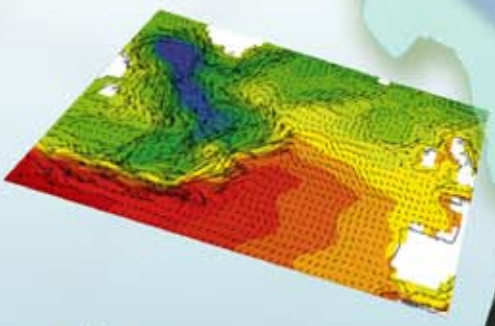
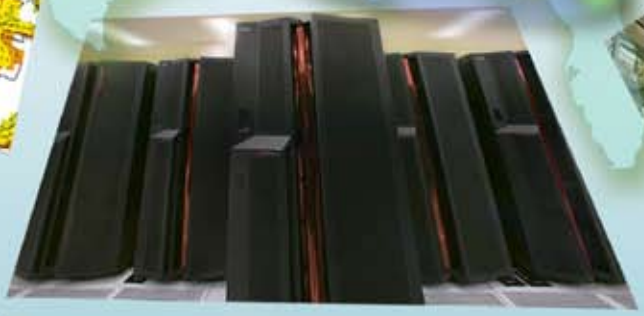
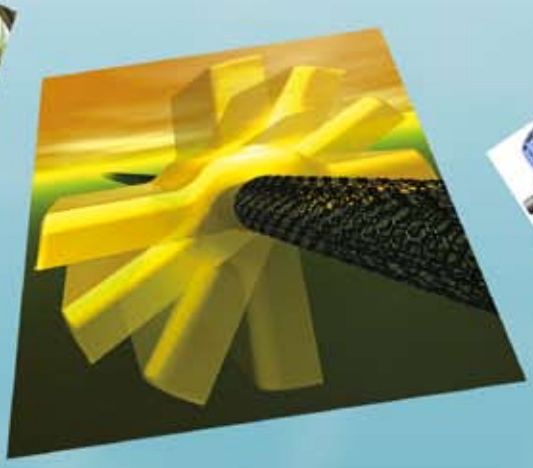
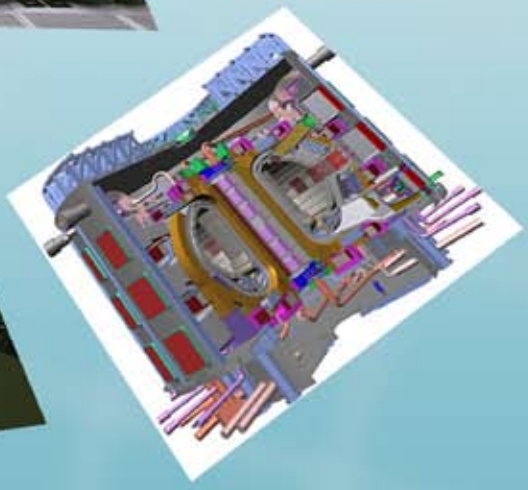
Science-Driven Network Requirements for ESnet

ESnet
Energy Sciences Network

Update to the 2002 Office of Science Networking Requirements Workshop Report
February 21, 2006

Berkeley Lab

Office of Science
U.S. Department of Energy

Science-Driven Network Requirements for ESnet

Update to the 2002 Office of Science Networking Requirements Workshop Report
February 21, 2006

Contributors

Paul Adams, LBNL (Advanced Light Source)
Shane Canon, ORNL (NLCF)
Steven Carter, ORNL (NLCF)
Brent Draney, LBNL (NERSC)
Martin Greenwald, MIT (Magnetic Fusion Energy)
Jason Hodges, ORNL (Spallation Neutron Source)
Jerome Lauret, BNL (Nuclear Physics)
George Michaels, PNNL (Bioinformatics)
Larry Rahn, SNL (Chemistry)
David Schissel, GA (Magnetic Fusion Energy)
Gary Strand, NCAR (Climate Science)
Howard Walter, LBNL (NERSC)
Michael Wehner, LBNL (Climate Science)
Dean Williams, LLNL (Climate Science)

Other Sources

LHC Operations Group Meetings
LHC Networking Workshops

Edited by Eli Dart, ESnet Engineering Staff
dart@es.net

Acknowledgements

ESnet is funded by the US Dept. of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program. Dan Hitchcock 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, Facilities Division.

This is LBNL report LBNL-61832

Contents

Chapter 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 and scientists that it serves. This focus has helped ESnet to be a highly successful enabler of scientific discovery for over 20 years.

In August, 2002 the DOE Office of Science organized a workshop to characterize the networking requirements for Office of Science programs. Networking and middleware requirements were solicited from a representative group of science programs. The workshop was summarized in two documents – the workshop final report and a set of appendixes.

This document updates the networking requirements for ESnet as put forward by the science programs listed in the 2002 workshop report. In addition, three new programs have been added. The information was gathered through interviews with knowledgeable scientists in each particular program or field.

Chapter 2 Updated Requirements from Science Programs and Facilities

2.1 *Introduction*

In August 2002, the DOE Office of Science organized a workshop to characterize the networking requirements for Office of Science programs. Networking and middleware requirements were solicited from a representative group of science programs chosen by the Office of Science. In a subsequent workshop in 2003, these requirements served as the basis for a new architecture for ESnet, which has been implemented to the extent of available funding.

Programs from five of the six DOE Office of Science program offices were characterized in case studies for the 2002 workshop: Chemical Sciences, Macromolecular Crystallography and the Spallation Neutron Source (SNS) from Basic Energy Sciences; Bioinformatics and Climate Science from Biological and Environmental Research; Fusion Energy Sciences; and High Energy Physics. This update to the 2002 workshop report adds three additional programs: the National Energy Research Scientific Computing (NERSC) Center and the National Leadership Computing Facility (NLCF) from Advanced Scientific Computing Research; and the Relativistic Heavy Ion Collider (RHIC) from Nuclear Physics.

The requirements updates were obtained through interviews, email and telephone conversations with senior scientists or staff in each particular program, facility or field. The interviewees were asked to enumerate the current state of their network requirements, and then project those requirements five years into the future. In addition, we asked for requirements coming beyond the five year time horizon. Some disciplines have such long-term plans that are enumerated here (e.g. the ITER facility for the Fusion program) while others (e.g. High Energy Physics) are in the middle of implementing large-scale projects that have already been the subject of years of hard work.

The case studies presented here are an updated set of networking requirements from a representative sample of the DOE Office of Science portfolio.

Note: in cases where data set sizes are identified and data rate requirements are not, an estimate of data rate requirements has been made. The estimate assumes that the data set must be moved in 8 hours.

2.2 *Bioinformatics and Life Sciences*

2.2.1 Executive Summary

The Pacific Northwest National Laboratory is building new Confocal microscopes with enhanced capabilities. These provide high resolution video of the subject samples, which are typically protein molecules. The microscopes also have a remote steering capability. Typical use of these microscopes is multidisciplinary, requiring the data stream to be multicast to multiple scientists at multiple remote institutions. Data rates are expected to be 78 megabytes (625 megabits) per second per microscope within 6 months. PNNL expects to have 20 of these microscopes in full production within two years. In 5 years, upgrades will increase the per-camera data rate by a factor of 20 to 12.5 gigabits per second per camera, for a total bandwidth requirement of 250 gigabits per second for the set of 20 microscopes. Genomes to Life (GTL) is expected to have a significant impact in the 2010-2012 time frame.

2.2.2 PNNL Bioinformatics Program

PNNL is building new confocal microscopes with enhanced capabilities, including remote steering. Current instruments have one camera sensor per microscope. A microscope with two sensors is currently in testing, and is expected to be in production use within 6 months. Twenty of these microscopes with two sensors each are to be deployed for production science use at PNNL within 2 years. The data stream is high-resolution video.

Each microscope sensor provides 39 megabytes per second of traffic load to the network. Since the microscopes that will be deployed for production science use will have two sensors each, we assume that one microscope will have a data rate of 78 megabytes per second, or 625Mbps. The control channel for remote steering is low bandwidth, but demands high reliability and bandwidth guarantees to ensure control stability. Typical use of these microscopes involves multidisciplinary collaboration, where specialists in multiple fields must view the data simultaneously from multiple institutions.

In 5 years, the per-microscope data rate is expected to rise by a factor of 20, resulting in a network bandwidth requirement of 12.5Gbps per microscope, in addition to the control channel.

In addition to the microscopes, there is a proteomics simulation program that will generate 0.5 petabytes of data in 2006. 5 years from now, proteomics simulations will be generating 5 petabytes per year.

In the 2010 to 2012 time frame, the Genomes to Life centers will be fully operational.

Table 2.2. Bioinformatics Requirements Summary

Feature			Anticipated Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Network**	**Network Services and Middleware**
Near-term	• Large-scale proteomics simulations generate datasets 500TB in size • Custom confocal microscopes begin operation • Confocal microscopes generate 625Mbps of video per microscope • 20 microscopes on line in two years	• Multidisciplinary microscopy collaboration	• Transfer of proteomics data sets to collaborator's sites for computation • Visualization of microscope video • Multicasting of microscope video streams • 12.5Gbps of microscope data in two years	• Remote steering applications
5 years	• Confocal microscopes generate 12.5Gbps each • 20 microscopes in full production use • Proteomics simulations generate 5PB/year	• Multidisciplinary microscopy collaboration	• QoS • Virtual Circuits • 250Gbps of microscope data • Multicasting of microscope video streams	• Authentication • Grid services
5+ years	• Genomes to Life	• Multiple GTL centers, each with computational capabilities and large-scale instruments		

2.3 *Chemical Science*

2.3.1 Executive Summary

The chemistry community is in the process of developing the tools to link experiments with simulation data, create data repositories, and collaborate in real time between geographically distributed sites. Over the next 5 years, simulation data set sizes are expected to increase by a factor of 10 from the current 10-30 terabytes to 100-300 terabytes. The creation of data repositories will increase the load on the network as the data sets are aggregated and then analyzed. The use of Grid middleware, Grid security services, and metadata management tools are expected to increase substantially. Required data rates, as derived from data set sizes, are 3 to 9 gigabits per second today, and 30 to 90 gigabits per second in 5 years.

2.3.2 Chemical Science Program

The chemistry community is extensive and incorporates a wide range of experimental, computational, and theoretical approaches to the study of problems, including advanced, efficient engine design; cleanup of the environment in the ground, water, and atmosphere; the development of new green processes for the manufacture of products that improve the quality of life; and biochemistry for biotechnology applications including improving human health. The advanced computing infrastructure that is being developed will revolutionize the practice of chemistry by allowing us to link high-throughput experiments with the most advanced simulations.

To overcome current barriers to collaboration and knowledge transfer among researchers working at different scales, a number of enhancements must be made to the information technology infrastructure of the community:

- A collaboration infrastructure is required to enable real-time and asynchronous collaborative development of data and publication standards, formation and communication of interscale scientific collaborations, geographically distributed disciplinary collaboration, and project management.

- Advanced features of network middleware are needed to enable management of metadata, user-friendly work flow for web-enabled applications, high levels of security especially with respect to the integrity of the data with minimal barriers to new users, customizable notification, and web publication services.

- Repositories are required to store chemical sciences data and metadata in a way that preserves data integrity and enables web access to data and information across scales and disciplines.

- Either tools now used to generate and analyze data at each scale must be modified or new translation/ metadata tools must be created to enable the generation and storage of the required metadata in a format that allows interoperable workflow with other tools and web-based functions. These tools also must be made available for use by geographically distributed collaborators.

- New tools are required to search and query metadata in a timely fashion and to retrieve data across all scales, disciplines, and locations.

- New tools and network services are needed to support the collaborative definition, execution and analysis of petascale simulation data, such as contained in the concept of a 'computational end station' for combustion research.

The advanced computing infrastructure that is being developed will revolutionize the practice of chemistry by allowing us to link high-throughput experiments with the most advanced simulations. Chemical simulations taking advantage of the soon-to-come petaflop architectures will enable us to guide the choice of expensive experiments and reliably extend the experimental data into other regimes of interest. The simulations will enable us to bridge the temporal and spatial scales from the molecular up to the macroscopic and to gain novel insights into the behavior of complex systems at the most fundamental level. For this to happen, we will need to have an integrated infrastructure including high-speed networks, vast amounts of data storage, new tools for data mining and visualization, modern problem-solving environments to enable a broad range of scientists to use these tools, and, of course, the highest-speed computers with software that runs efficiently on such architectures at the highest percentages of peak performance possible.

Table 2.3. Chemical Science Requirements Summary

Feature / Time Frame	Science Instruments and Facilities	Process of Science	Anticipated Requirements: Network	Anticipated Requirements: Network Services and Middleware
Near-term	• High data-rate instruments running for long times producing large data sets • Greatly increased simulation resolution- data sets ~10–30 terabytes • Geographically separated resources (compute, viz, storage, instmts) & people • Numerical fidelity and repeatability • Cataloguing of data from a large number of instruments • Large scale quantum and molecular dynamics simulations	• Distributed multi-disciplinary collaboration • Remote instrument operation / steering • Remote visualization • Sharing of data and metadata using web-based data services • Computing on the net by linking large scale computers	• Robust connectivity • Reliable data transfer • High data-rate, reliable multicast • Quality of service • International interoperability for namespace, security • Large-scale data storage needed both for permanent and temporary data sets. Can the network serve as a large scale data cache?	• Collaboration infrastructure • Management of metadata • High data integrity • Global event services • Cross discipline repositories • Network caching • Server side data processing • Virtual production to improve traceability of data • Data Grid broker / planner • Cataloguing as a service
5 years	• 3D Simulation data sets 30–100 terabytes • Coupling of MPP quantum chemistry and molecular dynamics simulations for large scale simulations in chemistry, combustion, geochemistry, biochemistry, environmental studies, catalysis • Validation using large experimental data sets • Analysis of large scale experimental data sets including visualization and data mining	• Remote steering of simulation, e.g., control of the time step, convergence of the SCF, introducing a perturbation in an MD simulation • Remote data sub-setting, mining, and visualization • Shared data/ metadata with annotation evolves to knowledge base	• 10s of gigabits for collaborative visualization and mining of large data sets	• Remote I/O • Collaborative use of common, shared data sets – version control on the fly • International interoperability for collaboratory infrastructure, repositories, search, and notification • Archival publication
5+ years	• Accumulation of archived simulation feature data and simulation data sets • Multi-physics and soot simulation data sets ~1 petabyte • Combustion simulations incorporating new uncertainty quantification algorithms ~> 1 petabyte • Large-scale MD simulations – 100s of terabyte to petabyte datasets	• Internationally collaborative knowledge base • Remote collaborative simulation steering, mining, visualization, and analysis	• 100+ gigabit for distributed simulations – computational quantum chemistry, molecular dynamics, CFD combustion simulations	• Remote collaborative simulation steering, mining, visualization, and analysis

2.4 *Climate Modeling Requirements*

2.4.1 Executive Summary

Climate modeling requires significant high-performance computing resources to run the modeling codes many times – this results in a large store of data that is then analyzed and compared, with the results fed back into the model. This means that both the rate of data production and the amount of data moved for analysis are proportional to the supercomputer time allocations given to climate analysts. Current data repositories contain a few hundred terabytes of data in total (75 to 180 terabytes at a few repositories). The climate community has gained significant scientific leverage by aggregating data into central repositories and allowing climate scientists to mine the aggregated data. While this places additional load on the network, the scientific payoff is huge – a good example of this is PCMDI at LLNL. Demand for robust, high-speed networks to carry climate model data to climate scientists will increase substantially over the next 5 years, as will the demand for Grid middleware to enable efficient access to repositories. As models increase in complexity and resolution, the data sets produced will grow by a factor of 10 over the next 5 years. In addition, the climate modeling community will require the integration of running simulations at geographically distributed computing sites, placing further load on network and storage infrastructure. Based on data set sizes, data rate requirements will reach 30 gigabits per second in 5 years.

2.4.2 Background

Climate change and its implication for human and ecological systems is among the most compelling issues of our time. The current generation of climate models has explained 20th century climate change quite well at large geographical scales. However, policymakers require information about climate change at much smaller scales than currently feasible. As computing technology advances, climate scientists can increase the resolution of models to better capture regional scale phenomena. High impact events such as climate extremes (heat waves and cold snaps), hurricanes, drought and precipitation pattern changes, require not only higher fidelity but far more simulations in order to fully describe the climate change signal within the ambient noise. Over the next five years, climate models will see an even greater increase in complexity with the addition of biogeochemistry and atmospheric chemistry submodels to existing descriptions of the atmosphere, ocean, sea ice and land. Hence, the expected output from US climate models (especially NCAR CCSM) is expected to be significantly increased. Moreover, due to the expense involved in producing this unique model dataset, distribution to a large worldwide set of climate model analysts is required.

2.4.3 Climate Modeling Today

To better understand climate change, we need better climate models – and to get those, we need to exhaustively analyze what's incorrect about today's models in order to improve them. The cycle of analysis → improved model → analysis is typical of climate model work generally. One thing we do know is that climate models today are too low in resolution to get some

important features of the climate right. Generally, the computing power will be there over the next 5-10 years, but to determine things like climate extremes (heat waves and cold snaps), hurricanes,[a] drought and precipitation pattern changes, and other potential changes as a result of climate change, we need better analysis. Currently, analysis is accomplished by transferring the data of interest from the computing site to the climate scientist's institution. The recent trend in climate modeling has been to share data from all of the world's leading climate models freely with the scientific community. This can be inefficient if the data volume is large, and several strategies to reduce the data volume before transfer have been developed. However, the scientific payoff has been enormous with over 200 refereed papers written in less than a year using analysis of the IPCC AR4 model data distributed from the Program for Climate Model Diagnosis and Intercomparison at LLNL.
Hence, faster networks to deliver more climate model data more efficiently are urgently required. Since climate models require large computing resources, there are only a few sites in the U.S. and worldwide that are suitable for executing these models at this time. In addition, for efficiency reasons, the data produced by these integrations are often stored at the same sites - however, climate scientists are scattered all over the globe, furthering the need for efficient data distribution.

2.4.4 The Next Five Years

Over the next five years, climate models will see an even greater increase in complexity than that seen in the last ten years. Influences on climate will no longer be approximated by essentially fixed quantities, but will become interactive components in and of themselves. The North American Carbon Project (NACP), which endeavors to fully simulate the carbon cycle, is an example. Increases in resolution, both spatially and temporally, are in the plans for the next two to three years. During 2004 and 2005, for the IPCC AR4 (Intergovernmental Panel on Climate Change Fourth Assessment Report), approximately 180 terabytes of data was generated by the NCAR Community Climate System Model (CCSM). This data was reduced in volume to about 15 terabytes to meet IPCC requirements, but the CCSM was only one of about 20 climate modeling centers worldwide that performed experiments for the AR4 and submitted data for analysis.. These much finer resolution models, as well as the distributed nature of computing resources, will demand much greater bandwidth and robustness from computer networks than is presently available. These studies will be driven by the need to determine future climate at both local and regional scales as well as changes in climate extremes - droughts, floods, severe storm events, and other phenomena. Climate models will also incorporate the vastly increased volume of observational data now available (and that available in the future), both for hind casting and intercomparison purposes. The end result is that instead of tens of terabytes of data per model instantiation, hundreds of terabytes to a few petabytes (10^{15} bytes) of data will be stored at multiple computing sites, to be analyzed by climate scientists worldwide. The Earth System Grid and its descendents will be fully utilized to disseminate model data and for scientific

([a]) Hurricane Katrina in 2005, cost over 1000 lives and many tens of billions in dollars in damage. Current climate models aren't quite good enough to resolve hurricanes, but research models driven by reasonably realistic future climate scenarios imply that Katrina-strength hurricanes striking the US will become more common. That implies many more billions in damage and more deaths.

analysis. Additionally, these more sophisticated analyses and collaborations will increase the needed network resources and infrastructure. It is expected that many climate scientists will examine the model data – many more than today. Bulk data transfer will be necessary, as well as tools like Access Grids and personal Grids.

As climate models become more multidisciplinary, scientists from fields outside of climate, oceanography and the atmospheric sciences will collaborate on the development and examination of climate models and their output. Biologists, hydrologists, economists and others will assist in the creation of additional components that represent important but as-yet poorly known influences on climate. These models, sophisticated themselves, will likely be utilized at computing sites other than where the climate model is executed. In order to maintain efficiency, dataflow to and from these collaboratory efforts will demand extremely robust and fast networks.

2.4.5 2010 and Beyond

In the following five years, climate models will again increase in resolution, and many more fully interactive components will be integrated. At this time, the atmospheric component may become nearly mesoscale (commonly used for weather forecasting) in resolution, 30 km by 30 km, with 60 vertical levels. Climate models will be used to drive regional scale climate and weather models, which require resolutions in the tens to hundreds of meters range, instead of the typical hundreds of kilometers resolution of the CCSM. There will be a true carbon cycle component, models of biological processes will be used, for example, simulations of marine biochemistry (which affects the interchange of greenhouse gases like methane and carbon dioxide with the atmosphere), and fully dynamic vegetation. These scenarios will include human population change and growth (which effect land usage and rainfall patterns) and econometric models, to simulate the potential changes in natural resource usage and efficiency. Additionally, models representing solar processes, to better simulate the incoming solar radiation, will be integrated. Climate models at this level of sophistication will likely be run at more than one computing center in distributed fashion, which will demand extremely high speed and tremendously robust computer networks to interconnect them. Data volumes almost certainly will reach several petabytes, if not more.

Table 2.4.2. Climate Requirements Summary

Feature	Science Instruments and Facilities	Process of Science	Anticipated Requirements	
Time Frame			Network	Network Services and Middleware
Near-term	• A few data repositories, many distributed computing sites • NCAR[(a)] - 180 Tbytes • NERSC[(b)] - 75 Tbytes • ORNL[(c)] – 75 Tbytes		• Authenticated data streams for easier site access through firewalls	• Server side data processing (computing and data cache embedded in the net) • Information servers for global data catalogues
5 years	• Add many simulation elements/components as understanding increases • 100 Tbytes / 100 model yrs generated simulation data – 1-5 Pbytes / yr (at NCAR) • Distribute in large datasets to major users/collaborators for post-simulation analysis	• Enable the analysis of model data by all of the collaborating community (major US collaborators are a dozen universities, several Federal Agencies, and international collaborators)	• Robust, secure access to large quantities of data (multiple paths)	• Reliable data/file transfer • Across system/ network failures
5+ years	• Add many diverse simulation elements/components, including from other disciplines - this must be done with distributed, multidisciplinary simulation as the many specialized sub-models will be managed by experts in those fields • 5-10 Pbytes/yr (at NCAR) • Virtualized data to reduce storage load	• Integrated climate simulation that includes all high-impact factors	• Robust networks supporting distributed simulation - adequate bandwidth and latency for remote analysis and visualization of massive datasets • Quality of service guarantees for distributed simulations	• Server side computation for data extraction/ subsetting, reduction, etc., before moving across the network • Virtual data catalogues for data generation descriptions, data regeneration planners, data naming and location transparency services for reconstituting data on demand

(a) NCAR = National Center for Atmospheric Research.
(b) NERSC = National Energy Research Scientific Computing Center
(c) ORNL = Oak Ridge National Laboratory

2.5 *High Energy Physics*

2.5.1 Executive Summary

The current focus of networking in the High Energy Physics community revolves around the Large Hadron Collider (LHC) at CERN in Switzerland. Two LHC experiments are of particular interest in the U.S: CMS (Compact Muon Solenoid) and ATLAS (A Toroidal LHC ApparatuS). The Atlas and CMS experiments are widely distributed collaborations composed of 2000 physicists from 150 institutions in more than 30 countries, including 300 to 400 US physicists from more than 30 universities as well as the major US high-energy physics laboratories. The CMS and ATLAS instruments will produce several petabytes of data per year in full production. The data will be distributed from CERN using a tree model of several tiers, with Tier0 being the central repository at CERN. Each experiment will have a Tier1 center in each participating nation-state or group of nation-states, and that Tier1 center is responsible for distributing its data to client Tier2 centers which will perform computational analysis on the data. In the U.S, the CMS Tier1 center is at the Fermi National Accelerator Laboratory, and the ATLAS Tier1 center is at the Brookhaven National Laboratory. Tier0-Tier1 data flows will require 10 to 20 gigabits per second of bandwidth per Tier1 by the end of 2007, and 40 gigabits per second per Tier1 by 2010. Tier1-Tier2 data flows have a similar requirement, since all the data that enters a Tier1 center is subsequently exported to its client Tier2 centers. Since all the U.S. Tier2 centers are universities and both U.S. Tier1 centers are DOE laboratories, ESnet will carry all the Tier0-Tier1 traffic and all the Tier1-Tier2 traffic. There is an additional requirement for significant network capacity outside the tiered data distribution model that is the subject of current study.

2.5.2 Large Hadron Collider

The High Energy Physics (HEP) community is currently focused primarily on preparing for the production operation of the Large Hadron Collider at CERN. There are several other HEP instruments (BaBAR, D0, etc) in production mode – their requirements are met by the current infrastructure and are not addressed here. In contrast, the LHC requirements are unmet by current infrastructure, and are a significant driving force for increased bandwidth and a more flexible network architecture.

The LHC experiments addressed here are CMS (Compact Muon Solenoid) and ATLAS (A Toroidal LHC ApparatuS). These are large experiments that are expected to produce several petabytes of data per year when in full production. In addition, collaborators encompass 2000 physicists from 150 institutions in more than 30 countries, including 300 to 400 US physicists from more than 30 universities as well as the major US high-energy physics laboratories.

The large data volumes and distributed nature of the data analysis result in a huge data management problem. The solution to this is to distribute the data from the instruments at CERN to the analysis systems in a tiered fashion. The Tier0 center is CERN itself. CERN will distribute the data to a set of Tier1 centers (typically one per participating

region or nation-state) that act as a distributed archive of the experiment data. The Tier1 centers then distribute the data to client Tier2 centers that perform the computational analysis. The United States is unusual in that it has a Tier1 center for both the CMS and ATLAS experiments – the CMS Tier1 center is located at the Fermi National Accelerator Laboratory and the ATLAS Tier1 center is located at the Brookhaven National Laboratory. In addition, the US Tier1 centers are unusual in that they have a significantly larger client base when compared to most Tier1 centers.

The Tier0 to Tier1 traffic flows for the US Tier1 centers will traverse a purpose-built transatlantic network (USLHCNet). In the initial deployment phase, USLHCNet will provision a 10Gbps link to Starlight in Chicago to serve the CMS Tier1 center at Fermi, and a 10Gbps link to New York to serve the ATLAS Tier1 center at Brookhaven. USLHCNet will also provision a 10Gbps circuit between Starlight and New York for redundancy. ESnet will provision 10Gbps connectivity between USLHCNet and the US Tier1 centers via Metropolitan Area Networks (MANs) in Chicago and New York. Within 5 years, the increasing bandwidth requirements of the LHC will drive the upgrade of the USLHCNet links and the MANs that connect USLHCNet to the US Tier1 sites from 10Gbps to 30-40Gbps.

In addition to the Tier0 to Tier1 flows, there will be significant traffic load generated by the distribution of data from the Tier1 centers to the Tier2 centers for analysis. There will also be significant traffic load stemming from data transfers outside the tiered distribution model – examples include Tier2 centers fetching data from multiple Tier1 centers, and Tier1 to Tier1 traffic. These flows are the subject of ongoing requirements gathering and characterization, and will become clear in the near future.

Feature			Anticipated Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Network**	**Network Services and Middleware**
Near-term	• Instrument based data sources • Hierarchical data repositories • Improved quality of videoconferencing capabilities • Final preparations for LHC	• Tiered data distribution • Computational analysis at Tier2 centers	• 10Gbps to BNL to support Tier0-Tier1 traffic • 20Gbps to FNAL to support Tier0-Tier1 traffic • Significant additional traffic from Tier2 centers • Significant additional Tier1-Tier1 traffic	• Grid for job submission, data movement, etc. • Collaboration tools • Deadline scheduling
5 years	• LHC experiments in full production	• Tiered data distribution • Science-driven data access (significant traffic outside of tiered distribution model) • Widely distributed computational analysis	• 30-40Gbps each to BNL and FNAL for Tier0-Tier1 traffic • Significant additional traffic from Tier2 centers • Significant additional Tier1-Tier1 traffic	• Continued reliance on Grid, Federated Trust services. • Collaboration tools • Deadline scheduling
5+ years	• 1000s of petabytes of data	•	• 1000 gigabits/sec	•

2.6 *Macromolecular Crystallography*

2.6.1 Executive Summary

Macromolecular crystallography is an experimental technique that is used to solve structures of large biological molecules (such as proteins) and complexes of these molecules. The current state-of-the-art implementation of this technique requires the use of a source of very intense, tunable, x-rays that are produced only at large synchrotron radiation facilities. The high operating cost of these facilities, coupled with the heavy demand for their use, has led to an emphasis on increased productivity and data quality that will need to be accompanied by increased network performance and functionality. Current requirements for average data transfer rate are 1 to 25 megabytes per second per station; it is expected that in five to ten years, this will increase by an order of magnitude to 10 to 250 megabytes per second per station. This is exacerbated further by the fact that most research facilities have from four to eight stations; this places a future requirement of 40 to 2000 megabytes per second per facility. In addition to raw bandwidth, there is the need for network services enabling remote instrument control, collaboration, and remote visualization.

2.6.2 Macromolecular Crystallography

Macromolecular crystallography is an experimental technique that is used to solve structures of large biological molecules (such as proteins) and complexes of these molecules. The current state-of-the-art implementation of this technique requires the use of a source of very intense, tunable, x-rays that are produced only at large synchrotron radiation facilities. In the United States, more than 36 crystallography stations are distributed among the synchrotron facilities and dedicated to macromolecular crystallography. The high operating cost of these facilities, coupled with the heavy demand for their use, has led to an emphasis on increased productivity and data quality that will need to be accompanied by increased network performance and functionality.

The data acquisition process involves several interactive online components, data archiving and storage components, and a compute-intensive offline component. Each component has associated networking requirements. Online process control and online data analysis are real-time, interactive activities that monitor and coordinate data collection. They require high-bandwidth access to images as they are acquired from the detector. Online data analysis now is limited primarily to sample quality assurance and to data collection strategy. There is increasing emphasis on expanding this role to include improved crystal scoring methods and real-time data processing to monitor sample degradation and data quality. Online access to the image datasets is collocated

and could make good use of intelligent caching schemes. Datasets from previously exposed samples are not required during online processing.

High-performance networking can play several roles in online control and data processing. Bob Sweet at the Brookhaven National Laboratory National Synchrotron Light Source has outlined several approaches to remote, networked, collaboratory operation. The datasets most often are transferred to private institutional storage. This requirement places a large burden on the data archiving process that transfers the data between online and offline storage units. Current requirements for average data transfer rate are 1 to 25 Mbytes/s per station; it is expected that in five to ten years, this will increase by an order of magnitude to 10 to 250 Mbytes/s per station. This is exacerbated further by the fact that most research facilities have from four to eight stations; this places a future requirement of 40 to 2000 Mbytes/s per facility. Advanced data compression schemes might be able to reduce these figures by a factor of 5 to 10.

In addition to increased raw network bandwidth, the next-generation high-performance networking infrastructure will need to provide tools and services that facilitate object discovery, security, and reliability. These tools are needed for low-latency applications such as remote control as well as high-throughput data transfer applications such as data replication or virtual storage systems.

Table 2.6. Macromolecular Crystallography Requirements Summary

Feature			Anticipated Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Network**	**Network Services and Middleware**
Near-term	• 10 experiments per day, 10's of GB of data per experiment at full resolution • Preliminary analysis of one experiment guides parameters of subsequent experiments		• Bulk transfer of resultant data sets to home institution for later analysis, ~1TB/day for large data sets	
5 years	• Additional beamlines come on line	• Detector upgrades increase resolution and data set size	• 5TB/day data transfer (500Mbps throughput sustained) • QoS to support remote steering, remote visualization	• Remote collaboration with on-site staff • Remote steering • Remote visualization
5+ years	• Detector and software improvements resulting in 3D data sets	• Increased biological understanding from more sophisticated analysis	• 10TB/day data transfer (1Gbps throughput sustained) • Increased reliance on QoS capabilities to support advanced middleware and services	• Increased reliance on remote visualization and steering • Live video for environmental monitoring of experiments

2.7 *Magnetic Fusion Energy Science*

2.7.1 Executive Summary

The Magnetic Fusion Energy program is composed of two major parts – an experimental component and a theoretical, simulation-based component. Fusion experiments in the U.S. are conducted at three large facilities. These experiments operate in a pulsed mode producing plasmas of up to 10 seconds duration every 15 to 20 minutes, with 25 to 35 pulses per day. Each pulse generates several gigabytes of data. These data are then subjected to analysis, the output of which is used to determine the input parameters for the next pulse. This duty cycle demands reliability and guaranteed bandwidth (200+ megabits per second), as well as remote collaboration capability. In 5 years the amount of data taken per pulse will have increased to 20 gigabytes, and the guaranteed bandwidth requirement will be 1 gigabit per second. When ITER enters production in 2015, the pulse frequency will be once per hour, and the data volume will be 1 terabyte per pulse. The present analysis duty cycle is expected to apply to ITER. The simulation component of the fusion program is distributed in nature, with simulation codes running at the major supercomputer centers (e.g. NERSC, NLCF) and post-simulation analysis occurring at about 20 sites. The post-simulation analyses have generated a distributed data archive several 10's of terabytes in size – this data requires middleware for efficient querying for further analysis. Both the experimental and simulation components of the fusion program rely on ESnet's PKI infrastructure for authentication and Grid support.

2.7.2 Magnetic Fusion Energy Science

The long–term goal of magnetic fusion research is to develop a reliable energy system that is environmentally and economically sustainable. To achieve this goal, it has been necessary to develop the science of plasma physics, a field with close links to fluid mechanics, electromagnetism, and nonequilibrium statistical mechanics. Fusion Energy Sciences is highly collaborative with a small number of large experimental facilities and a computationally intensive theoretical program, creating unique challenges for computer networking and middleware.

In the U.S., experimental magnetic fusion research is centered at three large facilities (Alcator C–Mod, DIII–D, NSTX). As experiments have increased in size and complexity, there has been concurrent growth in the number and importance of collaborations between groups at the experimental facilities and smaller groups located at universities, industry sites, and national laboratories. International collaborations, always an essential element of the program, have taken on added importance with the impending start of the ITER project. Though ITER, the International Thermonuclear Experimental Reactor, will not be operational for ten years, preparatory work is a large and growing component of the U.S. program. Joint experiments with our international partners have become a major focus of the U.S. facilities and the prospect of remote participation on ITER drives interest in this mode of operation.

Magnetic fusion experiments operate in a pulsed mode producing plasmas of up to 10 seconds duration every 15 to 20 minutes, with 25 to 35 pulses per day. For each plasma pulse, up to 10,000 separate "channels" are acquired, totaling several Gbytes of

data. It is anticipated that the data available between pulses will grow to the 10 Gbyte level within the next five years. Throughout the experimental session, adjustments to plasma conditions are debated and discussed among an experimental team, composed typically of 20–50 scientists, students and engineers, and implemented via a plasma control system. Most team members work on site in the control room while others participate from remote locations. Decisions for changes to each plasma pulse are informed by data analysis and visualization carried out within the roughly 20 minute between-pulse interval. This mode of operation places a large premium on rapid data analysis that can be assimilated in near–real–time by a geographically dispersed research team. The computational emphasis in the experimental science area is to perform ever more complex data analysis between pulses. Five years from now, analysis that is today performed overnight should be completed between pulses. The movement of data, driven by interactive use of advanced analysis and visualization servers, will place a severe burden on the network infrastructure.

Teaming with the experimental community is a theoretical and simulation community whose efforts range from applied analysis of experimental data, fundamental theory and the creation of non–linear 3D plasma models. Datasets generated by simulation codes will exceed the Tbyte level within the next three to five years. A new long-term initiative, called the Fusion Simulation Project (FSP) is attempting to integrate multiple physics modules into more complex models. The FSP will drive fusion computational science toward larger, more geographically diffuse collaborations. The first steps in the FSP, sometimes called focused integration initiatives, are already composed of large distributed teams and will require collaboration technologies and infrastructure similar to that needed for experiments.

The need for real-time interactions among large research teams and the interactive visualization and processing of very large datasets drive additional network requirements. Important components include an easy-to-use and easy-to-manage user authentication and authorization framework, global directory and naming services, distributed computing services for queuing and monitoring, and network quality of service (QoS) in order to provide guaranteed bandwidth at particular times or with particular characteristics.

Table 2.7. Magnetic Fusion Energy Requirements Summary

Feature			Anticipated Requirements	
Time Frame	Science Instruments and Facilities	Process of Science	Network	Network Services and Middleware
Near-term	• Each experiment only gets a few days per year - high productivity is critical • Experiment episodes ("shots") generate 2-3 Gbytes every 20 minutes, which has to be delivered to the remote analysis sites in two minutes in order to analyze before next shot • Highly collaborative experiment and analysis environment	• Real-time data access and analysis for experiment steering (the more that you can analyze between shots the more effective you can make the next shot) • Shared visualization capabilities		• PKI certificate authorities that enable strong authentication of the community members and the use of Grid security tools and services. • Directory services that can be used to provide the naming root and high-level (community-wide) indexing of shared, persistent data that transforms into community information and knowledge • Efficient means to sift through large data repositories to extract meaningful information from unstructured data.
5 years	• 10 Gbytes generated by experiment every 20 minutes (time between shots) to be delivered in two minutes • Gbyte subsets of much larger simulation datasets to be delivered in two minutes for comparison with experiment • Simulation data scattered across United States • Transparent security • Global directory and naming services needed to anchor all of the distributed metadata • Support for "smooth" collaboration in a high-stress environment	• Real-time data analysis for experiment steering combined with simulation interaction = big productivity increase • Real-time visualization and interaction among collaborators across United States • Integrated simulation of the several distinct regions of the reactor will produce a much more realistic model of the fusion process	• Network bandwidth and data analysis computing capacity guarantees (quality of service) for inter-shot data analysis Gbits/sec for 20 seconds out of 20 minutes, guaranteed • 5 to 10 remote sites involved for data analysis and visualization	• Parallel network I/O between simulations, data archives, experiments, and visualization • High quality, 7x24 PKI identity authentication infrastructure • End-to-end quality of service and quality of service management • Secure/authenticated transport to ease access through firewalls • Reliable data transfer • Transient and transparent data replication for real-time reliability • Support for human collaboration tools
5+ years	• Simulations generate 100s of Tbytes • ITER – Tbyte per shot, PB per year	• Real-time remote operation of the experiment • Comprehensive integrated simulation	• Quality of service for network latency and reliability, and for co-scheduling computing resources	• Management functions for network quality of service that provides the request and access mechanisms for the experiment run time, periodic traffic noted above.

2.8 *National Energy Research Scientific Computing Center (NERSC)*

2.8.1 Executive Summary

As DOE's flagship High Performance Computing (HPC) Center, NERSC has networking requirements which differ from many other DOE laboratories. NERSC supports approximately 300 DOE projects involving approximately 2400 scientists from government, education and private industry including 30 SciDAC and 3 INCITE projects. NERSC is also the home of the PDSF computational cluster which is used for high energy and nuclear physics research and the High Performance Storage System (HPSS) which currently stores approximately 2.2 PB of data including data sets of national interest. Since NERSC's user base is remote and widely dispersed, end to end high performance network connectivity is critical to the success of NERSC's mission. NERSC requires high speed connectivity to ESnet, and high-speed peerings between ESnet and the networks that serve NERSC's users (e.g. Abilene, CENIC). NERSC's current connection to ESnet is 10 gigabits per second. NERSC expects an upgrade to 20-40 gigabits per second to be required in late 2008. NERSC's mission also requires the ability of ESnet to support features that enable high single-stream bandwidth, such as jumbo frames. NERSC also makes use of Grid services that rely on ESnet's PKI infrastructure.

2.8.2 Introduction

As DOE's flagship High Performance Computing (HPC) Center, NERSC has networking requirements which differ from many other DOE laboratories. These requirements stem from the fact that NERSC supports approximately 300 DOE projects involving approximately 2400 scientists from government, education and private industry including 30 SciDAC and 3 INCITE projects. In addition, NERSC is the home of the PDSF computational cluster which is used for high energy and nuclear physics research and the High Performance Storage System (HPSS) which currently stores approximately 2.2 PB of data including data sets of national interest.

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 and the reliability and stability (amount of packet loss) from end to end.

The NERSC ESnet requirements are divided into the following areas:

- NERSC to/from (↔) ESnet link speed
- ESnet ↔ other networks peering speed
- End-to-end single stream bandwidth
- Additional services
- NERSC ↔ ESnet link funding

2.8.3 NERSC ↔ ESnet link speed

DOE Networking workshops and the NERSC Greenbook provide a detailed description of the scientific needs for computation, network and storage. The published data from three DOE networking workshops[2][3][4] that predict scientific data transfer rates for the next several years may be useful to estimate ESnet backbone requirements. For a single ESnet site, however, the amount of data that needs to be transferred to that site is heavily dependent on the DOE projects allocations at the site as well as the site's computing and data storage capabilities. At NERSC, the DOE project supported is based on allocations of computer time and storage.

The two main drivers of WAN network bandwidth at NERSC are incoming data from remote sites which are stored on the NERSC HPSS system and then used on NERSC computational systems and data generated on the computational systems which are usually stored on HPSS and 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 ↔ ESnet link bandwidth requirements.

The DOE-NNSA ASCI program uses the ratio of WAN network bandwidth to the peak system performance to predict required WAN bandwidth. The table below summarizes this ratio for several ASCI systems and NERSC.

[2] High Performance Network Planning Workshop, August 13-15 2002, Reston, VA.
[3] DOE Workshop on Ultra High-Speed Transport Protocols and Network Provisioning for Large-Scale Science Applications, April 2003, Argonne National Laboratory.
[4] DOE Science Networking: Roadmap to 2008 Workshop, June 3-5, 2003, Jefferson Laboratory.

Table 2.10.1: Past and future bandwidth and computing Peak Performance

System	FY	WAN bandwidth ($x10^9$ bits/s - Gigabits per second)	System Performance (Peak $x10^{12}$ flop/s - Teraflop/s)	Ratio (bits/flop)
ASCI Q	02	10	30	.0003
NERSC 3	02	0.622	5	.0001
NERSC 3E	04	2.4	10	.0002
ASCI Purple (est.)	05	100	100	.001
NERSC 3E + NCS	05	10	13	.0008
NERSC 3E + NCS + NCSb	06	10	20	.0005
NERSC 5 (est.)	07-08	40	75	.0005
NERSC 6 (est.)	10	100	150	.0007

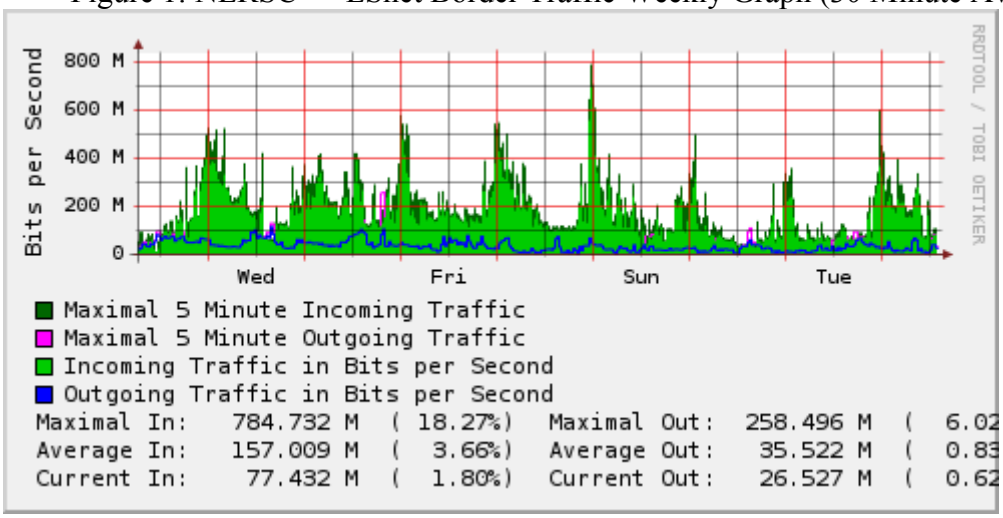
Note: The ASCI numbers are taken from an open literature publication[5]. To the best of our knowledge, the existing DisCom WAN bandwidth[6] between ASCI sites is 4.8Gb/s which would make the ratio .0002 for ASCI Q and .00005 for ASCI Purple.

It is clear from Table 1 that the NERSC WAN Bandwidth/Peak System Performance ratios have historically been in the range of .0002 to .0008 and are predicted to stay in this range throughout 2010. ASCI ratios also appear to be the same orders of magnitude[7]. The SNL ASC Highlights 2003 report[8] states that *"the DisCom WAN has provided users the capability to transfer large data files among laboratories at speeds up to 100 megabytes per second"* which is consistent with the ratios shown above. The NERSC future estimates assume NERSC baseline plan funding in which case the NERSC WAN traffic is driven primarily by incoming data being stored on HPSS. If the NERSC Capability plan is funded, we expect NERSC ESnet bandwidth requirements to accelerate in time but we doubt that the ratio will exceed .001 in the next 5 years.

In 2003, DOE requested that ESnet sites analyze their border traffic and categorize it into a few broad categories. The following graph shows a typical week of NERSC ↔ ESnet border traffic:

[5] ASCI Technology Prospectus, 2001, Sandia National Laboratory page 80
[6] ASC Highlights, 2003, Sandia National Laboratory, page 36
[7] ASCI Technology Prospectus - July 2001, http://www.nnsa.doe.gov/ASC/Files/Copy of Prospectus.pdf
[8] ASC Highlights, 2003, Sandia National Laboratory, page 36

Figure 1: NERSC ↔ ESnet Border Traffic Weekly Graph (30 Minute Average)



NERSC developed software to categorize the border traffic into broad categories. The following table summarizes three weeks of ESnet network border traffic for both NERSC and LBNL:

Table 2: NERSC and LBNL ESnet Border traffic by category

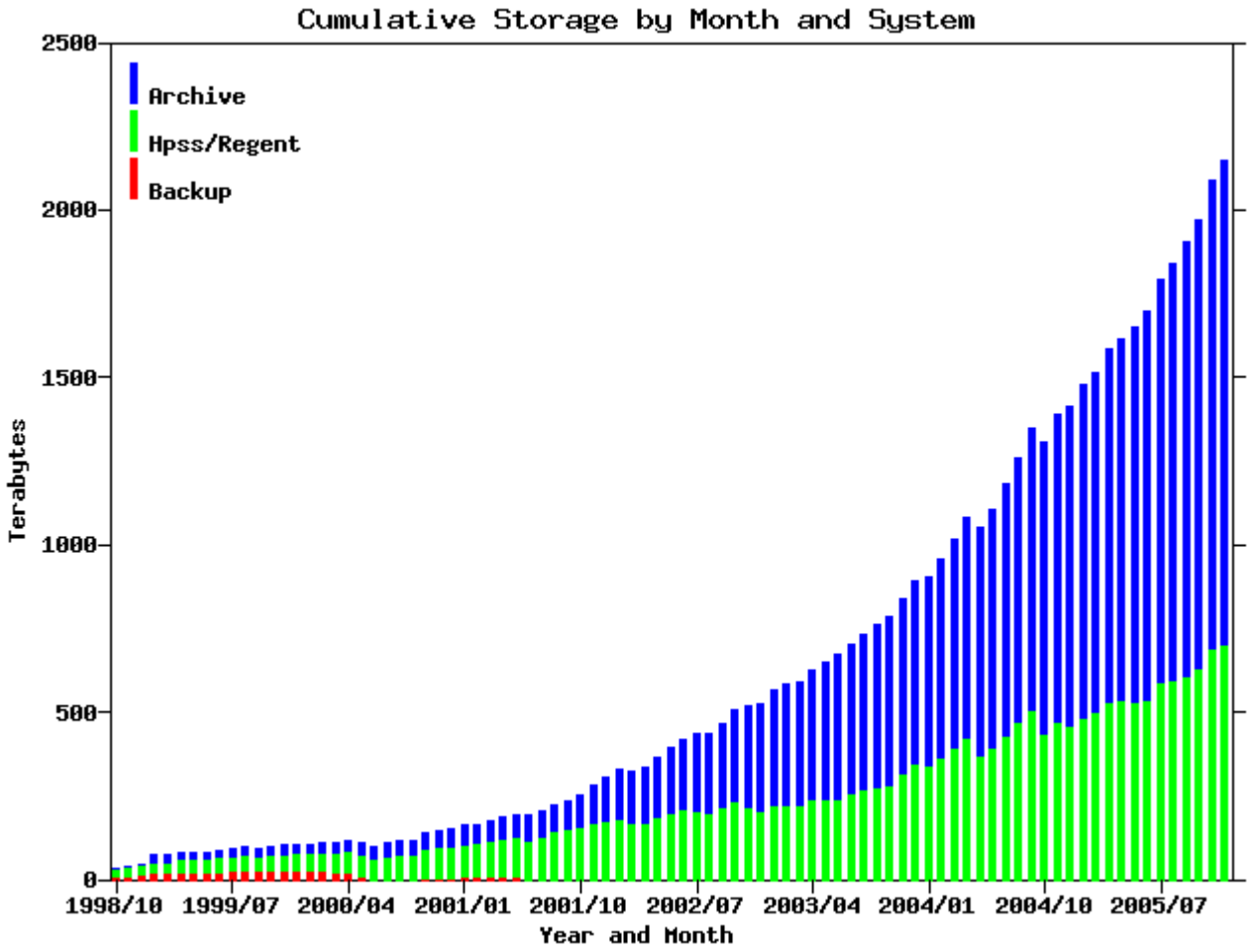
Type	NERSC	LBNL
Bulk Data (ftp, hsi)	85%	36%
Grid	7%	<1%
Computer System Services (DNS, iperf)	4%	14%
Interactive (ssh, kshell)	3%	4%
World Wide Web	<1%	41%
Mail	<1%	1%
Database	<1%	<1%
Uncategorized	1%	3%
Total	100%	100%

As shown in Table 2, the majority of NERSC border traffic is bulk and grid data which is significantly different than a laboratory site (LBNL is used as an example). Leading Edge Science that requires very large bulk data (>100GB data sets) is currently moved with TCP. To sustain high data rates for these large data sets it is important not to enter congestion control mechanisms in TCP. If one of these transfers stalls it may take hours to recover to the full transfer rate. NERSC has observed that in most cases a 2X headroom above the peak rate tends to be sufficient to avoid transfer stalls. NERSC currently averages 3TB of WAN traffic per day (~278Mb/s) while heavy throughput days approach 6TB/day (556Mb/s). Peak periods through the day are roughly twice the daily average (566Mb/s – 1.1Gb/s). Therefore, NERSC provisions 4X above the daily average (2X for peak and 2X for headroom). For NERSC's existing (10Gb/s) ESnet link, EEB will suffer when 4X the average rate (2.5Gb/s) approaches the link speed. An average rate of 2.5Gb/s corresponds to approximately 27TB per day. To summarize current traffic:

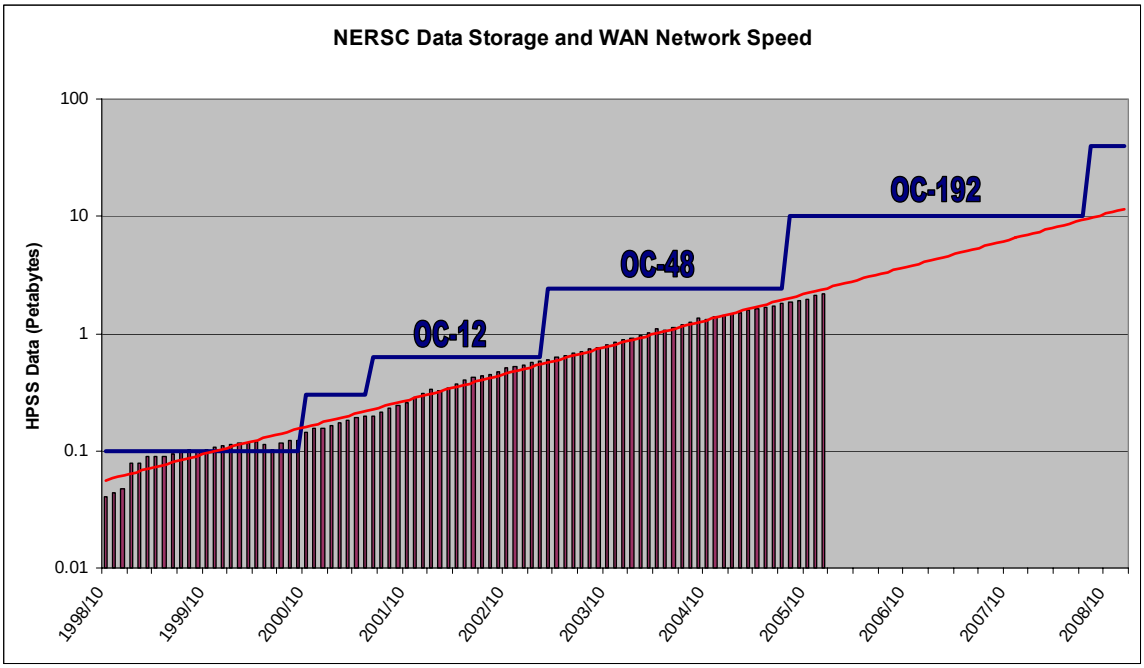
	Data moved	Average Rate required	Peak Rate required	Headroom to support large transfers
Average Day (today)	3 TB	278Mb/s	556Mb/s	800Mb/s
Heavy Day (today)	6 TB	556Mb/s	1.1Gb/s	2.2Gb/s
Upgrade Point	27 TB	2.5Gb/s	5Gb/s	10Gb/s

Note that it is as much the nature of the Leading Edge Science of just 10% of NERSC projects that use large data sets that requires the over provisioning as the nature of TCP in being very conservative in recovering from a stalled transfer. If either the large data sets or TCP are removed this level of over-provisioning could be reduced. If the traffic were not large bulk data or grid transfers then 4X over-provisioning would not be necessary. Any site that primarily processes large (millions) of small transfers will not be required to overprovision because any small transfer does not require high bandwidth rates.

This data is increasing at a rate of 1.7X per year and this rate has been consistent for the last 8 years. The following graph (http://www.nersc.gov/nusers/status/hpss/Summary.php) shows the actual amount of data stored on HPSS.



The data on this graph is replotted in the next graph with the Y-axis converted to a logarithmic scale:



OC-12 = 622Mbps, OC-48 = 2.5Gbps, OC-192 = 10Gbps

The trend line in red demonstrates that data is increasing at the rate of 1.7X per year. The blue line shows the NERSC ↔ ESnet link speed over the same time interval.

Assuming a 1.7X increase, NERSC is predicted to reach 27 TB/day in the Q4CY08. To minimize the chance of impacting scientific data transfers, the NERSC ↔ ESnet link should be upgraded beyond 10Gbps prior to this time. Assuming the computational capability at NERSC continues to increase at historical rates, the NERSC ↔ ESnet link will need to be upgraded to 20-40Gbps around October 2008.

If NERSC's computational capability was to increase at faster than historical rates, the NERSC ↔ ESnet link speed may have to be increased sooner. This is a distinct possibility with a significant increase in NERSC funding in the FY07 Presidential budget.

In addition to the production NERSC ↔ ESnet link discussed above, NERSC should have one or more 1-10Gbps non-production links for special projects such as the ESnet OSCARS (virtual circuits) project.

2.8.4 ESnet ↔ other networks peering speed

The majority of NERSC computational time is used by NERSC users located at university sites which do not have local ESnet connections. Two of the three FY04 and FY05 INCITE projects are also located at universities. These users access NERSC

through other national networks such as Abilene, CENIC and National LambdaRail. It is vital that ESnet peer with these other networks at backbone speeds.

2.8.5 End-to-end single stream bandwidth

Although TCP is relatively robust, in high-speed wide-area networks there are several issues. One primary problem is that the product of the bandwidth and the delay of the path is very large. To fully utilize the available bandwidth in such a path requires the amount of data in flight at any point in time be equal to the bandwidth delay product. For example, in order to fill a one-gigabit path which has a 100-millisecond round-trip time and a packet size (MTU) of 1,500 bytes, a TCP stream would have to have of the order of 8,000 packets in flight continuously—the equivalent of 12 megabytes. A 100 millisecond round-trip time is approximately an East-West coast transfer. The TCP protocol is designed with the premise that the random loss rate in network components is insignificant compared to the loss rate due to congestion. Thus, the TCP sending rate is a function of the packet loss rate (assumed congestion) in the network. The packet sending rate drops dramatically as a response to a congestion event (packet loss); then the sending rate increases slowly until the next congestion event is encountered.[9] The ESnet network must be designed to minimize end-to-end single stream TCP packet loss as well as support alternatives to TCP. NERSC requires

- Jumbo packet support and large interface buffers throughout ESnet
  TCP recovery happens faster with jumbo (9000 byte) packets. Also, because it minimizes CPU interrupts, larger packets are more efficient for end systems to send and receive. Network interfaces must be sufficiently large enough to minimize packet loss so that end hosts can realize there is congestion in the network and perform a graceful slowdown.

- Support for high bandwidth UDP streams
  Today, UDP is widely used as an alternative transport protocol for scientific data. In particular, remote visualization depends on using UDP to improve EEB as well as reduce latency and improve responsiveness for WAN distributed interactive graphics applications. One reported example showed EEB increasing from 25 to 88 percent of line rate for a multi-gigabit network.[10]

- Non-IP (e.g. Fiberchannel) based flows
  As part of the preparation for the data analysis for the Planck Satellite project, LBNL's Julian Borrill is already requesting that NERSC's computational systems and NASA's Columbia systems mount a common set of disks over the network. While this can currently be done using IP as the transport protocol, it performs

[9] "Deep Scientific Computing Requires Deep Data", William T. C. Kramer, Arie Shoshani, Deborah A. Agarwal, Brent R. Draney, Guojun Jin, Gregory F. Butler, and John A. Hules. IBM Journal of R&D Special Issue on Deep Computing, 2003

[10] "Grid-DistributedVisualizations Using Connectionless Protocols", E. Wes Bethel and John Shalf; IEEE Computer Graphics and Applications, Mar/Apr 2003, pages: 51- 59

poorly. Since NERSC has other projects considering such arrangements for distributed disks ESnet should provide the capability to transport protocols natively without the overhead and restrictions of the IP layer. As one example of the technical issues being solved, the Fiberchannel MTU is 64kB - a size that is not expected to be available in IP routers or Ethernet switches anytime soon.

- Flow tracking capability
  When a NERSC user has an end-to-end bandwidth problem, NERSC network staff assists them by tracking the flow from NERSC to their local machine. To do this, NERSC staff needs access to network data throughout ESnet. While ESnet projects such as NetInfo provide some information, it is not sufficient to track a single flow. NERSC requires an ESnet array of "network microscopes" that would permit NERSC staff into ESnet peering points and view flow rates and packet loss for an individual TCP sockets. We would also like the ability to generate policy based traffic snapshots for analysis.

2.8.6 Additional services

- DOE wide transport fabric for authentication
  NERSC wants 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.[11]

- 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 which would guarantee a minimum 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 CERN and ITER.

- Intrusion Detection (ID) monitoring and better response to Denial of Service attacks

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

ESnet should run ID systems at hubs and peering points to detect attacks and provide data to enable better response. Finer grained ways to respond rather than simply dropping the peering link need to be investigated.

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 DOEgrids certificates will be automatically trusted by these browsers.

2.8.7 NERSC ↔ ESnet link funding

Over the past 5 years, the NERSC program has paid ESnet approximately $1.3M for the NERSC network link to ESnet. As demonstrated in the "NERSC and LBNL ESnet Border traffic by category" Table 2 above**,** the network traffic of a DOE site with its several thousand workstations is very different than the network traffic of an HPC center. HPC Centers like NERSC must optimize their network and computer security for their HPC users and resources. This is best accomplished by having their own ESnet link independent of their home-site link and the cost for the link should be covered by ESnet.

Feature	**Science Instruments and Facilities**	**Process of Science**	**Anticipated Requirements**	
Time Frame			**Network**	**Network Services and Middleware**
Near-term	• Large supercomputer center • Broad user base • Large HPSS storage system • NERSC5 system (~75 TFLOPs) in 2008	• Large data transfers requiring low packet loss • Non-TCP (UDP) data transport protocols	• 10Gbps • Additional 10Gbps link for special projects / dedicated bandwidth • 20-40Gbps in 2-3 years • Jumbo Frames • DoS mitigation • Native transport for non-IP traffic (e.g. fibrechannel)	• Grid / PKI infrastructure • Network and computational co-scheduling • Flow tracking capability for troubleshooting • Distributed infrastructure for One Time Password hardware tokens
5 years	• NERSC6 (~150 TFLOPs) in 2010	•	• 100Gbps	•
5+ years	•	•	•	•

2.9 *National Leadership Computing Facility (NLCF)*

2.9.1 Executive Summary

The National Leadership Computing Facility (NLCF) at the Oak Ridge National Laboratory is one of two large DOE supercomputer centers, and houses the largest unclassified supercomputers in DOE. To enable NLCF users to fully utilize the NLCF's capabilities, the NLCF staff feel that the NLCF should be connected to ESnet at the same speed as ESnet's backbone.

2.9.2 The National Leadership Computing Facility

The NCCS currently houses a 5000 node Cray XT3 with a disk subsystem capable of saturating its 40+ Gigabits/second of network connectivity. As this year's NLCF allocations scale to fully use the XT3, we expect the NCCS users to need the full extent of this capacity. To accommodate this bandwidth, the NCCS has made a substantial investment in local-area network capacity. Likewise, ORNL has made a substantial investment in wide-area capacity to accommodate the NLCF's geographically diverse user base. Although we cannot identify the specific geographic locations from which our users will come, we do know that they will make heavy use of other DOE facilities. For this reason, we believe that ORNL should be connected at no less than ESnet backbone rates. This is the first step in guaranteeing that users can efficiently use the NLCF resources while incurring the smallest number of bottlenecks

Feature			**Anticipated Requirements**	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Network**	**Network Services and Middleware**
Near-term	• Major supercomputer center • Broad user base		• Backbone bandwidth parity	•
5 years	•	•	• Backbone bandwidth parity	•
5+ years	•	•	•	•

2.10 *Nuclear Physics*

2.10.1 Executive Summary

The Relativistic Heavy Ion Collider (RHIC) is the largest facility of its kind to date, and it is becoming the world leader in the scientific quest toward understanding how mass and spin combine into a coherent picture of the fundamental building blocks nature uses for atomic nuclei. It is also providing unique insight into how quarks and gluons behaved collectively at the very first moment our universe was born. The main RHIC experiments, Phenix and STAR, are collaborations spanning many countries and are composed of hundreds of collaborators each. The wide scope of collaboration points directly to a requirement for high-quality, flexible connectivity between collaborators to enable effective communication and data analysis. In addition, the scale of data movement required by the RHIC experiments relies on the availability of high-bandwidth, production quality network infrastructure. Phenix relies heavily on the ability to sustain large data transfers (the recent movement of 6.8 million events, or 270TB of data in 11 weeks from BNL to the CC-J facility in Japan is an example) to remote institutions for analysis. STAR makes extensive use of Grid computing, using a data-grid model with tools adapted from the Particle Physics Data Grid (PPDG). RHIC is sited at the Brookhaven National Laboratory. ESnet provides all of BNL's network connectivity, and the RHIC experiments assume the ability to move large quantities of data, collaborate efficiently, and effectively distribute computation in the service of the experiments. ESnet provides these capabilities. The RHIC scientists have estimated that the combined network requirements of the RHIC experiments will exceed 20Gbps in 2007, and will reach 70Gbps in 2010.

2.10.2 Relativistic Heavy Ion Collider

The Relativistic Heavy Ion Collider (RHIC) program is a Nuclear Physics program composed of several experimental halls, a world-class scientific research facility located at the Brookhaven National Laboratory (BNL). RHIC is the biggest facility of its kind to date. It is becoming the world leader in the scientific quest toward understanding how mass and spin combine into a coherent picture of the fundamental building blocks nature uses for atomic nuclei. It is also providing unique insight into how quarks and gluons behaved collectively at the very first moment our universe was born. The main RHIC experiments, Phenix and STAR, are collaborations spanning over many countries and thousands collaborators.

Currently reaching the Petabyte scale data recording overall per year (10^{12} bytes), the raw data rates envisioned by the RHIC experiment's program will grow by an order of magnitude by 2008, reaching an online data acquisition rate of 1 GB/sec and making data management and distribution an ever growing challenge with an overall increase of the problem by an order of magnitude. To face the challenges caused by the size of those datasets while preserving the Physics quality and turn around, the RHIC experiments have resorted to either a distributed computing model or an allocation of remote dedicated or opportunistic resources. The Phenix experiment for example has recently moved 6.8 billion events (or 270 TB of data) to their Japanese CC-J facility over the 11

weeks of running where it was further reconstructed and analyzed. The STAR computing model has for the past few years evolved around a data-grid model with help from tools developed from with the Particle Physics DataGrid (PPDG) collaboratory: processed data is made immediately available to remote sites where computing resources may be available. More recently, the onset of user based analysis on the STAR/Grid has raised questions of access to massive data sets, their availability and the problem of data movement and caching at smaller but numerous Tier2 sites.
However, this model remains at modest scale (mainly Tier0 to one Tier1 center per large RHIC experiment) and its future in the RHIC-II era is uncertain as the data pool augments. Multiple production passes, and therefore increased science quality and scientific deliverables on shorter time scales, could not be envisioned without strong coast to coast network connectivity. The ability of harvesting remote resources built by our international collaborators for either data mining or user analysis coupled to the global approach of today's computing world demands network consolidation to ensure and maintain world leadership of the RHIC program without taking any short cuts or prioritization of the science. In addition, and while the RHIC program is a world leader in its kind, several of the RHIC remote institutions do not have easy access to the data due to two factors: network latencies and network backbone infrastructure. Those would highly benefit from a model involving data on demand such as the one developed in STAR for transferring data coast to coast. Data on demand would provide data proximity allowing for scientific equity across our many institutions and help, to some extend, break the digital divide.
To estimate the resources needed by the RHIC mid-term program and RHIC-II era, algorithms were developed for estimating for anticipated RHIC running scenarios. The network estimates were based on the acceptable fractional data transfer rates comparing to the experiment's data acquisition capabilities. The result and summary of these algorithms, derived primarily on the expected amount of raw data collected, are showed in the table below.

Table 2.10.1 RHIC and RHIC-II planning derived data sets and estimated WAN needs

	FY05	**FY06**	**FY07**	**FY08**	**FY09**	**FY10**	**FY11**	**FY12**
STAR Data (TB/year)	300	365	350	540	1360	1915	2610	2610
PHENIX Data (TB/year)	500	400	800	1000	1000	1500	1500	1500
Total Annual Raw Data (TB/year)	800	765	1150	1540	2360	3415	4110	4110
Required WAN bandwidth (Megabytes per second)	276	1500	2737	3485	6066	8719	11029	12185

Table 2.10.2 RHIC Requirements Summary

Feature			Anticipated Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Network**	**Network Services and Middleware**
Near-term	• RHIC at BNL with 4 current experiments (STAR, PHENIX, BRAHMS, PHOBOS)	• PHENIX data transferred to Japan for analysis (RIKEN) • STAR data transferred to Tier1 site (PDSF) • STAR analysis uses Data Grid model (OSG) • Support of two to five STAR Tier2 emerging sites, including path to South America	• 12Gbps Wide Area bandwidth	• Grid infrastructure • Reliable transfer of large datasets coast-to-coast and to Japan.
5 years	• RHIC (STAR, Phenix) and RHIC-II support	• Enable data analysis for STAR's collaborators spanning over 53 institutions and 12 countries (pre-paced data) especially South America, Russia, India, China and institutions in the EU • Expand to dynamic migration of hot-spot datasets (data on demand) • All simulation needs for RHIC-II done on Grid infrastructure	• 70Gbps Wide Area bandwidth	• Bandwidth and Service guarantees, quality of service • Network bandwidth predictions, guaranteed high bandwidth (accounting) • Secured access to data • Grid infrastructure, schedulers, brokers, planners, co-scheduling • Object level access • Distributed database
5+ years	•	• Real-time data analysis, event visualization • Underlying object-on-demand oriented analysis (rather than file) stabilizing network needs • Extensive use of data caching • Interoperability with other grids	• 100Gbps Wide Area bandwidth	• Resource discovery • Database access on Grid • Dynamic data grid, computational grid relies on "closest location for data" (coupling CPU and network)

2.11 *Spallation Neutron Source*

2.11.1 Executive Summary

The Spallation Neutron Source (SNS) is a new facility at the Oak Ridge National Laboratory that is scheduled to come into production in 2006. The SNS is expected to generate 160 gigabytes of data per day, or about 50 terabytes per year. The bulk of this data will be analyzed using remote computational resources, so bulk data movement capabilities are a significant requirement. Productivity is critical, since a particular scientist will get at most 2 days of instrument time per year. This means that network capacity and reliability are very important enablers for the effective use of this new resource, especially since real-time analysis is used to fine-tune the experiment and therefore make more effective use of the instrument. Network bandwidth needs in the near term are 100 to 160 megabits per second sustained (640 megabits per second peak), and in 5 years the requirements are expected to jump to 2 gigabits per second sustained, with peaks of 10 gigabits per second likely.

2.11.2 Spallation Neutron Source Network Requirements

Six DOE laboratories are partners in the design and construction of the Spallation Neutron Source (SNS), a one-of-a-kind facility at Oak Ridge, Tennessee, that will provide the most intense pulsed neutron beams in the world for scientific research and industrial development. When completed in early 2006, the SNS will enable new levels of investigation into the properties of materials of interest to chemists, condensed matter physicists, biologists, pharmacologists, materials scientists, and engineers, in an ever-increasing range of applications.

The SNS supports multiple instruments that will offer users at least an order of magnitude performance enhancement over any of today's pulsed spallation neutron source instruments. This great increase in instrument performance is mirrored by an increase in data output from each instrument. In fact, the use of high-resolution detector arrays and supermirror neutron guides in SNS instruments means that the data output rate for each instrument is likely to be close to two orders greater than a comparable U.S. instrument in use today. This, combined with increased collaboration among the several related U.S. facilities, will require a new approach to data handling, analysis, and sharing.

The high data rates and volumes from the new instruments will call for significant data analysis to be completed offsite on high-performance computing systems. High-performance network and distributed computer systems will handle all aspects of post-experiment data analysis and the approximate analysis that can be used to support near real-time interactions of scientists with their experiments.

Each user is given a specific amount of time (0.5 to 2 days) on an instrument. The close to real-time visualization and partial analysis capabilities, therefore, allow a user to refine the experiment during the allotted time. For the majority of SNS user experiments, the material or property being

studied is novel, and this capability is essential for the experimentalist to focus in on the area of interest and maximize the science accomplished in the limited amount of beam time.

In this scenario, the combined data transfer between the 12 SNS instruments and a distributed computer network for real-time data mapping is estimated to be a constant 2 Gbits/sec (assuming 50% of users using real-time visualization). The return data stream to servers managing the visualization and analysis tasks as well as communicating to the users across local area networks (LANs) and/or the Internet likely will be around 280 Mbits/sec (dominated by the four- and three-dimensional response maps). The servers (one for each instrument) would generate selected views of the response function as well as send the response function back out to the distributed computer network for quick/partial analysis.

It is anticipated that analysis of experimental data in the future may be achieved by incorporating a scattering law model within the iterative response function extraction procedure. These advanced analysis methods are expected to require the use of powerful offsite computing systems, and the data may transit the network several times as experiment/experimenter/simulation interaction converges to an accurate representation.

Table 2.11. Spallation Neutron Source Requirements Summary

Feature / Time Frame	Science Instruments and Facilities	Process of Science	Anticipated Requirements	
			Networking	Middleware
	(Facility comes on-line in 2006)			
Near term	• The 12 instruments at the SNS will operate about 200 days/year and generate an aggregate 160 Gbytes/day • The data analysis will be accomplished mostly on computing systems that are remote from the SNS		• 100-160 Mbits/sec sustained • 640 Mbits/sec peak	• Workflow management • Reliable data transfer
5 years	• Neutron scattering instruments operate 24 hr 7 days a week during facility run periods, real time data visualization, some real time analysis capabilities, and security to modify experiment conditions by a user at his/her hotel via an internet browser will be required.	• Real-time data analysis and visualization will enhance the productivity of the science done at SNS, which runs 24 hr/day.	• 2 Gbits/sec sustained	• Security (authentication and access control) to permit direct interaction with the instrument remotely.
5-10 years	• Statistical scattering models will be incorporated into analysis code requiring supercomputer levels of remote computing.	• Iterative analysis of the data with the use of models running on supercomputing systems will produce much more accurate results and understanding.		