10 Gbps network capacity
2-5 years	•	•	•	• 100+ Gbps network capacity
5+ years	•	•	•	

8 Earth System Grid

8.1 Background

The Earth System Grid Center for Enabling Technologies (ESG-CET) is about building a *science gateway* that enables a *researcher-friendly* infrastructure that provides distributed access to petabytes of observation and climate simulation data. This next generation problem-solving environment will allow access to distributed federated data, information, models, analysis, visualization tools, and computational resources. It is based on the Internet and the need to access and analyze large-scale data anywhere in the world. The scale of data that needs to be analyzed will require that many analysis functions should be performed near the data, therefore reducing the need for network traffic. However, there will still be a need to move vast amounts of data to and from sites for scientific purposes. For example, large-scale data that reside at various sites around the world will need to be assembled at one location for ensembles or multiple model data intercomparisons. Data replication for data backup and additional access also argues the case for large-scale data movement.

Because ESG-CET is a distributed data repository for vast amounts of climate related data and products, it has (and will have in the future) tens of thousands of users accessing it from various forms of user interfaces (e.g., web browsers, direct file transfers with transfer tools, and climate analysis tools). The multitude of users requesting tens of terabytes of data simultaneously can be daunting and can greatly impact the network performance of ESnet and other potential ESnet science users.

8.2 ESG-CET Architecture

The new ESG-CET architecture is based on three tiers of data services (Figure 17). The three levels of data services are Tier 1—Global; Tier 2—Gateway; and Tier 3—Data Node. Three ESG Gateways are planned initially, at LLNL, ORNL, and NCAR. The figure also shows where data users and data providers gain access. The three tiers are further described below:

- **Tier 1—Global metadata services for search and discovery:** comprises a set of services providing shared functionality across the worldwide ESG-CET federation. These services include user registration and management, security services and access control, metadata services for describing and searching the data, notification and registry, and global monitoring. All ESG-CET sites share a common database, so that a user only has to register once in order to access resources across the whole system and can find data throughout the federation, independent of the site at which a search is launched.
- **Tier 2—Data gateways as data-request brokers:** comprises a limited number of ESG Data Gateways which act as brokers handling data requests to serve specific user communities. Services deployed on a Gateway include the user interface for searching and browsing metadata, for requesting data (including analysis and visualization) products, and for orchestrating complex workflows. Gateways will be operated directly by ESG-CET engineering staff.

- **Tier 3—ESG nodes with actual data holdings and metadata accessing services:** includes the actual data holdings and reside on a (potentially large) number of federated ESG nodes, which host those data and metadata services needed to publish data onto ESG and execute data-product requests through an ESG Gateway. Personnel at local institutions will operate ESG Nodes. A single ESG Gateway serves data requests to many associated ESG nodes: for example, more than 20 institutions are expected to set up ESG data nodes as part of the Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment.

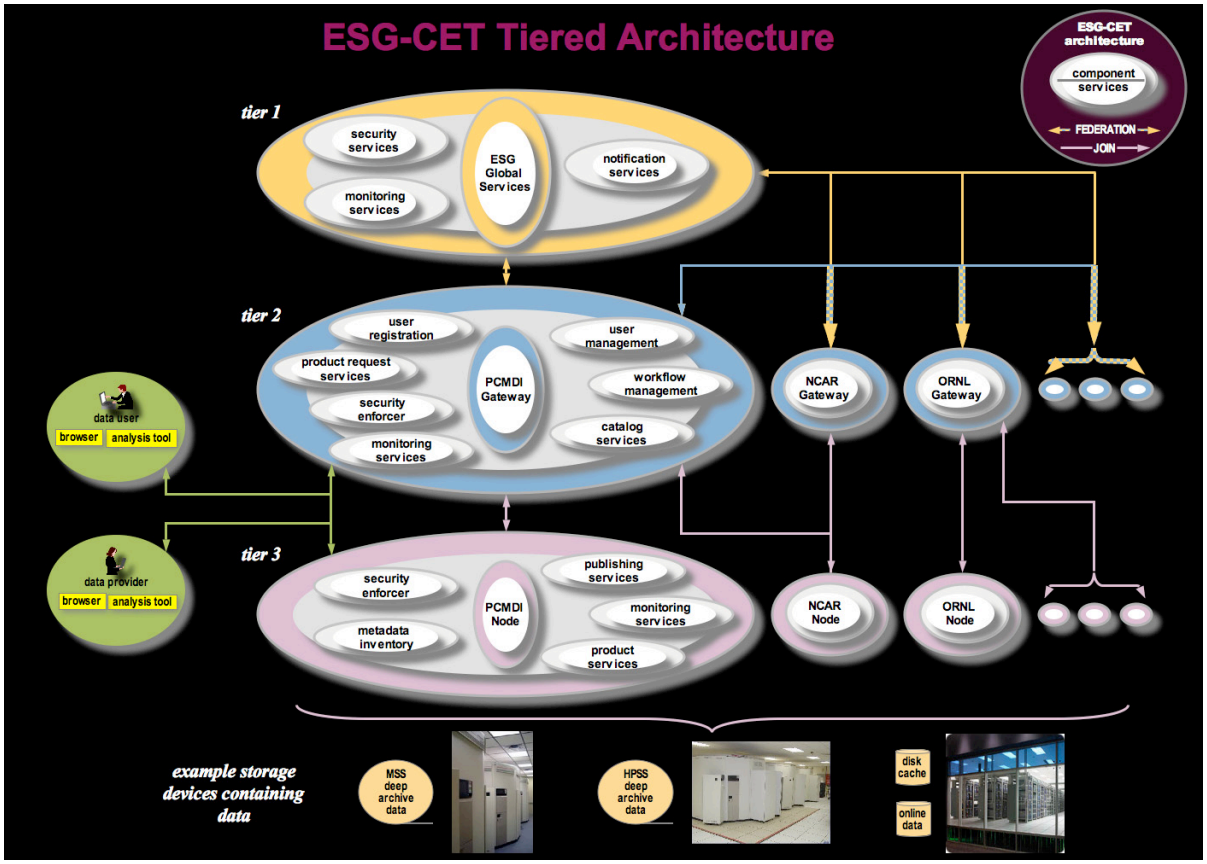


Figure 17: ESG-CET tiered architecture

8.3 Key Remote Science Drivers

8.3.1 Instruments and Facilities

ESG-CET climate data nodes typically will have 100 to 1,000 TB of disk storage and multiple systems with 10 Gbps and 100 Gbps network connection each. Data nodes are typically Linux front-end systems running a minimal software stack consisting of transfer server(s) (e.g., HTTP, GridFTP, OPeNDAP), Postgres (relational database), MyProxy server, THREDDS data server (describes datasets as a collection of XML catalogs), and a customized ESG data publisher. Data storage can be either on tape (e.g., HPSS, MSS) or

rotating disks. For example, LLNL has a 1 PB Sun Fire x4600 (8 Dual-Core AMD Opteron — Model 885) running Solaris 10x64. This is a scalable storage rotating disk system. In the near future, this is expected to grow to 2 PB and the core data will be archived at NERSC's deep storage.

8.3.2 Process of Science

A sketch of what we expect the future ESG-CET/ESnet network needs to look like is shown in Figure 18. We describe below three use cases of large-scale data movement between nodes to one location for ensemble and/or model inter-comparison, mirroring a core dataset, and the use case of multiple large-scale data access by thousands of users.

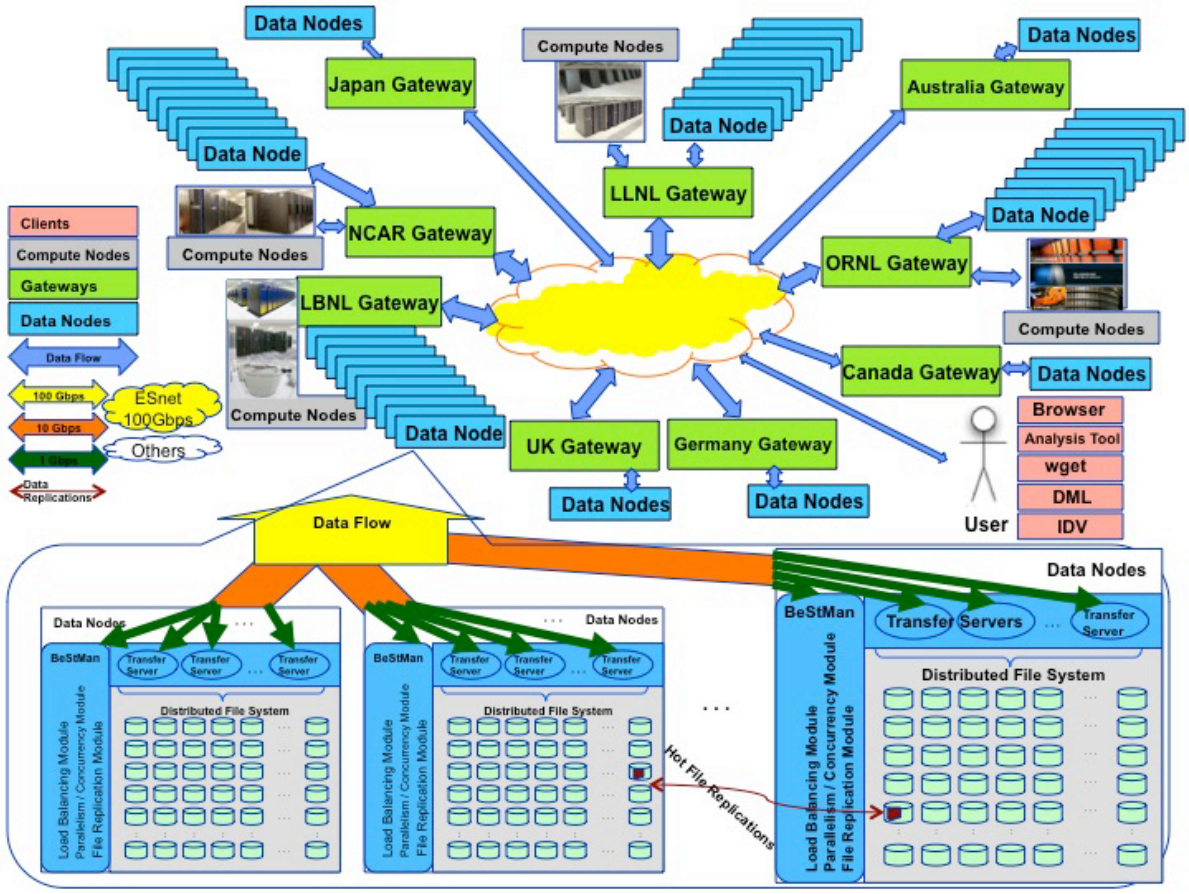


Figure 18: The envisioned federated topology of the ESG-CET enterprise system utilizing 100 Gigabits per second (Gbps) network connections. A network of geographically distributed Gateways and Data Nodes is built into a globally federated "built-to-share" scientific discovery infrastructure. By federating these Gateways using a fast network, independent data warehouses deliver seamless access to vast data archives to scientists and their specialized client applications. Experts (e.g., model developers, climate researchers) and non-experts alike need fault-tolerant end-to-end system integration and large data movement, and benefit from rich data exploration and manipulation — in the process moving vast amounts of data to and from sites around the world.

8.3.3 Scenario 1: Generation of Data and Replication and Movement

- Data generation sites, with computers specialized for running models: LLNL, NCAR, Japan, UK, Germany, ORNL, Australia, and Canada.
- Authorized user logs onto LLNL's Gateway and issues a request to collect large-scale data from multiple source nodes (i.e., LLNL, NCAR, Japan, UK, Germany, ORNL, Australia, and Canada) in order to generate a temperature ensemble of the global models.
- The target node initiates the data transfer by pulling data from the multiple source nodes.
- A requestID is returned to the user, and transfer starts asynchronously.
- The user can check status of transfer requests with requestID.
- Size: approximately 1 PB, depending on the data set

8.3.4 Scenario 2: Mirroring a Core Dataset

- Core dataset needs to be mirrored from source node to target nodes.
- For example, BADC node (Germany) needs to mirror all or part of core dataset from LLNL node.
- A requestID is returned, and transfer start asynchronously at a target node by pulling data from the source node.
- The user can check status of transfer request using requestID
- Size: approximately 1.2 PB at the maximum depending on what target node requests to mirror.

8.3.5 Scenario 3: Many Users Access Many Data Subsets

- Thousands of users log onto the LLNL Gateway to search and browse data.
- They simultaneously request data subsets consisting of hundreds of thousands of files located on the LLNL data node.
- Transferable URLs are returned to each user and users start transfers concurrently.
- The given size that any one user can access and download is approximately 10 TB.
- The system manages the I/O requests in parallel and balances loads on transfer servers.

8.4 Remote Science Drivers — the Next Two to Five Years

To enable the climate scientists executing these activities, ESG-CET is producing an infrastructure and toolkit for the next-generation ESG architecture, which is based on federating dozens of archive and Gateway sites around the world. This federation will provide access to vast climate data holdings; in aggregate these collections will comprise tens of petabytes of information. Both raw and processed data will be obtainable using the web and local analysis clients. Remote processing will be available, including subsetting, concatenation, re-gridding, and filtering. ESG-CET will also provide tools for common analysis and intercomparison procedures.

The Center's current research program is aimed primarily at scaling the ESG to meet the challenges facing its primary stakeholders, who are engaged in complex climate science activities, including:

- The Coupled Model Intercomparison Project, Phase 5 (CMIP5) for scientists contributing to the IPCC Fifth Assessment Report (AR5)
- The development of the Community Climate System Model (CCSM)
- The SciDAC-2 climate application entitled A Scalable and Extensible Earth System Model for Climate Change Science
- The Computational Climate End Station (CCES) at the ORNL Leadership Computing Facility
- The Global Organization of Earth System Science Portals (GO-ESSP)
- The North American Regional Climate Change Assessment Program (NARCCAP)
- The Parallel Ocean Program (POP) at Los Alamos National Laboratory
- Other wide-ranging climate model evaluation activities.

8.5 Beyond Five Years — Future Needs and Scientific Direction

Climate model datasets are growing at a faster rate than the dataset size for any other field of science. Based on current growth rates, these datasets will be hundreds of exabytes by 2020. To provide the international climate community with timely access to these data in order to maximize scientific productivity, these data will need to be replicated and cached at multiple locations around the globe. Unfortunately, establishing and managing a distributed data system presents several significant challenges not only to system architectures and application development, but also to the existing wide area and campus networking infrastructures. For example, transport technologies currently deployed in wide area networks do not cost-effectively scale to meet the scientific community's projected aggregate capacity requirements based on the growth rates for dataset size. Even if backbone network technology improvements increase link speeds from the current 10 Gigabits per second to 100 Gigabits per second and are in production service by 2012, as anticipated, more efficient use of networking resources will be essential. Efforts are underway to develop hybrid networks with dynamic virtual circuit capabilities, such as those currently being deployed by research and education networks like ESnet that have active network research and development activities. Although dynamic virtual circuits allow high-capacity links between storage and computer facilities to be created as needed and then deactivated quickly to free up network capacity for other users, much work is still required to optimize and harden the software.

8.6 Outstanding Issues

Timely and efficient data transport across wide area networks is not the only networking challenge to be faced over the next five to ten years. In the use case mentioned above, the policymaker downloads 1 PB of data to their site. To achieve fast and efficient

downloads and better performance for the ordinary user, training is key. Often there is too little staff or outreach for this effort—leaving the user of the system frustrated.

8.7 Summary

Table 10 summarizes the key science drivers and anticipated network requirements for the ESG.

Table 10: ESG network requirements

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• ESG data node at LLNL and LBNL and scalable Linux front-end with 10 x 10 Gbps connections	• CMIP5 (IPCC AR5) data access	• 1 GigE/ 10 GigE	• 10 x 10 Gbps network connection
2-5 years	• Extend ESG data nodes out to NCAR, ORNL, ANL, and LANL with 100 Gbps connection	• CCSM and CCES data access	• Expand use of 10 GigE	• 100+ Gbps network connection
5+ years	• Extend ESG data nodes to EU and Asian partners with all US network connections to 10 x 100 Gbps	• CMIP6 (IPCC AR6) and other climate simulation and observation data	• Expand use of 100 GigE	• 10 x 100 Gbps network connection

9 Findings

The following issues were reported and discussed at the workshop.

Long-Term Planning

Due to the compute cycle allocation process, the supercomputer centers at LBNL, ANL, and ORNL have no way to know who their users will be in the future, and so talking to current users about future needs may or may not be useful. However, NERSC has many years of data that shows that trend analysis from the past is a very good predictor of future needs in terms of data volume.

Long-term software maintenance was of concern for many attendees. Scientists rely heavily on a large deployed base of software that does not have a secure long-term funding model. Such software is typically maintained through short-term funding for the development of additional features, but stable program-level funding for many of the software tools on which DOE scientists rely is currently lacking. Software packages for which this is true include data transfer tools such as GridFTP as well as identity management and other software infrastructure that forms a critical part of the Open Science Grid and the Earth System Grid.

Cross-Site Authentication

Currently users who need to transfer data between ALCF and OLCF need to use two different one-time password (OTP) devices, one for each end of the transfer. This is a headache for users, and causes access and compatibility problems for scientists who need to move data between the supercomputer centers, or utilize resources at multiple supercomputer centers. Today this mainly affects data transfers, but in the future it will affect coupled jobs and may make automated data placement services more difficult to deploy. To help characterize the problem, a suggestion was made that the three supercomputer centers generate a white paper explaining how each site currently does authentication.

One solution for this problem is to build a system that would allow the supercomputer centers to authenticate a user against the OTP server at the user's home institution, rather than against an OTP server managed by the supercomputer center. This would allow supercomputer centers to use the OTP tokens that users already have, rather than issuing a new OTP token to each user for the user's account at each supercomputer center. A proof-of-concept system for accomplishing this was demonstrated several years ago, but a production service was not funded. A request to fund and deploy such a system was made at the workshop. A contemporary solution using SAML-based services, such as a Shibboleth federation, would address this need. Such a federation already exists in the US university space (InCommon), and an effort to develop and deploy a compatible federation of the DOE science laboratories (the Science Identity Federation) is currently underway. In addition to the deployment of the Science Identity Federation, the data transfer and data management tools on which the scientists rely need to be updated so as to be compatible with contemporary identity and trust federation protocols.

Data Movement Services / End-to-End Performance

Creation of a working group to address end-to-end network performance issues is very important. The working group needs to include people representing the site LANs, site storage systems, and ESnet. A successful example of such a working group, the Data Transfer Working Group or DTWG, was established in early 2009 and included NERSC, ORNL, and ESnet staff. The DTWG has proven invaluable in establishing dedicated high-performance data transfer systems at NERSC and ORNL. ANL has joined this working group and is currently building dedicated systems to support high-performance wide area data transfers. A key aspect to maintaining WAN performance on these dedicated systems is regular test and measurement of real disk-to-disk data transfers between sites. This provides users with an expectation of performance, and provides supercomputer center staff with early notification of problems if they should arise. One of the significant benefits of the DTWG has been collaboration on host and network performance tuning — a great many of the difficulties encountered during the setup and deployment of the data transfer servers were related to host and site network performance tuning. The knowledge gained from these experiences has been added to ESnet's performance tuning site http://fasterdata.es.net/.

GridFTP is available on the dedicated transfer systems, but several users and projects prefer to use BBCP due to its greater ease of use (BBCP data transfers can be authenticated with the same credentials as are used for SSH login, and so users do not need to get Grid certificates as they do for GridFTP). GridFTP supports SSH authentication in the current version, but the GridFTP installations at several centers are older and do not currently support this authentication mode. There was a discussion at the workshop about some of the features that might be beneficially added to GridFTP. One feature that was deemed important was the support of rsync semantics (i.e., data set mirroring). SSH typically performs poorly as a wide area data transfer tool, and one of the common reasons scientists move data over SSH is for rsync (rsync uses SSH for transport in most cases).

Ease of use is a key determinant of the scientific utility of network-based services. Users will do the thing that is easy for them to do, even if it might perform less well than some other more complex solution. Therefore, a key enabling aspect for scientists' beneficial use of high performance networks is a widely deployed, well-maintained toolset at the network endpoints that allows scientists to easily utilize the services the network provides. Ease of use of advanced services brings both increased network utilization and enhanced scientific productivity.

Tools that increase network visibility are helpful to users in determining whether performance problems are due to network issues such as packet loss and congestion, or due to some other factors (e.g., host configuration, system load). A request was made for tools that will provide increased visibility into network state. The perfSONAR infrastructure is emerging as the standard platform for deploying such tools and services, and ESnet is active in the development and deployment of perfSONAR.

Identity Management Issues

Several issues were identified related to identity management, including future support of OpenID, integration with Shibboleth and other federations, credential lifetime, policy enforcement (e.g., password length, one time password authentication requirements), identity management system interoperability, and others. It was decided that ESnet would host a workshop on this topic.

Connectivity and Bandwidth

The following paths were identified as having particularly high bandwidth requirements. For increased bandwidth management capabilities, SDN circuits for each of these, is desirable as well.

- ORNL↔NERSC
- ORNL→GFDL/Princeton (for NOAA data, including wide-area Infiniband)
- ORNL→ANL (GTC data)
- PCMDI (LLNL)↔ORNL (Climate data)
- PCMDI↔NCAR (Climate data)

Some other connectivity-related items of note include:

- ANL has direct connectivity to Internet2 — therefore traffic between Internet2 sites and the ALCF will not normally transit ESnet
- ORNL has direct connectivity to NLR — therefore OLCF traffic to NLR sites will not transit ESnet

The top non-DOE sites for ESG/Climate Data include:

- NCAR
- TACC at UT Austin
- UC Davis
- University of Michigan

Co-Scheduling Services

The idea of co-scheduling of compute, network, and storage resources has been discussed by CS researchers for a long time. However, none of the centers have any current users who are currently requesting co-scheduling. Despite this, a number of people still think this might be a requirement in the future. For example, the fusion community, which is a large consumer of supercomputer cycles, proposes coupled models (which would require co-scheduling of resources at multiple supercomputer centers) and coupled simulation and experiment (which would require co-scheduling computational resources and experimental facilities) as part of the Fusion Simulation Project (FSP). In order to link these resources together, co-scheduling of the network is also a likely requirement.

Collaboration Services

The OSG community depends heavily on the certificate infrastructure supported by the DOEGrids Certificate Authority (CA) run by ESnet. The supercomputer centers also rely on ESnet CAs. OSG also views the ESnet Collaboration Services (ECS — audio and video conferencing provided by ESnet to the science community) as indispensable. In

contrast, the supercomputer centers seemed largely unaware of the ESnet supported audio and video conferencing tools. More outreach in this area is needed.

Future Needs for Remote Steering

ANL predicts that as supercomputers' CPU speeds continue to increase, the need for remote steering of simulations will become more important. This will likely require guaranteed bandwidth services to many more destinations (e.g., the home institutions of the scientists).

10 Requirements Summary and Conclusions

Authentication, user access, and security issues were of significant concern for all attendees. While this is not something that ESnet can address directly, consistent authorization and access control policies have a significant impact on whether the users of the supercomputer centers can effectively use the network that interconnects the supercomputer centers. Another aspect of this is identity management. It was clear at the workshop that continued collaboration between ESnet and the supercomputer centers is needed in order to balance authentication, access, security, and related issues.

The supercomputer centers expect to need 100 Gigabit network connectivity as soon as it is financially viable to deploy the technology, as do other communities such as OSG/HEP and the Earth System Grid. The continued growth of scientific data sets is at least in part driven by the continued growth in scale of the computational resources available to the science community. In addition, several science communities and the supercomputer centers themselves seek dedicated bandwidth between major resources. ESnet will work with the supercomputer centers and others to deploy these circuits.

The need for high-performance, easy-to-use data movement services has been highlighted in previous network requirements workshops. The supercomputer centers have made significant progress in this regard due to a collaborative effort called the Data Transfer Working Group. Part of this success is due to the enhanced diagnosis and troubleshooting capabilities available through the perfSONAR infrastructure. PerfSONAR is being actively deployed by ESnet, the DOE supercomputer centers, and many other sites and networks around the world. However, more work is needed to develop and deploy data transfer tools that are easy for scientists to use, reliable, high-performance, and robust. Good data transfer tools are a critical part of the infrastructure that enables scientists to effectively use the network to move data and collaborate with optimal efficiency.

ESnet Collaboration Services (ECS) and ESnet's identity and trust management services are very important to several science communities.

Long-term maintenance of important software tools in the form of long-term funding was discussed as a major concern for many attendees. This is a topic that will need to be addressed further in the future.

Action Items

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

- ESnet will host an Identity Management Workshop with OSG
- ESnet will work with sites to set up the following SDN circuits:
    - Between ANL, NERSC and ORNL in support of bulk data transfers
    - ORNL to GFDL in support of Climate Research
    - Between LLNL, ORNL and NCAR in support of ESG
- ESnet will continue to assist sites with perfSONAR deployments

- ESnet will continue to assist sites with network and system performance tuning

In addition, ESnet will continue development and deployment of the ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS) to support the Science Data Network.

11 Acknowledgements

This work would not have been possible without the contributions and participation of those who provided information and attended the workshop. ESnet would also like to thank the ASCR program office for their help in organizing the workshop and providing insight into the facilities supported by the ASCR program. In addition, the LBNL conference support and logistics staff was very helpful.

ESnet is funded by the US 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 US 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.

This is LBNL report LBNL-2495E