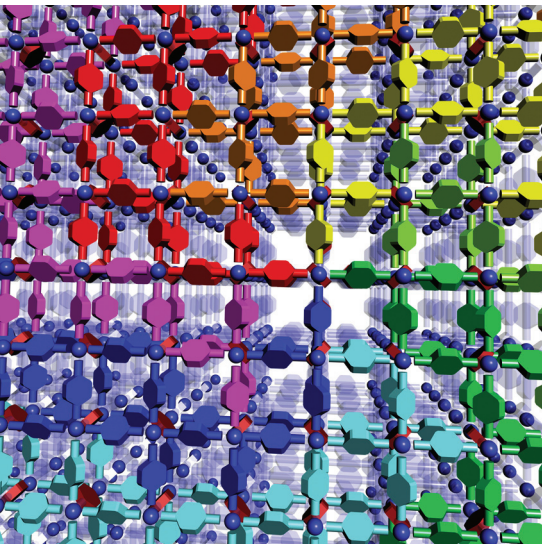
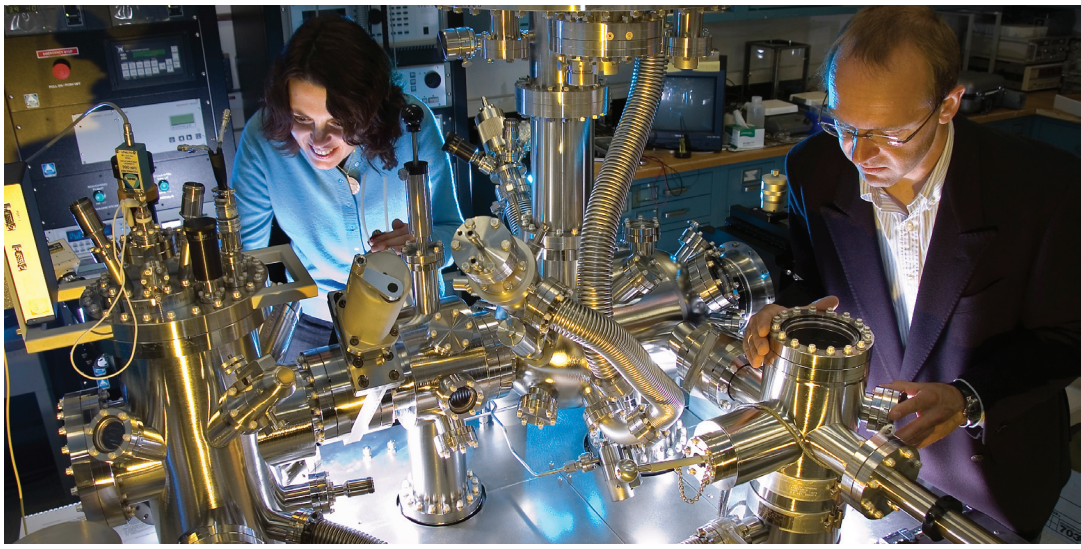
BES Science Network Requirements

Report of the Basic Energy Sciences
Network Requirements Workshop

Conducted September 22 and 23, 2010

ESnet
Energy Sciences Network





BERKELEY LAB
Lawrence Berkeley
National Laboratory

U.S. DEPARTMENT OF
ENERGY
Office of Science

BES Network Requirements Workshop

Office of Basic Energy Sciences, DOE Office of Science
Energy Sciences Network
Gaithersburg, MD — September 22 and 23, 2010

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, Facilities Division, and the Office of Basic Energy Sciences.

This is LBNL report LBNL-4363E

Participants and Contributors

Alan Biocca, LBNL (Advanced Light Source)

Rich Carlson, DOE/SC/ASCR (Program Manager)

Jackie Chen, SNL/CA (Chemistry/Combustion)

Steve Cotter, ESnet (Networking)

Eli Dart, ESnet (Networking)

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

Jim Davenport, DOE/SC/BES (BES Program)

Alexander Gaenko, Ames Lab (Chemistry)

Paul Kent, ORNL (Materials Science, Simulations)

Monica Lamm, Ames Lab (Computational Chemistry)

Stephen Miller, ORNL (Spallation Neutron Source)

Chris Mundy, PNNL (Chemical Physics)

Thomas Ndousse, DOE/SC/ASCR (ASCR Program)

Mark Pederson, DOE/SC/BES (BES Program)

Amedeo Perazzo, SLAC (Linac Coherent Light Source)

Razvan Popescu, BNL (National Synchrotron Light Source)

Damian Rouson, SNL/CA (Chemistry/Combustion)

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

Bobby Sumpter, ORNL (Computer Science and Mathematics and Center for Nanophase Materials Sciences)

Brian Tierney, ESnet (Networking)

Cai-Zhuang Wang, Ames Lab (Computer Science/Simulations)

Steve Whitelam, LBNL (Molecular Foundry)

Jason Zurawski, Internet2 (Networking)

Editors

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

Table of 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 (SC), 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 September 2010 ESnet and the Office of Basic Energy Sciences (BES), of the DOE Office of Science, organized a workshop to characterize the networking requirements of the programs funded by BES.

The requirements identified at the workshop are summarized below, and described in more detail in the case studies and the Findings section.

There are significant changes coming to the BES facilities. Data volumes are set to increase significantly, and the science process will change dramatically for many collaborations. The current science process for a majority of users at BES facilities involves collecting experimental data at the facility and then physically transporting the data to a home institution using portable media – the network is not used in the transfer of the scientific data. However, data volumes are going to increase significantly in the next few years (to 70TB/day or more) – much faster than portable media is likely to increase in capacity. This means that data must be transferred over the network. While there are other examples of scientific communities that can transfer these data volumes (supercomputer centers, HEP), the BES facility community does not have institutional knowledge of the design and use of high-performance data transfer systems. This problem can easily be solved, but it will require a conscious effort to build institutional knowledge and capability at the BES facilities.

The second change coming to the BES facilities is the adoption of remote control and data streaming technologies. These are likely to significantly increase scientific productivity. However, in order for these technologies to work well, multiple institutions must coordinate the scheduling and configuration of their networking and computational resources. ESnet has tools and technologies (such as the Science Data Network and the OSCARS circuit scheduling system) that are well-suited to the task, but the sites and facilities will need to work with ESnet to develop operational models that work for the BES facility community.

The BES community has a need, shared with many other communities, for data transfer tools that are easy to use, well-supported, and permitted by the cybersecurity organizations at the sites and facilities. The only tool that is easy to deploy and use for many collaborations is the SSH suite, consisting of SSH, SCP, and SFTP (rsync over SSH is also used). The SSH suite has well-understood performance limitations that make it unsuitable for use with the large data sets characteristic of modern science. There is a clear and present need for an alternative that is widely deployed and well-supported. Other tools such as GridFTP could meet this need, but documentation and expertise are

often lacking, and cybersecurity organizations typically oppose the deployment of Grid tools.

Several attendees discussed the use of cloud services, such as those operated by Amazon. This is in line with the activities of other communities, such as Genomics.

Collaboration tools are a recurring theme in this community – there is an ongoing need to communicate with people on the beamlines at BES facilities, or to monitor the health of the experiments. Tools such as Skype and EVO are used, but there was a need expressed for better-supported and more capable collaboration tools.

Simulation science continues to generate ever-greater data volumes. At this stage, simulation-based science (such as combustion and computational materials science) can generate data sets of essentially arbitrary size. The analysis of these data sets is an ongoing challenge that is shared with other communities.

2 Workshop Background and Structure

The strategic approach of the Office of Advanced Scientific Computing Research (ASCR – ESnet is funded by the ASCR Facilities Division) and ESnet for defining and accomplishing ESnet's mission involves three areas:

1. Work with the DOE Office of Science (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, including the data characteristics of scientific instruments and facilities, regarding what data will be generated by instruments and supercomputers coming on-line over the next 5-10 years. Also by examining the future process of science: how and where will the new data be analyzed and used, and how the process of doing science will change over the next 5-10 years.
>
> B) Observing current and historical network traffic patterns and trying to determine how trends in network patterns predict future network needs.

The primary mechanism of accomplishing (A) is the 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 repeats every three years:

- Basic Energy Sciences (materials sciences, chemistry, geosciences) (2007, 2010)
- Biological and Environmental Research (2007, 2010)
- Fusion Energy Science (2008)
- Nuclear Physics (2008)
- Advanced Scientific Computing Research (2009)
- High Energy Physics (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 September 2010 ESnet and the Office of Basic Energy Sciences (BES), of the DOE Office of Science, organized a workshop to characterize the networking requirements of the science programs funded by BES.

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 Basic Energy Sciences

Basic Energy Sciences (BES) supports fundamental research to understand, predict, and ultimately control matter and energy at the electronic, atomic, and molecular levels in order to provide the foundations for new energy technologies and to support DOE missions in energy, environment, and national security. The BES program also plans, constructs, and operates major scientific user facilities to serve researchers from universities, national laboratories, and private institutions. The BES program is one of the Nation's largest sponsors of the natural sciences and it funds experimental, computational and theoretical research at more than 160 research institutions through three divisions.

The *Materials Sciences and Engineering (MSE) Division* supports fundamental experimental and theoretical research to provide the knowledge base for the discovery and design of new materials with novel structures, functions, and properties. This knowledge serves as a basis for the development of new materials for the generation, storage, and use of energy and for mitigation of the environmental impacts of energy use. The *Chemical Sciences, Geosciences, and Biosciences (CSGB) Division* supports experimental, theoretical, and computational research to provide fundamental understanding of chemical transformations and energy flow in systems relevant to DOE missions. This knowledge serves as a basis for the development of new processes for the generation, storage, and use of energy and for mitigation of the environmental impacts of energy use. The *Scientific User Facilities (SUF) Division* supports the R&D, planning, construction, and operation of scientific user facilities for the development of novel nano-materials and for materials characterization through x-ray, neutron, and electron beam scattering; the former is accomplished through five Nanoscale Science Research Centers and the latter is accomplished through the world's largest suite of synchrotron radiation light source facilities, neutron scattering facilities, and electron-beam microcharacterization centers.

The Office of Basic Energy Sciences in the U.S. Department of Energy’s Office of Science has also established 46 Energy Frontier Research Centers (EFRCs). These Centers involve universities, national laboratories, nonprofit organizations, and for-profit firms, singly or in partnerships, and were selected by scientific peer review and funded at $2-5 million per year for a 5-year initial award period. These integrated, multi-investigator Centers will conduct fundamental research focusing on one or more of several “grand challenges” and use-inspired “basic research needs” recently identified in major strategic planning efforts by the scientific community. The purpose of these Centers will be to integrate the talents and expertise of leading scientists in a setting designed to accelerate research toward meeting our critical energy challenges. In addition to the EFRCs, the BES-Funded Joint Center for Artificial Photosynthesis (JCAP) is led by the California Institute of Technology (Cal Tech) in partnership with the U.S. Department of Energy's Lawrence Berkeley National Laboratory (Berkeley Lab), will bring together leading researchers in an ambitious effort aimed at simulating nature's photosynthetic apparatus for practical energy production.

As highlighted throughout the following pages, the data sets generated and used by BES scientists continue to increase in size and Internet access to these data sets continues to be a vital requirement that is fulfilled by ESnet. The studies herein illustrate this case for theoretical, computational and experimental fields of inquiry and provide a means for understanding how availing future data to the larger community is coupled to future network needs. This workshop has identified a number of issues that ESnet can help address for the coming years.

4 Advanced Light Source at LBNL

4.1 Background

The Advanced Light Source (ALS) is a national user facility at Berkeley Lab that generates intense light for scientific and technological research. The ALS is one of the world's brightest sources of ultraviolet and soft x-ray beams. The ALS operates 45 beamlines and hosts more than 2,000 distinct users annually. ALS is a 24-hour operational facility and beams are available to users in excess of 4,000 hours per year.

4.2 Key Local Science Drivers

4.2.1 Instruments and Facilities

With 45 ALS beamlines collecting data simultaneously, network data flow is significant, and improvements to detectors, sample robotics and experimental control positioning and software continue to increase the data rates and overall data volume. The ALS assists users both on-site and off, and datasets are shared over the network and via portable media. The data rates and size can vary widely depending on the experiment. Not all experiments are data intensive, but there are a number of detectors struggling with high data accumulation rates. Examples of these include programs in x-ray Microdiffraction, x-ray Scattering, x-ray Tomography and Protein Crystallography. Future instruments with high data rates include the COSMIC Nanosurveyor. Of the existing programs, the most data intensive at this stage are x-ray tomography and protein crystallography. Coming online soon are high bandwidth fast readout detectors for x-ray Scattering and Microdiffraction that will produce up to 5Gbps of data that will need to be streamed to remote supercomputers and for processing the resulting images returned to the ALS in near real-time. New detectors for Protein Crystallography may increase the peak data rate to nearly 0.6 Gbps per beamline over the next few years, and ALS has many PX beamlines so the total PX network rate could be 4 Gbps.

Other local resources include LBNL computational clusters such as Lawrencium, and various local clusters on the beamline and soon in the new User Support building server room that will come online in a few months.

The LBNL network core is presently 10Gbps. Multiple 1Gbps networks (~20 each) feed the ALS experimental floor and control systems and several beamlines are already in the process of obtaining 10Gbps network feeds to meet their specific high bandwidth requirements.

The new User Support Building Server room is being outfitted now to provide a better environment for computing equipment and to remove noise and heat from the experimental areas. It is being wired with more than 400 fibers from the experimental floor and network core for data transmission. Experiments with high data rate requirements are using multiple 1Gbps fiber paths to carry the load.

4.2.2 Process of Science

X-ray Microdiffraction

X-ray micro-diffraction on BL12.3.2: The detector is a Pilatus 1 M, with around 50 frames / sec, of 1e6 pixels (probably using a 20 bit word). This is the maximum continuous acquisition rate. The detector records diffraction patterns (1M pixels) for each point in an image, which might be 10,000 – 100,000 image pixels. Total datasets therefore can be 400GB. This data must be sent to supercomputer centers for analysis, with results sent back to the ALS and to users' sites.

Scanning X-ray microdiffraction (uXRD) is a synchrotron technique, which consists of raster scanning a sample under a focused X-ray beam (typically less than 1 μm in size) and collecting at each location an x-ray diffraction pattern using an area detector. The analysis of the produced arrays of diffraction pattern yields maps of different properties of the sample, such as phase distribution, grain orientation, strain/stress distribution, and dislocation density distribution. This technique has a wide range of applications in the study of the micro-mechanics and characterization of polycrystalline materials

uXRD relies on savvy beamline optics, state-of-the-art detector technology and high CPU performance as the analysis can be computationally intensive. A typical uXRD scan contains thousands of data points, each consisting of a 4 MB diffraction image, resulting in the rapid generation of gigabytes of data that need to be analyzed. Laue diffraction patterns are typically far more complex than the average monochromatic x-ray beam, Electron diffraction or Kikuchi lines patterns, necessitating the use of more sophisticated routines for background fit and removal, peak search and fitting and multigrain indexing and resulting in significant downtime in the calculation process, especially for low-symmetry phases and highly deformed samples.

Until June 2010, the area detector used for uXRD at the ALS was a fiber-optic coupled x-ray CCD detector and the data collection and analysis speeds were limited by the readout time of the detector. With the new Dectris Pilatus detector, with negligible readout time, better signal/noise performances and often subsecond exposure time per point, the data collection has become an order of magnitude faster than the data analysis. It can take a few minutes to collect the data and a few hours to process them, which is more than a mere inconvenience when allocated beamtime is limited and fast decision making is required (for instance in the case of time-resolved experiments). uXRD would therefore benefits from real-time analysis with diffraction data fed to a fast computing machine as they are taken.

To answer the challenge, the code for data processing called XMAS (for X-ray microdiffraction Analysis Software) initially developed at the ALS for a single or dual processor PC desktop, has been adapted to work on a LINUX cluster (each node of the cluster analyzing a different image). A new version of the code (called FOXMAS or Fast On-line XMAS, the "S" standing now for "services" rather than "software") written in C and running on a cell blade cluster system at SHARCNET is being developed at the University of Western Ontario (UWO) by a team of physicists and computer scientists (http://www.anise-project.com/foxmas.php). This application has been written using the framework of the ANISE (Active Network for Information from Synchrotron Experiments) project, intended to provide streaming analysis software for synchrotron

experiments. Initial testing indicates that the data processing can be speed up by at least an order of magnitude compared to a regular PC.

However since the core code is running at UWO, the data taken at the beamline needs to be rapidly sent to the institution for processing. This requires a relatively large amount of bandwidth. A virtual circuit is presently being established between the ALS and the UWO (initially at 1Gbps but can be easily upgraded to 10Gbps), as a collaborative effort between LBLnet, ESnet, and CANARIE to stream the data directly from the ALS data storage computers to the UWO SHARCNET facility that runs the code. A similar virtual circuit has already been established between the VESPERS beamline at the Canadian Light Source (CLS) and UWO. Besides the ones with the ALS and CLS, similar virtual circuits are in project between other synchrotron facilities equipped with a X-ray microdiffraction beamline. Science Studio is the visual GUI interface at the beamline that allows one to set initial calculation parameters, perform the data transfer, and collect the results in a way completely transparent to the users. The ANISE project is however limited in time with initial funding for another one and half year but shows how fast networking can be useful to perform experiments at a beamline that requires real time data interpretation obtained by analysis on remote clusters.

Small Angle X-ray Scattering (SAXS)

The BL7.3.3 SAXS beamline has installed two high rate detectors similar to the Pilatus 1M. The most demanding experiments are time resolved, recording data as a sample changes. Here the experiment is 'non-imaging' so that the total data set is much smaller than uXRD, but still presents difficulty. The issue in this case is the computational analysis, which needs to be done remotely.

BL12.3.1 SAXS for biological sciences is relatively new and has limited computational resources. Datasets are collected from a CCD detector as raw images that must be extracted and integrated into intensity files using beamline specific computer programs. The current experimental configuration generates maximally 2 GB/day of data. Following data integration, datasets are reduced and analyzed collectively to assess data quality. This requires the user to maintain an open network connection with the beamline computers to transfer files for analysis and visualization. SAXS data analysis is highly interactive. Following data reduction, modeling of the data is performed using web applications that transfers and queues the reduced datasets to the beamline computational cluster. Following modeling, data is transferred back to the user and additional round(s) of modeling is performed for model refinement. Unlike PX, SAXS data analysis requires both beamline and user specific computational resources to be involved in an active network connection.

SAXS is experiencing a renewed interest amongst the scientific community as the technique provides structural information on every type of sample, e.g. folded, flexible and unfolded proteins. User time at beamline 12.3.1 is highly requested through a peer-reviewed proposal system where there is an excess of 6 proposals per user shift. As such, there is an enormous need to commission new beamlines dedicated to SAXS throughout the international scientific community. We anticipate a large increase in the network requirements at beamline 12.3.1 due to the development of web applications for SAXS data analysis and the launching of the SAXS database BioIsis.net. Beamline 12.3.1 is

currently developing a family of web applications for SAXS data analysis in the near-term that will be available for use by the international community. This will undoubtedly place enormous bandwidth requirements on the network due to the real time data analysis the beamline cluster will be performing.

Protein Crystallography (PX) Detector Improvement

Data output of a typical protein crystallography beamline is currently 30GB per day. This will increase significantly with a new detector now under development. It is expected that other PX beamlines will adopt this technology.

With this new detector each image will be 32MB, and will be recorded and saved to disk in 0.1 second. A complete data set will be about 2000 images, with no delay during the recording process. In 200 seconds each data set of 64GB will be transferred to disk. Users may process up to 100 crystal samples in one day, so data transfer and storage requirements could be up to 6.4TB per day just for the raw data. Processed data will add an additional data volume.

4.3 Key Remote Science Drivers

4.3.1 Instruments and Facilities

One new trend in ALS experiment evolution is a new requirement for low latency transfers to a remote computational facility for near real time processing. Beamlines with high rate detectors such as uXRD, SAXS, and PX have difficulty or are unable to provide enough computation for local real-time processing and therefore must depend on remote processing at facilities such as NERSC, OLCF, ALCF and the University of Western Ontario. This must be near real time to evaluate the experimental progress and make any needed adjustments at the beamline.

Some users travel long distances to use ALS facilities due to specific beamline capabilities or to the unavailability of beam time at a more local facility. The user data transfers are thus not exclusively “regional” and many involve long network paths to their home institutions.

Many users are part of collaborations that have distant members, so again the data may not only be sent long distances, but it may be delivered to more than one destination.

4.3.2 Process of Science

Same as above.

4.4 Science Drivers – the next 2-5 years

4.4.1 Instruments, Facilities and Science Process

The "nanosurveyor" is a new instrument in development for the recently funded COSMIC beamline. This instrument includes a detector collecting megapixel images at 200 Hz producing sustained data rates of 1.5 terabytes per hour. Ideally the experiment requires the detector data be processed in near real time to provide feedback to the experiment and to remote collaborators for rapid evaluation.

There are other detectors producing similar amounts of data at ALS such as fast readout large CCDs, and other Pilatus detectors. In 5 years the aggregate data rate will probably be well over 10 terabytes per hour. Similar experiments will be taking place at SLAC. These experiments are quite new now so the data is captured and processed slowly and locally. Over time this will be improved to take scientific advantage of valuable beam time and the detector's full capabilities.

4.4.2 Local Network

The local network upgrades will continue as needed. The ALS Accelerator Controls Group works closely with LBL network services to plan and respond to growing facility network needs. The next few years will see upgrades from the existing 1Gbps to 10Gbps rates on many ALS local networks. The ALS connection to the LBL network core is already at 10Gbps and will likely be increased on this time scale as faster equipment becomes advantageous and cost effective. These upgrades are based on both network traffic monitoring as well as coordinated planning.

4.4.3 Remote Network

The COSMIC Nanosurveyor discussed above is an example of an ALS experiment that will drive the remote network requirements for capacity and latency. Aggregate rates of approximately 10 TB per hour (45Gbps) are expected.

4.5 Beyond 5 years – future needs and scientific direction

Detectors continue to improve (similar to Moore's law, data throughput is doubling roughly every 18 months) and the quantity of high rate detectors in use at ALS will increase as experimental stations are upgraded and additional detectors are installed. The increase in data throughput is a function of funding for these upgrades.

New Facilities such as the planned Next Generation Light Source (NGLS) are under development now but will not go online for ten or more years.

Scientists designing experiments for the NGLS are planning to capture Megapixel images at megahertz rates to record dynamics. This will require significantly improved local and remote networking. Data rates for a single camera could reach 2 Tbps. The beamlines will likely operate thousands of hours per year.

4.6 Middleware Tools and Services

The ALS staff and users are expected to use tools such as GridFTP and bbcp to enhance the throughput of large data transfers in the future. Use of these tools is minimal at this time (as far as we know). Education and convenience are likely the reason.

Use of cloud services is being evaluated on several fronts. The new ALS Web server is being developed in the Amazon cloud. Some testing is also underway for Amazon compute cluster services where a set of machines in the cloud is requested that have optimized local connectivity between them for computational purposes. Use of cloud storage such as Amazon S3 is also being tested.

Authentication and data protection are currently handled on a per beamline ad-hoc basis. There may be value in moving to a more standardized solution in the future.

4.7 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• 45 ALS Beamlines.	• SAXS, uXRD, PX	• Detectors producing 5 Gbps of image data to be processed remotely and returned dominate the network dataflow	• Real-time to local servers for relay, analysis, and storage	• Near real-time • Some Data are transferred to NERSC for processing and returned to ALS and other institutions
2-5 years				
• Quantity of detectors with high data rates (above) will increase by approximately a factor of two or more	• More external cloud processing and storage and collaborations, COSMIC Nanosurveyor	• Aggregate data rate approx 10 Gbps	• Real-time	• Near real-time • Data are transferred to multiple remote sites
5+ years				
• NGLS facility may have detectors of Megapixel size with Megahertz readouts.	• Use of external data Processing and storage is expected to increase	• Data rate 2Tbps for one camera	• Real-time	• Near real-time • Many remote sites

5 Linac Coherent Light Source, SLAC National Accelerator Laboratory

5.1 Background

The Linac Coherent Light Source (LCLS), located at SLAC, began operation in 2009 as the world's first x-ray free-electron laser, producing ultra-fast pulses of coherent x-rays with unprecedented brightness. Six different x-ray instruments for exploiting the unique LCLS scientific capabilities are being developed.

Two of these instruments are already operating (Atomic, Molecular & Optical Science and the Soft X-ray Material Science), one is being commissioned (X-ray Pump Probe) and three are in the installation phase (Coherent X-ray Imaging, X-ray Correlation Spectroscopy and Matter in Extreme Conditions).

Three instruments (AMO, SXR and XPP) are located in the Near Experimental Hall (NEH) and three instruments (XCS, CXI and MEC) are located in the Far Experimental Hall (FEH).

5.2 Key Local Science Drivers

5.2.1 Instruments and Facilities

Each instrument has two or more 1Gbps edge switches for the controls network and one 10Gbps switch for the data acquisition traffic (DAQ switch). Each instrument has a dedicated controls network and a dedicated DAQ network. All edge switches connect to one router, which also provides connectivity to the SLAC central computing facilities and to the accelerator side.

Each instrument has a dedicated online data cache with one interface on the instrument DAQ switch and one interface on a shared 10Gbps switch (analysis switch), which connects to the short-term storage.

The short-term storage is shared among the instruments, it's disk based and its size is currently 2 PB. Each PB has a maximum throughput of 5Gb/sec. The science data files are kept in the short-term storage for one year and are available on tape for 10 years.

The short-term storage is in the NEH building and communicates with the tape staging system in the SLAC central computing facilities through one dual 10Gbps link. An additional dual 10Gbps link between the NEH and the SLAC central computing facilities is used to transfer the data off-site.

The offline processing is based on a batch farm and an interactive farm. Both are in the NEH and are currently made of 60 and 24 eight-core nodes respectively. Each chassis in the batch farm contains 20 blades and one switch with 1Gbps connections to the blades and one 10Gbps uplink to the analysis switch. Each chassis in the interactive farm contains 12 blades and one switch with 10Gbps connections to the blades and two 10Gbps uplinks to the analysis switch.

5.2.2 Process of Science

The scientists can monitor the data on the fly using the online system, and they can analyze the data after the files are stored in the short-term storage using the LCLS offline system. Access to the data for each experiment is granted only to the members of that experiment. The experimenters are allowed to transfer their data files to their home institution if they decide to do so.

The online monitoring is implemented by snooping on the multicast traffic between the readout nodes and the online cache. The data files are copied as fast as possible from the online cache to the short-term storage where they are made available for offline analysis and for off-site transfer.

5.3 Key Remote Science Drivers

Up to now, some of the AMO collaborations have copied their data files to the Max Planck Institute in Munich and to DESY in Hamburg. While the MPI collaborators have been able to copy their data from SLAC to Munich at an average rate of about 80MB/sec, the DESY collaborators achieved only between 2 and 20MB/sec.

5.4 Local Science Drivers – the next 2-5 years

5.4.1 Instruments and Facilities

The online system used by the AMO, SXR and XPP instruments in the NEH building will be replicated in the FEH for the XCS, CXI and MEC instruments during FY11.

The offline system will initially be an extension of the existing NEH system, but some of the future storage and processing servers may be located in the FEH if the system grows beyond the space constraints of the NEH building.

We plan to increase the short-term storage from 2 PB to 6 PB and to expand the processing farm from about 500 cores to at least 1000 cores.

5.4.2 Process of Science

Many of the instruments are expected to increase the size of their detectors in the next couple of years. The event size will go from about 2.5 MB to more than 10 MB. Data selection and data compression will be become critical.

The repetition rate will increase from the current 30-60 Hz to 120 Hz by the beginning of 2011. The typical peak data rate will increase from 100MB/sec to more than 1GB/sec.

5.5 Remote Science Drivers – the next 2-5 years

5.5.1 Instruments and Facilities

More institutions are expected to copy their data off-site during the next couple of years. Many of these will likely be in the United States and Europe, especially Germany, France and Sweden, but future collaborations are in no way limited to these countries and could come from extremely diversified areas around the world.

5.6 Beyond 5 years – future needs and scientific direction

In the long term, the offline system may grow beyond the space constraints of the NEH and FEH buildings. No decisions have been taken at this time regarding which facilities shall be used if these long-term expansions become necessary. Possible candidates that are being evaluated are the SLAC central computing facilities, some of the on-site unused experimental areas and other national labs.

5.7 Middleware Tools and Services

The file system adopted for the short-term storage is Lustre. The SLAC application `bbcp` is currently used for transferring the science data from the online cache to the offline and from the offline to remote institutions. HPSS is used for the tape staging system.

5.8 Summary Table

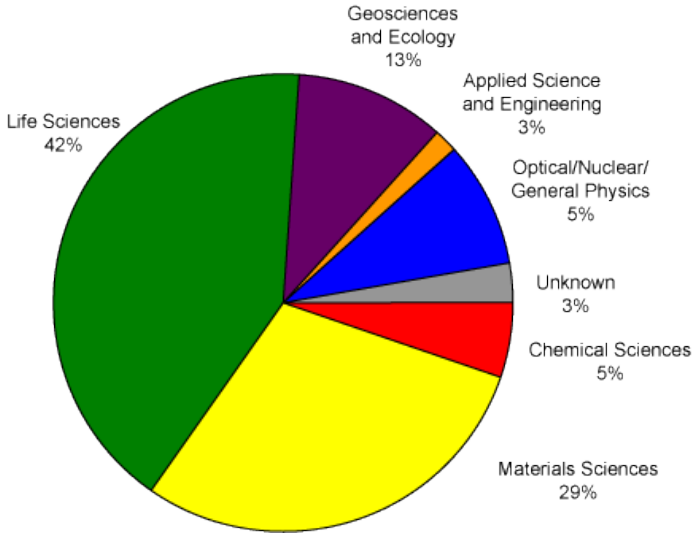
Key Science Drivers			**Anticipated Network Needs**	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• AMO • SXR • XPP • XCS • CXI • MEC	• Online monitoring • Offline analysis	• Between 2TB and 100TB per experiment • Two experiments per week	• Peak rate: 100MB/sec to 1.5GB/sec • Sustained: 200MB/sec	• One experiment data-set per week • 100-200MB/sec
2-5 years				
• Unknown	• Unknown	• Unknown	• Unknown	• Unknown

6 National Synchrotron Light Source at BNL

6.1 Background

The National Synchrotron Light Source is a national user facility that operates two electron storage rings providing intense and tunable X-ray, UV and IR synchrotron light for approximately 2200 scientists per year from more than 400 academic, industrial and government institutions. NSLS is an important national and international research resource, supporting research in diverse fields such as biology, physics, chemistry, geology, medicine, environmental and materials science (see Figure below). With about 1000 papers published annually, NSLS is one of the most prolific scientific facilities in the world; the facility's staff has won nine R&D 100 Awards and two Nobel Prices in Chemistry have been awarded based partially on work performed at NSLS: in 2003 and 2009.

NSLS Users by Field of Research (2009)



The facility has 64 experimental stations across the X-ray and UV Rings, runs 24x7 throughout the year, except during periods of maintenance and accelerator studies, for an average of about 44 weeks per year. A wide range of experimental techniques, from diffraction, protein crystallography, imaging, to all ranges of spectroscopy, generate a steadily increasing volume of data. The enhancements in beam intensity and detector performance, the advent of modern area detectors of increased resolution, brought the average data throughput to about 1TB/day or 250-300 TB per year. The actual data management strategy relies on the Internet and portable media to make experimental datasets available to collaborators at home institutions for further detailed analysis.

The discovery potential of photon sciences at BNL will be expanded greatly by the construction of a new state of the art third generation synchrotron light source – NSLS-II.

The new facility is currently in an advanced phase of its construction and will be commissioned and become operational in mid 2014. NSLS-II will provide ultra high brightness and flux –10,000 times brighter than the brightest source at NSLS, and exceptional beam stability. Relying on advanced insertion devices, optics, detectors and robotics, the new facility will enable studies of material properties and functions with a spatial resolution of ~1nm, an energy resolution of ~0.1meV and the ultra high sensitivity necessary to perform spectroscopy of a single atom.

When fully built, NSLS-II will accommodate more than 58 beamlines, which given the facility's significantly enhanced capabilities, will support a substantially larger community; about 3500 users per year. Given the anticipated developments in detector technology and the wide use of automated experimental setups, the expected data volumes will increase by more than two orders of magnitude relative to NSLS. The aggregate data volumes may reach up to 500TB/day or several PBytes per week, making entirely impractical the actual use of portable media for disseminating experimental results. In addition, the complexity of particular analysis algorithms compounded with the increased data volumes will create a strong requirement for a sizeable shared computing facility located close to the data sources. This is the optimal solution for controlling network bandwidth demands and providing an economic response to the need for significant computation cycles. Consequently, secure and reliable remote access to shared computing services, remote instrument control and real-time collaboration support, will gain an unprecedented role in assuring that the NSLS-II facility reaches its scientific potential.

6.2 Key Local Science Drivers

6.2.1 Instruments and Facilities

Data intensive experimental activities at NSLS involve a variety of techniques and instruments and rely on modern high throughput detectors. Typical high data volume programs include Macromolecular/Protein Crystallography (PX), Quick EXAFS, imaging and spectroscopy using multi-element detectors. Among them, Protein Crystallography dominates the data production.

NSLS has ten PX beamlines, including eight bending magnet (BM) beamlines employing 4Kx4K CCD or 2Kx2K CCD, and two undulator beamlines using ADSC Q315 6Kx6K CCD. Depending on the studied sample and detector resolution, the typical BM beamline produces 20-50GB/day, for an aggregate across the eight BM beamlines of 160-400GB/day. The two undulator beamlines with Q315 detectors, at readout rates of ~1sec for a full 36Mpixel frame, produce 250GB/day per beamline or about 500GB/day across the two. The average daily data production of the PX program is thus between 700-900GB/day.

The PX experimental data is stored on local disk arrays dedicated to the PX program, physically located on the NSLS experimental floor and within the BNL's Biology Department. The two locations are connected with a dedicated network link.

Other large contributors to experimental data production that together can add between 200-300GB/day include:

- The quick EXAFS program with continuous scan in photon energy, involving fast ADCs, can generate about 20GB/day.
- Undulator beamline for SAXS using area detectors – 20-50GB/day
- Multi-element Si detector array (384 elements) for microprobing with microbeam scanning applications – 80Gb/day
- Microfocusing beamline using CCD detector – 50-80GB/day.
- Two CCD detectors available on request – 10-30GB/day per detector, depending on experiment type.

Except the PX program, all other NSLS experimental setups use primarily local, dedicated storage and processing facilities.

Overall, NSLS produces on average 1TB/day, or 250-300TB/year.

6.2.2 Process of Science

The typical data acquisition cycle involve an initial sample setup followed by several quick preliminary detector acquisitions to confirm the correctness of the experimental setup; once the setup parameters are properly established, several full duration measurements are taken, varying diverse parameters, as the specific technique requires. The output of these measurements is carefully monitored to confirm the validity of the results; invalid or otherwise valueless datasets are discarded. The cycle restarts with a sample change or other major intervention on the experimental setup.

Typically, the commercial detector systems are provided fully equipped with data acquisition electronics and the associated front-end computer system. The raw data flow generated by the detector is buffered by the front-end computer and the resulting data delivered to the user as files within the local operating system. The user can choose how to handle the data depending on its volume and the level of analysis that can be performed locally. If the data volume permits, many acquisition cycles can be stored and analyzed locally, up to the entire ensemble of an experimental run. Later, the data is flushed to external media to be carried by the scientists to their home institutions for further analysis, or transferred to facility file servers or FTP servers. Within the NSLS community the use of portable media for transferring experimental data between organizations is widely spread.

For larger volumes of data or computationally intensive analysis techniques, the data is transferred to external storage arrays shared by a specific group (e.g. the PX program). The data is processed with the computing resources provided by the group and can be held in store in accordance to the group's storage policies. Ultimately, the raw and the processed data sets are copied on portable media and carried to home institutions or to a lesser extent, are transferred to those locations using the wide area network.

An experimental run, defined by the beamline time allocation, usually lasts between 3-5days. At the end of this period the experimental resources, including data storage and

processing, must be returned to their default state, ready for the next team. Consequently, all local data processing must be completed and all storage resources must be freed up. Given the average daily data volumes enumerated in section 1.2.1, for the vast majority of techniques and setups, carrying the data on portable media remains very practical as it rarely involves more than 1TByte. On the contrary, transferring several hundred GBytes over the network, primarily to academic institutions, had proven difficult more often than not. The main difficulties are insufficient network provisioning at the non-BNL end (especially the last mile) and ineffective or difficult to use transfer utilities (issue largely exacerbated by cybersecurity restrictions).

Each individual NSLS beamline is configured as its own private local network in order to comply with BNL cybersecurity requirements while not imposing impractical constraints onto visitor equipment. Well-defined traffic is allowed through the network boundaries, as is the case for storage access to group resources (PX), or traffic to/from the facility's FTP server. Together with other cybersecurity protections implemented lab wide, these measures impose restrictions and create substantial complexity that works against wide deployment of collaborative utilities and the use of remote instrument control. However, sufficient flexibility was built into the system to allow remote control for certain setups that have requested it.

6.3 Key Remote Science Drivers

6.3.1 Instruments and Facilities

The NSLS experimental data volumes, even for the more data intensive techniques or programs, remain relatively easy to handle by the current technology. Portable disks are easily available at capacities between 1-2TB, very generously proportioned for the actual data sets. Consequently, dispersed scientific collaborations make wide use of data replication using portable media, copying data sets in as many locations as necessary, to comfortably pursue local analysis as required. With this convenient alternative available the pressure on deploying network resources, where insufficient, is weak and was rarely acted upon.

The dominant analysis model is based on individual standalone workstations running local copies of commercial or community developed software, accessing a local copy of the raw and pre-processed data sets. Given the data volumes and the compute power of the typical workstation (PC), the model is practical for a large majority of situations.

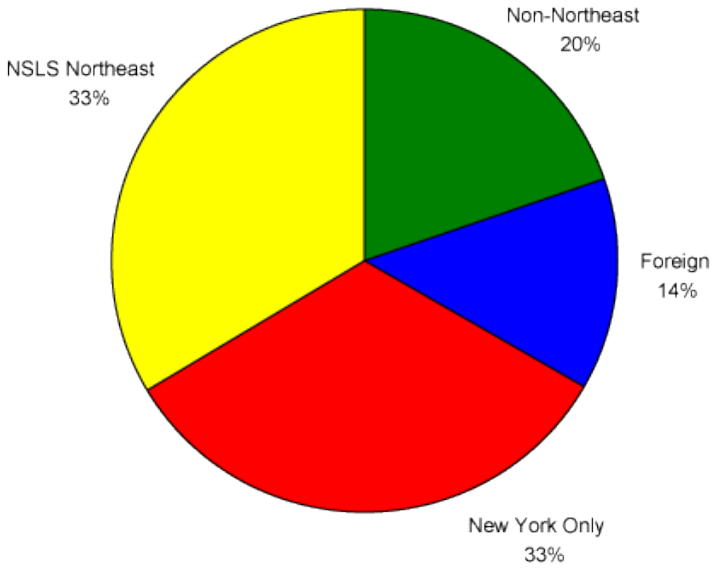
The improvements in detector technology threaten this distributed but disconnected model, by creating data volumes that will become impossible to handle by portable media or standalone workstations; on the other side, computer technology advancements counter this risk.

6.3.2 Process of Science

Two thirds of the NSLS user community is based in the Northeastern United States (Fig.2) making travel to BNL easily acceptable. The use of teleconferencing is not widely spread, an exception being the intra-group ad-hoc point-to-point working meetings mediated frequently by tools like Skype and using beamline local equipment (laptops).

Remote instrument control is used by few beamlines and typically consists of access to video cameras broadcasting details of the experimental setup (via commercially packaged video servers, e.g. AXIS) and remote access to control and data acquisition computers (terminal sessions via ssh or Windows remote desktop).

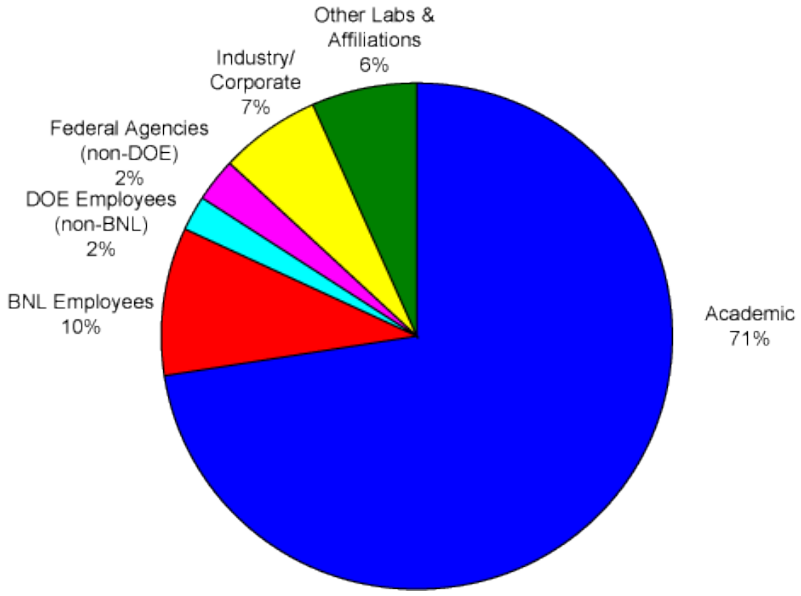
NSLS Geographical User Distribution (2009)



Three quarters of the NSLS scientific community originates at academic institutions (Fig.3). Traditionally, synchrotron research did not require the kind of computing resources, including network resources, demanded by more compute heavy sciences as nuclear and particle physics, or earth sciences. Therefore, frequently, university research groups using synchrotron facilities found themselves lacking sufficient local and wide area network provisioning, or the local expert support necessary to handle large scale transfers, modern secure access, tele-presence or remote instrument control.

Transferring large amounts of data require the support of modern utilities that can make effective use of large network bandwidth – tools newer than the traditional FTP, like bbftp or GridFTP, which can control multiple data streams and fill high-speed links. At their turn, grid utilities require the use of PKI and a certain familiarity with non-Windows environments. The steep learning curve associated with PKI and grid tools was a relatively high entry threshold limiting their wide deployment in the synchrotron research community. In addition, effective large scale transfers sustaining high transfer speeds for prolonged periods of time may require, at least initially, expert support from local network engineers and qualified system administrators to debug the network paths and properly tune high performance storage servers.

NSLS Users by Affiliation (2009)



As a result of:

- Modest experimental data volumes and widely available TByte range portable media
- Acceptably compute-intensive analysis algorithms and powerful CPUs in standard, low-cost workstations
- Ease of regional US travel and instability and complexity of multifunctional tele-presence solutions, compounded with high cost and complexity of increased beamline automation
- Effectiveness of carrying the data on portable media, high expertise necessary to tune high speed network transfers, complexity of tools and lack of proper provisioning of last mile network connections

the dominant mode of operation is that of mobile agents (individuals) physically carrying expertise and data, between service providing synchrotron facilities and local analysis centers. The role of data networks in carrying knowledge and products of scientific discovery is for the moment secondary to more traditional means of transportation.

However, the augmented capabilities and the increased performance of the new generation of facilities will break this paradigm and require the use of data networks to fulfill their productivity potential. NSLS-II will have to use a novel approach based on automation, remote instrument control, telepresence and specialized computing facilities.

6.4 Local Science Drivers – the next 2-5 years

6.4.1 Instruments and Facilities

With brightness 10,000 times superior to NSLS, with unique characteristics such as nanometer size beams and ultra high energy resolution, NSLS-II will completely change the way synchrotron science is done at BNL. The latest future generation of detectors will be used, increasing the data throughput by several orders of magnitude and producing at certain experimental station multiple times the entire NSLS daily data production.

To illustrate the current leading edge technology: one of the high performance hybrid area detectors of today: PILATUS 100k delivers 100kpixel frames at 300Hz for an equivalent raw throughput of more than 600Mbps. Also available commercially, the PILATUS 6M can assemble 60 elementary modules of the 100k version to increase the raw throughput to 1500Mbps. Medipix2's 256x256 photon counting matrix can produce today a throughput of 1Gbps. Fast CCD cameras can nowadays collect up to 6000fps at resolutions of 800x600pixels, producing almost 50Gbps internally.

Within the next few years, it is anticipated that NSLS-II could benefit from area detectors as large as 6Mpixel with readout rates between 1-4kHz or higher. The raw data rates produced by this class of detectors would be between 100-200Gbps. An example of current state of the art, that argues for these data rates, is the development of the EIGER detector – an advanced prototype proven to operate at 20kHz will undergo trials at NSLS within few months (as of June 2010); an advanced production version with resolution up to 4Mpixel and readout rates of 10-20kHz is expected to be available within 4 years. As in current installations, these rates will be buffered by the front-end systems which will deliver the file based datasets for archival and analysis. However, at these volumes, the front-end systems will require substantially higher performance and also high speed, large volume specialized storage systems.

Using this future class of detectors with the ultra bright NSLS-II beams, depending on technique and scientific program, several NSLS-II experimental stations will yield considerable amounts of scientific data. Examples of high throughput experiments define a range of data production between 2TB/day/beamline for coherent diffraction imaging or X-ray fluorescence microscopy, through 15-20TB/day/beamline for protein crystallography or time-resolved X-ray spectroscopy, up to 70TB/day/beamline for high-speed X-ray tomography.

While not all the NSLS-II beamlines will produce vast amounts of experimental data, aggregating the data flows over the ensemble of planned NSLS-II beamlines yields a total daily data volume of about 500TB/day per entire facility, or up 3PB per week. Handling such a large data volume will not be possible with the current NSLS model employing standalone groups processing data in quasi-isolation. Supporting the scientific mission of the NSLS-II community will require facility provided high performance storage systems, a medium/long term archival facility and dedicated analysis resources.

Dedicated 10Gbps network connections are planned to be deployed between each high throughput experimental station and the facility core, which will be linked with a local computing facility for data storage and processing, at speeds between 80-100Gbps.

Depending of the fraction of data assumed to require processing at remote locations: currently between 10-20%, the wide area network connectivity needed for NSLS-II ranges between 6-12Gbps.

6.4.2 Process of Science

The process of science at NSLS-II can be separated into two phases. The first phase occurs during the actual beam time allocation when the scientific group performs experimental measurements and a lighter pass of data analysis. This is followed by the second phase, or post-measurements, which is focused on completing the analysis of the acquired data. The structure of the process is similar to the one at NSLS. However in the NSLS case, due to the lower data volumes, most of the analysis is completed during the beam time allocation using beamline-local resources, The second phase of analysis relies completely on the resources available to the home institution, with practically no computing support from the NSLS facility. From the network perspective, in the NSLS case, phase 1 is local, while phase 2 takes place remotely but it is almost entirely disconnected (the exception being the few cases using the network to transfer data). The NSLS-II model brings in a strongly connected phase 2, where users rely substantially on facility provided resources to complete tasks otherwise far beyond the reach of institutionally available resources. The NSLS-II computing facility will use a tiered model where resource allocation will follow the separation between these two phases of the discovery process, in order to correctly manage system performance and service availability while maintaining an optimal total cost.

For the first phase of the process, the NSLS-II data acquisition cycle structure remains the same as for NSLS but the substantially increased data volumes will require higher performance storage and processing systems. These systems must have the necessary speed and volume to sink the raw detector data and simultaneously provide access for data reduction, on-line monitoring and fast analysis. In the NSLS-II two tier model, these large data flows requiring low access latency will be better supported by beamline-local computing resources dedicated to the local activities and available only to the current experimental group. This layout will also avoid overloading the central data storage systems and the facility local network. The required storage bandwidth of each beamline system is limited to the needs of a single experiment, thus reducing the system's required level of performance and its cost. In addition, data reformatting and removal of invalid datasets reduces the volume of data that needs transferring to central systems, reducing the requirements and cost of the LAN. So far, for Phase 1 activities, both NSLS and NSLS-II models are structurally similar providing the same types of services but at different performance scales.

As with the current NSLS model, the beamline systems must be returned to default state before the experimental run is completed – this includes completing the data processing and flushing the local storage system data to the central archive. To control the LAN requirements and not create a high peak bandwidth demand by concentrating the traffic during the final period of the run, an automated process will schedule background data transfers to central storage for the entire duration of the run, so the traffic profile is flattened.

All the analysis activities that cannot be completed during the experimental run, which are the object of Phase 2 of the process, will be supported by the resources of the NSLS-II computing facility. This is a major change in strategy from NSLS. Exceptions will probably be the lower data volume experiments, which may continue to process data at home institutions. Data will be held on-line (on disk arrays) for certain duration and then archived, in accordance with agreed storage policies. Data analysis will be supported by dedicated application servers running commercial and community-developed software, and by computing farms supplying raw cycles. NSLS-II community users will be able to remotely access the facility's services in order to complement their local resources. Given the substantially larger volume of experimental data, the computing resources provided by the NSLS-II facility will be essential for the proper completion of the discovery process.

Relevant fractions of the large experimental datasets, or complete sets for lower volume techniques, will be available for transfer to outside institutions via high performance networks.

6.5 Remote Science Drivers – the next 2-5 years

6.5.1 Instruments and Facilities

The current analysis model including primarily standalone workstations that process local copies of raw and pre-processed data will lose its dominance and will be used only by few relatively low data volume experiments. An upper boundary for its use may be illustrated by techniques generating up to 10TB per experiment -- coherent diffraction imaging or fluorescence microscopy experiments, which, if we consider 4-5 years of computer technology advancements, might still be practical to handle with inexpensive equipment. A probable change of the model above will involve replicating widely only certain datasets and leaving the rest of the experimental data within the facility's archive to be transferred over the network when needed. Another adjustment of the current model may involve the deployment at home institutions of a small set of mid-range computing systems (group servers), hosting a shared storage system (several tens of TBytes) as well as analysis applications, which can be used collectively by the local group.

The more radical departure from the current model will be required by experiments producing between 100-500TB of data. Handling several datasets of this volume will require resources that are uneconomic to deploy at a large number of institutions. Instead, a shared central facility deployed at NSLS-II makes more economic sense and can provide support for multi-petabyte long term archival of experimental data via the use of volume optimized tape media and also offer shared access to powerful application servers and compute farms for intensive data processing.

The deployment of the NSLS-II computing facility will limit the need for wide data dissemination that would require considerable amounts of WAN bandwidth. Assuming that 10-20% of data requires remote replication, the NSLS-II remote connectivity would be in the 5-15Gbps range. Given the absence of remote processing facilities, the data transfers will occur between NSLS-II and a multitude of mostly academic institutions,

presumably connected to Internet2. None of the individual transfers will substantially exceed 1Gbps.

The increased use of remote access will put emphasis on the need for an effective authentication/authorization solution that would cover securely and effectively multiple IT services. Drawing from the experience of the Open Science Grid (OSG) collaboration, one may use the existing ESnet's DOEGrids PKI and CA services to support the deployment of certificate based access control. In addition, one could also use the OSG middleware to control access to computing resources (via gateways and workload management systems) and support data movement requirements (GridFTP).

6.5.2 Process of Science

Remote access and automation will play an important role at NSLS-II. The increased luminosity of the facility will shorten the sample exposure time and will emphasize the need for fast, automated setups in order to maximize the facility's productivity. Sample handling automation and the streamlining of the measurement process will create a strong incentive for expanding the use of remote instrument control, moving towards a mode of operation that reduces on-site visits, relying instead on mailed in samples, remote control of instrumentation, pipelined data analysis processes and return of scientific results via electronic delivery.

Not all programs and experimental techniques may benefit from this kind of standardization. A substantial number will retain their experimental nature and will require physical presence and collaboration with colleagues from home institutions. In this case the role of tele-presence will be very important. Beyond support for high quality multi-party video conferencing it will be important that the solution provides reliable support for sharing access to computer applications and integration of multiple video sources. The typical use will consist of video conferencing multiple participants from multiple locations plus sharing view or/and access to experimental computer control and multiplexing several video cameras capturing images of the experimental setup.

For effective use of both tele-presence and remote instrument control it will be critical that ESnet (together with major peers as Internet2) provide end-to-end reliable, low latency connections.

Data processing at the NSLS-II computing facility will require secure and reliable system access, possibly based on a global authentication mechanism similar to the Grid PKI. The existing OSG tools and utilities could be appropriate for controlling access to computing resources, for scheduling large processing jobs, for performing usage accounting and transferring data. However the steep learning curve will require a well-organized training effort.

6.6 Beyond 5 years – future needs and scientific direction

Past the 5-year timeframe, NSLS-II will continue to mature its experimental installations and detector setups, creating a steadily increasing data flow. Advancements in detector technology and automation, increased sophistication of experimental techniques will produce larger data volumes and will demand more analysis resources. On the other side improvements in computer technology will somewhat temper this trend.

Increased power of commodity computing may boost the interest in transferring more data to home institutions for detailed analysis and add to the WAN bandwidth demand.

Several beamlines will substantially increase their productivity via sample mail-in and streamlined processing, widening the use of remote instrument control.

6.7 Middleware Tools and Services

ESnet's DOEGrids CA together with the collective experience, tools and utilities developed, tested and deployed by the Open Science Grid will be invaluable for deploying a distributed solution as the one necessary for NSLS-II. Even before that time, NSLS can benefit from utilities like GridFTP, as long as the learning curve associated with PKI can be addressed via expert support provided by the facility to all the users.

ESnet's Collaboration Services are so far relatively unknown to the NSLS community and the freely available services rarely used. A locally driven education effort may change that and make tools like EVO more popular.

6.8 Summary Table

Key Science Drivers			**Anticipated Network Needs**	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• 64 experimental stations at NSLS using high throughput X ray detectors • NSLS beamline local storage and data processing systems • Standalone/group operated workstations for data analysis at home institutions, processing local copies of experimental data	• NSLS local data acquisition and data analysis using local storage systems for the duration of the experiment • Data transfer to home institutions primarily via portable media • Data analysis using group or department shared resources • Limited remote instrument control • Point-to-point teleconferencing	• Total facility production: 1TB/day, or 250TB/year to 300TB/year • Max size of experimental data sets is ~1TB	• Beamline private networks that link storage and processing resources are 1Gbps. • Facility network infrastructure is 1Gbps	• Maximum 1TB per day at 50% utilization factor: 200Mbps • 1Gbps facility WAN connection is sufficient

2-5 years				
• 64 experimental stations at NSLS with detectors that have higher performance compared to today's instruments • 58 experimental stations at NSLS-II using ultra high throughput detectors • Beamline local storage and processing resource plus facility computing center for long term archiving and data intensive analysis • Group operated analysis systems and home institutions	• Beamline local data acquisition and preliminary data analysis on beamline resources (for the duration of the experiment) • Data transfer to facility hosted archive • Complete data processing on facility hosted resources • Remote access to computing facility for data analysis • Data transfer of main analysis results to home institutions • Extensive use of remote instrument control • Multicast teleconferencing with support for application sharing • .	• Total facility production (NSLS-II): ~500TB/day • Max size of experimental data sets: ~300TB • Max size of single experimental daily dataset: ~70TB	• Transfer to facility archive: 70TB/day at 80% utilization factor = ~8Gbps • Requires 10Gbps links between experimental station and facility core • 100Gbps facility backbone	• (Assumes that 10-20% of total data volume must be transferred to home institutions) • 500TB/day times 10-20% at 80% utilization factor = 6Gbps to 12Gbps • Required WAN connectivity: ~20Gbps • Many remote endpoints, no single connection exceeding 1Gbps • Reliable, low latency connections for remote instrument control
5+ years				
• NSLS-II full beamline set with increasing detector performance and wider use of automation	• Enhanced computing facility with expanded capacity • Extensive use of telepresence and remote instrument control	• Total facility production of up to 1PB/day • Max expected daily data volume per experiment: 100TB/day to 200TB/day	• Complete deployment of 10Gbps to each experimental station • Upgrade certain beamlines to 20Gbps or 40Gbps • Upgrade facility backbone to 200Gbps	• WAN connectivity up to 40Gbps

7 Neutron Scattering Science User Facilities at ORNL

7.1 Background

The Spallation Neutron Source (SNS) and the High Flux Isotope Reactor (HFIR) are both DOE experimental user facilities at ORNL. These facilities are used to support open research as experiments are performed in diverse science areas such as structural biology, crystalline materials, polymers, magnetism and superconductivity, chemistry, and disordered materials.

In general, experiments are performed by placing a representative sample of the material of interest within the direct neutron beam of an instrument. The neutrons interact with the sample material, and a portion of these neutrons are scattered and are detected by the instrument detector systems. These scatter patterns can be analyzed to characterize the structure and/or the dynamic processes of a material. The instruments are designed and optimized to statistically measure a particular range of size scales ranging from atomic, to molecular, to macro-molecular structures.

Data acquisition for these instruments can be characterized as collecting either traditional histogram data, or in a newer event data format in which the pixel location and time of flight information are identified for each individual detected event concatenated into live event data streams or as event list files. Upon completion of an experiment measurement, the data are pre-processed and formatted into an HDF-5 file format called NeXus, and a process called "live cataloging" is invoked in which the data are cataloged and copied to a centralized data archive at SNS. A single measurement is called a "run", and an experiment is typically composed of a number of runs.

By the nature of the two facilities SNS and HFIR, SNS has the ability to produce significantly larger data sets. For the lower range of SNS, a run data set may be on the order of tens to hundreds of MBytes, and on the upper end, a run data set may be many GBytes. A new SNS instrument called NOMAD, which is designed to study disordered materials, has the potential to produce 1TByte of data daily. An imaging and tomography beam line is being planned which will have potential to produce similarly large sized data sets, perhaps larger.

The current facility data networking paradigm is largely intra-ORNL at present. Data flows from the instruments to a centralized data management system and are processed either locally on SNS computers or on other ORNL computers. It is also useful to note that some computing for SNS takes place on NSF TeraGrid computers as well. However since the data volume users are producing typically surpasses the computing capacity readily available to these researchers, the centralized data management and computing infrastructure resident at SNS is quite useful for facilitating analysis and research.

New horizons in science give us insights into more network-centric science research environments. For example, crystallographers often perform experiments at X-Ray sources such as APS or NSLS prior to coming to SNS where they will perform neutron scattering measurements to co-refine the structures using their data obtained from both techniques. Also intriguing is that SNS has the ability to stream live data during an

experiment, and depending upon the network capabilities, this data could be streamed to local or remote high performance computing resources to perform near-real-time data analyses which the researchers could use in-situ to help them cultivate scientific inferences during an experiment. Near-real-time computing during an experiment can be used for determining structure, orienting the sample within the beam, or for controlling or sequencing more complex experiments – especially in the complex cases for creating the sample *within* the neutron beam while collecting data.

Thus one can imagine a number of scenarios for inter-facility networking among DOE's experimental and computational user facilities, though establishing the new science cases to support these connectivity scenarios is still ongoing. To pursue these science cases in today's DOE cyber-environment can be a daunting challenge leaving one to face many obstacles. However we live in an ever-increasing network-centric world, which thrives on the creativity offered by the flexibility it presents – which leaves us optimistic of quantum scientific gains, which could be made by making these inter-lab networking investments.

7.2 Key Local Science Drivers

7.2.1 Instruments and Facilities

SNS has approximately 70+ analysis computers and servers within its facilities. A typical analysis computer has between 16 to 32 cores and 64 to 256 GB of RAM. The centralized storage server is an HP SAN with approximately 100TB+ of data storage. NFS is heavily utilized to support mounting disks, though the UDP protocol is utilized for data streaming. The instrument analysis computers are reserved primarily for those performing experiments on-site. A portal infrastructure provides remote users access to both data and computing resources. This portal is based in the Open Research Network (ORN), which is a Low-Low-Low – Confidentiality-Availability-Integrity network, and serves approximately 50+ users per day.

In general, data flows from the instruments to the centralized data management system. Once there, these data can be accessed homogeneously via the other computers on the Open Research Network via a common NFS mount structure established across the computers. This method also facilitates basic queuing as we have implemented SLURM queues running computing jobs, which are largely data reduction jobs at present.

7.2.2 Process of Science

As data are the primary products of user facilities, data movement and data processing are core needs for scientists and researchers once they perform their experiments and then need to analyze their data. For SNS, the process utilizes the above described resources and can be characterized as listed below.

- In some cases, facility users may undertake pre-experiment planning facilitated by performing computer simulations in order to better anticipate experimental results. As simulation tools continue to improve, performing these simulations may likely become a routine component to performing experiments.

- Measuring and collecting data is a primary activity of users while at experimental user facilities – the measurement techniques vary widely across the suite of instruments at the facility, but they have some common steps involving data as listed below:
    - Acquiring data
    - Data movement, cataloging, and archiving
    - Data treatment and data reduction
    - Visualization of the raw and reduced data
    - Batch processing of data when possible
    - Fitting to models and analysis of the processed data for structure or dynamics information

7.3 Key Remote Science Drivers

7.3.1 Instruments and Facilities

As SNS and HFIR facility users come from across the globe, the need for collaboration tools is growing. This ranges from tools for locating and accessing data, running computing jobs remotely, to electronic collaboration (email, chat, Skype, remote desktop, etc.), to collaboratively working on publications together. Having the ability for researchers to leverage a feature-rich integrated resource pool would be helpful.

7.3.2 Process of Science

Scientists will begin to look more at the portfolio of networking and computing resources available to them to help support performing their research. Researchers may work at a number of experimental facilities nationally and perhaps internationally, and will want to have autonomy in being able to move and locate these data, utilize computing resources, define collaborations, and publish their results.

7.4 Local Science Drivers – the next 2-5 years

7.4.1 Instruments and Facilities

Both the SNS and HFIR facilities will continue to add instruments within the next 2 to 5 years, however the number of instruments being added is gradually reducing with time. The commissioning of existing instruments and these new instruments will continue at some level during this timeframe.

A focused effort is underway to upgrade the Open Research Network at SNS to 10Gbit. However ORNL has not yet been able to follow suit and provide 10Gbit access from the Open Research Network to the ORNL perimeter firewall. A next step goal to target late within this time period would be to provide multiple 10Gbit links, and perhaps higher per connection bandwidth such as 40 or 100Gbps.

7.5 Remote Science Drivers – the next 2-5 years

7.5.1 Instruments and Facilities

Researchers may start to utilize what could be characterized as "Virtual Instruments", as computing, modeling, and simulation continues to grow within the neutron scattering community. These virtual instruments are intended to enable virtual experiments which would enable researchers to focus more specifically on particular aspects of their research that they could use to gain insights for performing actual experiments. These virtual resources would need to be accessible remotely to researchers and collaborators. The simulations that these virtual experiments facilitate could be quite large, thus bringing the same data movement and storage issues to the neutron scattering community as already faced by the supercomputing communities.

7.5.2 Process of Science

In the next 2-5 years, users will expect there to be something perhaps which could be characterized as "Google Science" available to help facilitate their research. This will probably be a cloud computing based environment in which users can utilize various social networking tools adapted to science as well as high-end tools for facilitating publications. For this mode of working to be successful, it will be important to have access to responsive network-enabled research tools with low latency and high bandwidth connectivity.

7.6 Beyond 5 years – future needs and scientific direction

At the beyond 5-year mark, facility users will likely benefit from laptops with significantly more capacity and performance than the original instrument analysis computers that by this point would be 10+ years old. This being the case, users will come to rely more upon their own computing resources to perform tasks they once relied on remote computing resources to accomplish, and will require high network bandwidth capacity to move data to these computing resources they possess. This is not to say that facility provided computing would be obsolete, far from it. Performing experiments will always require local computing capacity, and research will continue to grow into the computing capacity at hand. It is also anticipated that the current generation of computing savvy researchers will take advantage of HPC resources to better prepare for experiments by performing simulations prior to arrival for their experiments as well as to utilize HPC resources while on-site to help them steer their experiments and for drawing scientific inferences from their experiment data. Following the completion of their experiment measurements, these computing savvy researchers could then take advantage of new high-powered modeling and simulation tools to analyze their data.

In some regards, one could think of a neutron scattering instrument as capable of creating complex data composed of significantly more atoms and molecules than currently computed by today's simulations. Thus the future holds the promise of leveraging extreme scale computing coupled with experimentation as these coverage scales converge.

Users will come to expect high performance, high-reliability access to both data and computing resources. They will want to work in a virtual world held together by networking and intelligent data management tools, which blurs the need for one to keep track of where resources and data reside. Cloud computing holds promise here, however these new capabilities have not yet been explored for use in supporting the research of SNS and HFIR users.

7.7 Middleware Tools and Services

- GridFTP is utilized for data movement between SNS computing resources as NSF TeraGrid computing resources.
- A number of workflow tools for helping with data movement and job management have been examined, though discovering the best tools for the job can be a challenge for a user facility to determine by itself as there are a variety of differing user needs and analyses by instrument beam line.
- Cloud computing is not yet in mainstream planning for HFIR and SNS, however these could hold promise for facilitating research if sufficient bandwidth between these resources and the data repository can be achieved.
- There is an increasing interest in the leveraging of collaboration tools such as skype, remote desktop, and remote application sharing.
- The high-volume data producing instruments give reason to rethink centralized data storage to also include a more distributed data storage model with more storage and computing capacity resident at the instruments.

7.8 Outstanding Issues

The neutron scattering science community is an international community with facilities broadly located across the globe. This being the case, there is interest in networking not only between DOE user facilities, but also with international facilities. This desire brings with it a number of challenging technical and policy issues; most notable is perhaps user identification and authentication. There are currently no methods for (easily) integrating user authentication systems across these facilities – either within the US or internationally. A challenge such as this can be a difficult obstacle for individual user facilities to address, however successfully overcoming these challenges will empower users to work in an autonomous way leveraging a massive portfolio of resources.

A good place to start could be the establishment of a User Facility network (UFnet) layered on top of ESnet specifically targeted and designed to support inter-facility data movement. UFnet could leverage the Data Transfer Nodes currently under development by ESnet to facilitate points of presence at each facility. A software stack could then be developed specifically intended to support the data movement needs of experimental user facility researchers.

7.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• SNS will have ~10 to 17 instruments in the user program	• Data reduction • Visualization • Experiments performed on-site by researchers at the facility	• 10's of GB/day • Could reach 1TB/day • Collecting 100 to 1000 files per day	• 1Gbps networking utilized • Initiate 10Gbps networking	• 1Gbps via ORNL border firewall available to portal users
2-5 years				
• ~16 to 20 instruments in the user program	• .	• 1 to 2 TB/day • 1000 to 10,000 files per day	• 10Gbps networking standard • Multiple 10Gbps lines supporting high-volume instruments individually	• 10Gbps available to portal users • May see data transfers occurring more regularly between SNS, APS, and NSLS • Initiate international collaborations at 1Gbps
5+ years				
• 20 to 24 instruments in the user program • Could begin building the second target station in this timeframe, which would eventually double the number of instruments at SNS.	•	• 3 to 5 TB/day • 10,000+ files per day	• Upgrade internal networking to 40 Gbps or higher	• Utilize 40 Gbps or higher between DOE user facilities • Utilize 10Gbps between international facilities

8 Center for Nanophase Materials Sciences (CNMS) and Computer Science and Mathematics Division (CSMD) at ORNL

8.1 Background (CNMS)

The Center for Nanophase Materials Sciences (CNMS), one of the five Nanoscale Science Research Centers (NSRC) supported by the Department of Energy (DOE) Office of Basic Energy Sciences (BES) Division of Scientific User Facilities (DSUF), began full-scale operations for users and science at the start of FY 2006. From the outset, the vision for CNMS has been to create a world-leading scientific research and user environment for nanoscale science that will accelerate the pace of scientific discovery and drive advances in nanotechnology. The collocation of the CNMS with the Spallation Neutron Source (SNS), and the availability of DSUF sponsored user facilities including the High Flux Isotope Reactor (HFIR) (http://neutrons.ornl.gov/) and the Shared Research Equipment (SHaRE) User Facility (http://www.ms.ornl.gov/share/index.shtml) provide opportunities to incorporate forefront neutron scattering and microscopy/characterization techniques into nanoscale research. Similarly, the location at Oak Ridge National Laboratory (ORNL) of the National Center for Computational Sciences (NCCS) has directly benefitted CNMS research in computational nanoscience, providing opportunities to expand the range and complexity of tractable problems in nanoscale theory, modeling and simulation. The CNMS is driven by the needs of users in nanoscience research. The internal research program is focused on areas where CNMS is at the forefront of nanomaterials sciences, attracting users worldwide to take advantage of new capabilities and expertise. These research areas were originally designed around needs identified in a series of planning workshops with the user community and are designed to evolve with user needs. There are currently three mutually supporting internal science themes:

- Origins of Functionality at the Nanoscale
- Functional Polymer Architectures
- Understanding Emergent Behavior

These CNMS themes are organized to define research problems and bring together appropriate staff expertise from across the organization to address those issues, under the leadership of senior scientific staff. Themes include expertise in CNMS specialty areas including synthesis of model systems, functional characterization over appropriate length and time scales, and theoretical and modeling approaches, and while staff members are primarily focused on a single science theme, collaborations across themes are common. Additionally. The CNMS research staff are engaged in the user program through user training, hands-on scientific interactions, and outreach to prospective users.

As a DOE user facility, the increasing visibility of energy-inspired research is reflected in current science within the themes. Origins of Functionality at the Nanoscale includes scanning probe based research into phase change phenomena at the level of individual defects, extending to multiscale investigation of complex nanomaterial interfaces relevant to advanced photoelectricsystems. Functional Polymer Architectures research addresses

fundamental questions of self-assembly at the nanoscale and resulting function in soft materials including complex conjugated polymers and synthetic polypeptides, which are relevant to energy conversion and optoelectronics, and to novel bioinspired materials. Research in Understanding Emergent Behavior has led to new knowledge related to the nanoscale fluctuations driving pairing behavior of electrons in high-temperature superconductors, as well as the relationships between fluctuations and deterministic processes in natural nanoscale systems. In addition to providing a core capability for the Understanding Emergent Behavior theme, theory, modeling and simulation are important components in all themes, with computational nanoscientists resident in each of the themes. Finally, research in the Nanofabrication Research Laboratory is closely coupled to all three themes and holds a central role in both pioneering capabilities available for CNMS users and for advancing the internal science program. CNMS staff members also are a significant percentage of ORNL-based users of the neutron scattering facilities, integrating neutron sciences into forefront nanomaterials research.

Nanomaterials Theory Institute (NTI) at the CNMS: A major focus of the NTI is on the understanding of emergent behavior at the nanoscale. The overarching research goal of this theme is to understand, predict, design, and exploit complex behavior that emerges at the nanoscale. This is a long-term goal with significant near-term opportunities over the next three years. The theme focuses on phenomena that are driven by a combination of the following two unique characteristics: the crossover from weak to strong electronic correlations, and the dominance of fluctuations and non-equilibrium states over static properties. In addition, significant effort is focused on understanding and predicting closely related phenomena that arise from integration across length and time scales, and their impact on functionality. This goal will be realized by research addressing the following three questions: How does the crossover from strong to weak electronic correlations influence a nanostructured material‘s behavior, where does it occur, and how can it be controlled? How do atomic scale structure and quantum mechanical effects impact electron transport and other electronic processes within nanostructures and across interfaces? What are the rules for compositional and temporal fluctuations in nanoscale systems and how does function emerge from fluctuations, particularly when they are large? The concept of fluctuations as defining function at the nanoscale is a common element running through this theme.

This research complements the current user-driven science at the CNMS. The interdisciplinary research is intended to add depth to the science at the CNMS and will lead to the development of new capabilities that will be available for future user projects.

Scanning Probes and Multiscale Functionality groups at the CNMS: The objectives of these groups and the underlying theme research are to develop new tools, both experimental and computational, capable of identifying functionality on the nanoscale; to use these tools to gain insight into a range of critical questions in nanoscale phenomena and interactive functionality; and to disseminate both the experimental results and the novel tools to the user program and the broader nanoscience community. To understand the nanoscale origins of multiscale functionality, three specific aims of research addressing this overarching goal are:

- First, to understand how the atomic and molecular structure of **nanomaterial interfaces** govern fundamental phenomena central to functionality, and how to control these interfaces during synthesis and processing.
- Second, to understand how architectures of nanostructures and their interfaces influence collective energy transfer phenomena to govern multiscale functionality.
- Third, to develop the new techniques that bridge characteristic length and time scales required to enable aims one and two and reveal unexpected phenomena.

Background (CSMD)

The Computer Science and Mathematics Division (CSM) is ORNL's premier source of basic and applied research in high-performance computing, applied mathematics, and intelligent systems. Basic and applied research programs are focused on computational sciences, intelligent systems, and information technologies.

Our mission includes working on important national priorities with advanced computing systems, working cooperatively with U.S. Industry to enable efficient, cost-competitive design, and working with universities to enhance science education and scientific awareness. Our researchers are finding new ways to solve problems beyond the reach of most computers and are putting powerful software tools into the hands of students, teachers, government researchers, and industrial scientists.

Computational Chemical and Materials Sciences: The Computational Materials Science (CMS) group develops and applies modern computational and mathematical capabilities for the understanding, prediction and control of chemical and physical processes ranging from the molecular to the nanoscale, to full-size engineering applications, using a multidisplinary approach that integrates chemistry, physics, materials science, mechanical engineering, and biology. Additionally, the CMS group is the core of the Nanomaterials Theory Institute at the Center for Nanophase Materials Sciences, where work is focused towards using theory and multiscale simulations and modeling for providing interpretive and predictive frameworks for virtual design and understanding of novel nanoscale materials with specific and/or emergent properties.

Current research areas include: Ab-initio materials simulation, applied mathematics, bio-nano science, computational biology and biophysics, correlated electron materials, energy storage materials, engineering and transportation technology, magnetisim and magnetrotransport in nanostructures, mechanics of materials, mesoscale models of deformation and dislocation, nanoscale charge transport, soft materials (polymers), and superconductivity.

Computational Astrophysics: The primary scientific and computational focus is on tera- to exa-scale simulation of supernovae of both classes in the Universe.

Core collapse supernovae are the death throes of massive stars, more than 8-10 times the mass of our sun. They are a dominant source of elements in the Universe, without which life would not be possible. The group is focused on ascertaining the core collapse supernova mechanism - i.e., how the explosions of these stars are initiated. Core collapse supernovae are three-dimensional, multi-physics events. Three-dimensional general relativistic radiation magnetohydrodynamics simulations must be performed to ascertain

definitively the supernova mechanism and to predict all of the associated supernova observables. Core collapse supernovae are driven by neutrinos (radiation) and perhaps magnetic fields. Thus, the group has developed discretizations, solution algorithms, and codes for the solution of the multi-dimensional neutrino (radiation) transport equations and the three-dimensional magnetohydrodynamics and Poisson equations, the latter for the star's self gravity. This is ultimately a petascale to exascale computational problem.

Other work involves research on Type Ia supernovae - in particular, the mechanism whereby white dwarf stars (the endpoint of stellar evolution for stars less than 8 times the mass of the Sun) end their lives in stellar explosions as well. Understanding these stellar explosions is particularly important in light of the fact they provide the means to probe the evolution of the Universe as a whole. Indeed, observations of Type Ia supernovae and conclusions based on them have led to the startling fact our universe is expanding at an accelerated rate, which has significant implications for its future and ultimate fate. Type Ia supernovae are driven by thermonuclear runaway. The challenge here is the modeling of a turbulent flame, centimeters thick, in a white dwarf star the size of the Earth. Thus, three-dimensional simulations of chemically reactive flows are required, with realistic sub-grid models.

Computational Earth Sciences: The Computational Earth Sciences group conducts research in computational methods for the prediction of global and regional climate. An emphasis is placed on the atmospheric dynamics and hydrology of earth systems.

Current Group Projects include:

- DOE Biological and Environmental Research Climate Change Prediction Program and the DOE Office of Science SciDAC Program
- A Scalable and Extensible Earth System Model
- Earth System Grid project
- DOE Office of Science SciDAC Program
- Multiscale Subsurface Reactive Flows
- Performance Engineering for the Next Generation Community Climate System Model
- NASA Carbon Assimilation for the Orbiting Carbon Observatory
- USDA Ecosystem Modeling
- ORNL LDRD - Economic modeling and climate feedbacks
- LDRD - Assessing Decadal Prediction of the Earth System after major volcanic eruptions

Computational Engineering and Energy Sciences: The overarching mission is to develop and implement high performance numerical algorithms to solve key problems in the areas of computational energy and engineering science.

Group members are involved in the development of codes, computational physics, and numerical mathematics appropriate for science-based analysis of key energy and national security problems of national interest.

Future Technologies: The Future Technologies group performs basic research in core technologies for future generations of high-end computing architectures, including experimental computing systems. Using measurement, modeling, and simulation, we

investigate these technologies with the goal of improving the performance, efficiency, reliability, and usability of these architectures for our sponsors. Accordingly, we develop new algorithms and software systems to effectively exploit the specific benefits of each technology.

Statistics and Data Sciences: Focus is on the development of cutting-edge statistical and information technologies and to bring quantitative rigor and efficiency to scientific investigations. We conduct research in the analysis and exploration of data, the collection and organization of data, and decisions based on data. Our collaborative work concerns all stages of the scientific life cycle and utilizes computing platforms ranging from the desktop to large clusters and supercomputers. We team on projects ranging from small single discipline efforts to large multi-disciplinary and multi-institution partnerships.

Areas of application have included: Chemistry, Biology and Genomics, Astrophysics, Climate, Environmental Science, National Security, Forensics, Simulation Science, Epidemiology, Fusion Science, Transportation and Automotive, Health and Safety, Grid Technologies, Manufacturing, Future Combat Systems, Remote Sensing, and Computer Network Security.

Computational Mathematics: This group is devoted to the development, analysis and application of efficient numerical algorithms for solving large-scale scientific and engineering problems on advanced computer architectures. Principal Research areas include:

- Boundary element method
- Dense matrix computations
- Direct methods for sparse matrix computations
- Iterative methods for linear systems
- Algorithms for solving differential equations
- Large eigenvalue computations
- Computational geometry and mesh generation

Complex Systems: The Complex Systems group conducts basic and applied research ranging from cooperating robots to quantum communications, including design and analysis of man-machine interfaces.

Center for Molecular Biophysics: The UT/ORNL Center for Molecular Biophysics performs research at the interface of biological, environmental, physical, computational, and neutron sciences. The goal is to study and understand the function of biologically relevant molecular systems by employing high-performance computer simulations in combination with biophysical experiments. The different groups focus on various aspects of this research and are headed by UT professors and ORNL scientists.

Extreme Scale Systems Center: The Extreme Scale Systems Center's (ESSC) primary goal is to help enable the best and most productive use possible of emerging peta-/exa-scale high-performance computers. Of particular interest are the systems expected from the DARPA High Productivity Computing Systems (HPCS) program. The ESSC is intended to foster long-term collaborative relationships and interactions between DOD, DOE, DARPA, NRL and ORNL technical staff that will lead to improved and potentially revolutionary approaches to reducing time to solution of extreme-scale computing and

computational science problems. The ESSC will support the major thrust areas required to accomplish this goal.

8.2 Key Local Science Drivers

8.2.1 Instruments and Facilities

Oak Ridge National Laboratory and the UT-ORNL Joint Institute for Computational Sciences

Computer Facilities: ORNL operates two petascale computing facilities: The Oak Ridge Leadership Computing Facility (OLCF) manages the computing program at ORNL for the Department of Energy while the National Institute for Computational Sciences (NICS) runs the computing facility for the National Science Foundation. Each has a professional, experienced operational and engineering staff comprised of groups in HPC operations, technology integration, user services, scientific computing, and application performance tools. The ORNL computer facility staff provides continuous operation of the center and immediate problem resolution. On evenings and weekends, operators provide first-line problem resolution for users with additional user support and system administrators on-call for more difficult problems. Primary systems include the following:

Jaguar is a Cray XT5 system consisting of 37,376 AMD six-core Opteron processors providing a peak performance of over 2.3 PF and 300 TB of memory. 192 service I/O (SIO) nodes provide access to our 10PB Spider parallel file system at over 240GB/sec. External login nodes (decoupled from the XT5 system) provide a powerful compilation and interactive environment utilizing dual-socket quad core AMD Opteron processors and 64 GB of memory. Jaguar is the world's most powerful computer system and is available to the international science community through the DOE INCITE program.

Kraken currently has a total of 16,512 XT5 compute processors, 129 TB of memory, and over 3 PB of disk space. Kraken and a 170 TF Cray XT4 named Athena are the largest university based computing systems in the world and the largest resources on the NSF TeraGrid.

The ORNL Institutional Cluster (OIC) has come together in two phases. The Original OIC consists of a bladed architecture from Ciara Technologies called VXRACK. Each VXRACK contains two login nodes, three storage nodes and 80 compute nodes. Each compute node has dual Intel 3.4 GHz Xeon EM64T processors, 4 GB of memory and dual gigabit Ethernet Interconnects. Each VXRACK and its associated login and storage nodes are called a block. There are a total of nine blocks of this type. Phase 2 blocks were acquired and brought online in 2008. They are SGI Altix machines. There are two types of blocks in this family:

• Thin Nodes (3 blocks): Each Altix contains one login node, one storage node and 28 compute nodes within 14 chassis. Each node has eight cores. There are 16 GB of memory per node. The login and storage nodes are XE240 boxes from SGI. The compute nodes are XE310 boxes from SGI.

• Fat Nodes (2 blocks): Each Altix contains one login node, one storage node and 20 compute nodes within 20 separate chassis. Each node has eight cores and 16 GB of memory. These XE240 nodes from SGI contain larger node-local scratch space and a much higher I/O to this scratch space because the space is a volume from four disks.

Eugene is an IBM BG/P. It consists of two racks of compute and I/O nodes and three racks of accessory systems. Each compute rack contains 32 I/O nodes, each of which is connected to 32 quad-core compute nodes, for a total of 1024 compute nodes per rack, or 2048 nodes (8192 core) in total. Individual compute nodes have 2 GB of memory, a single quad-core 850 MHZ PowerPC processor and are capable of running at 13.6 GFLOPS. Eugene is available to ORNL researchers to test scalability of applications to thousands of processors.

Frost (SGI Altix ICE 8200) consists of three racks totaling 128 compute nodes, five service nodes (one batch node and four login nodes), two rack leader nodes, and one administration node. Each compute node has two Intel Quad-Core Xeon X5560 at 2.8 GHz (Nehalem) processors, 24 GB of memory, one gigabit Ethernet connection, and two 4x DDR Infiniband connections. Each rack of compute nodes contains 8 Infiniband switches (Mellanox InfiniScale III MT47396, 24 10-Gbps Infiniband 4X ports) that are used as the primary interconnect between compute nodes and for connection to the Lustre file system. The center-wide Lustre file system is the main storage available to the compute nodes. The Frost cluster is available for ORNL staff and collaborators.

Network Connectivity. The ORNL campus is connected to every major research network at rates of 10 gigabits per second or greater. Connectivity to these networks is provided via optical networking equipment owned and operated by UT-Battelle that runs over leased fiber-optic cable. This equipment has the capability of simultaneously carrying either 192 10-gigabit per second circuits or 96 40-gigabit per second circuits and connects the LCF to major networking hubs in Atlanta and Chicago. Currently, 16 of the 10-gigabit circuits are committed to various purposes, allowing for virtually unlimited expansion of the networking capability. Currently, the connections into ORNL include ESnet, TeraGrid, Internet2, and Cheetah at 10 gigabits per second as well as Science Data Net at 20 gigabits per second and National Lambda Rail at 40 gigabits per second. ORNL operates the Cheetah research network for NSF. To meet the increasingly demanding needs of data transfers between major facilities ORNL is participating in the Advanced Networking Initiative (ANI) that will provide a native 100 gigabit optical network in a loop which includes ORNL, Argonne National Laboratory and other facilities in the northeast.

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

Visualization and Collaboration. ORNL has state-of-the-art visualization facilities that can be used on site or accessed remotely. ORNL's Exploratory Visualization Environment for REsearch in Science and Technology (EVEREST) is a 30-ft wide by 8-

ft high Powerwall for data exploration and analysis. The facility has a 600 ft^2 projection area and a 1000 ft^2 viewing area known as the EVEREST lab, a venue that serves both as a visualization center and a place for scientists to meet, hold discussions, and present their work. The ORNL visualization team has developed a suite of middleware software tools that offers an intuitive interface with which to operate the Powerwall and manage multimedia content. Twenty-seven projections are seamlessly edge-matched for an aggregate resolution of 11,520 by 3,072 pixels. This projection environment is driven by an 18-node cluster named Everest. Each node in the Everest cluster contains four dual-core AMD Opteron processors, 4GB of memory, dual NVIDIA GeForce 8800GTX graphics cards, and an Infiniband network. A dedicated Lustre file system provides high bandwidth data delivery to the EVEREST Powerwall. ORNL also provides Lens, a 32 "fat node" cluster dedicated to data analysis and visualization. Each node in Lens contains four quad-core AMD Opterons; 64 GB of memory; two graphics cards, an NVIDIA 8800 GTX and a 4GB NVIDIA Tesla C1060; and an Infiniband network. The Lens cluster is a resource of the OLCF and performs a variety of visualization-related functions, including computation, analysis, and rendering, including support for remote visualization for off-site customers. The Lens cluster has been demonstrated with a variety of commercial off-the-shelf software and open-source visualization tools including VisIt, Paraview, CEI Ensight, and AVS-Express. The Everest cluster rendering environment utilizes Chromium and Distributed Multi-Head X (DMX) for tiled, parallel rendering. The Lens cluster cross mounts the Center-wide Lustre file system to allow "zero copy" access to simulation data from other OLCF computational resources.

High Performance and Archival Storage. To meet the needs of ORNL's diverse computational platforms a shared parallel file system capable of meeting the performance and scalability requirements of these platforms has been successfully deployed. This shared file system based on Lustre, DDN and InfiniBand technologies is known as Spider and provides centralized access to Petascale datasets from all major computational platforms. Delivering over 240GB/sec of aggregate performance, scalability to over 26,000 file system clients, and storage capacity of over 10 Petabytes, Spider is the world's largest scale Lustre file system. Spider consists of 48 Data Direct Networks (DDN) 9900 storage arrays managing 13,440 1TB SATA drives, 192 Dell dual-socket quad-core I/O Servers providing over 14 Teraflops in performance and over 3 Terabytes of system memory. Metadata is stored on 2 LSI Engino 7900s (XBB2) [12] and is served by 3 Dell quad-socket quad-core systems. ORNL systems are interconnected to Spider via and InfiniBand system area network that consists of four 288-port Cisco 7024D IB switches and over 3 miles of optical cables. Archival data is stored on the center's High-Performance Storage System (HPSS) developed and operated by ORNL. HPSS is capable of archiving hundreds of petabytes of data and can be accessed by all major leadership computing platforms. Incoming data is written to disk and later migrated to tape for long term archival. This hierarchical infrastructure provides high-performance data transfers while leveraging cost effective tape technologies. Robotic tape libraries provide tape storage. The center has 3 SL8500 tape libraries, holding up to 10,000 cartridges each and is in the process of deploying a 4th SL8500 this year. The libraries house a total of twenty-four T10K-A tape drives (500 gigabyte cartridges, uncompressed) and thirty-two T-10K-B tape drives (1 terabyte cartridges, uncompressed). Each drive has a bandwidth

of 120MB/sec. ORNL's HPSS disk storage is provided by DDN storage arrays with nearly a Petabyte of capacity and over 12GB/sec of bandwidth.

Joint Institute for Computational Sciences (JICS): The University of Tennessee (UT) and Oak Ridge National Laboratory (ORNL) established the Joint Institute for Computational Sciences (JICS) in 1991 to encourage and facilitate the use of high-performance computing in the State of Tennessee. When UT joined Battelle in April 2000 to manage ORNL, the vision for JICS expanded to become a world-class center for research, education, and training in computational science and engineering. JICS advances scientific discovery and state-of-the-art engineering by:

- Taking full advantage of the petascale and beyond computers housed at ORNL and in the in ORNL's Leadership Computing Facility (OLCF) and
- Enhancing knowledge of computational modeling and simulation through educating a new generation of scientists and engineers well-versed in the application of computational modeling and simulation to solving the world's most challenging scientific and engineering problems.

Nanomaterials Theory Institute (NTI) Computational Facilities

NTI Beowulf cluster (~12 TFlops)

- NTI Developmental Cluster (120 CPU cores and 8 Tesla/ FERMI cards)
- National Energy Research Supercomputing Center (NERSC) allocation –high-end capacity computing
- National Leadership Computing Facility (NLCF) allocation –capability computing
- 16-screen video wall and 16-quad-processor-node visualization cluster/data storage

See figure below for overview of the interactions between different facilities and the NTI

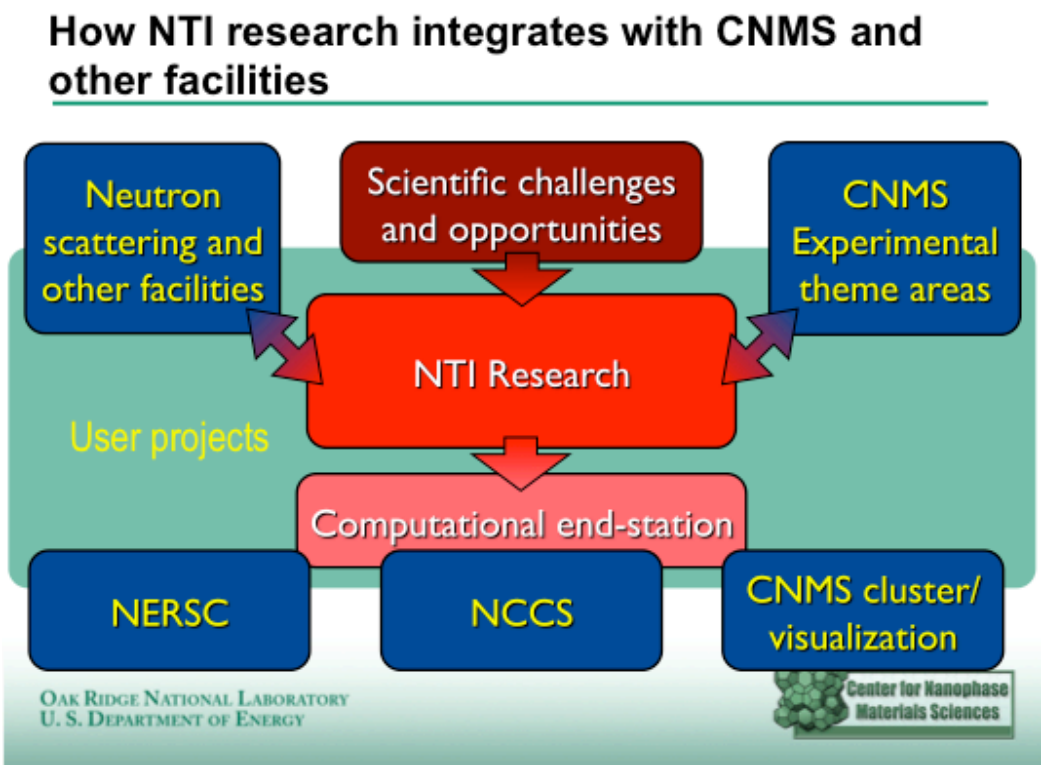


Figure showing how the NTI integrates with CNMS, NERSC, NCCS and ORNL neutron facilities

8.3 Local and Remote Science Drivers – the next 2-5 years

Over the next few years, the size of simulation data sets will increase toward the terabytes. This will be a result of both the deployment of enhanced computational resources and improved scalable algorithms to utilize those resources, which will permit the simulation of considerably larger nanostructured systems and phenomena on longer time scales. In addition, there will be a commensurate enhanced demand for remote access, visualization and control, and analysis of simulation and experimental data that will require greater network abilities. At the same time we should expect the large neutron facilities to be heavily used and likewise demand a greater need for networking and data storage.

8.4 Beyond 5 years – future needs and scientific direction

During the next 8-10 years, it is anticipated that large-scale computing systems will continue to dramatically increase in capacity and capability. Simulation at exascale levels presents both a tremendous opportunity and challenge for computational science to accelerate the development and application of this technology. However, the complexities associated with the anticipated advanced architectures require developing scalable, fault-tolerant, and robust application software tools. These tools must be created using new programming paradigms capable of addressing emerging novel heterogeneous computer architectures along with the various types of memory hierarchies, bandwidths and latencies.

Computational science at extreme scales (e.g., exaflops) necessitates a paradigm shift in the design and development of computer science and applied mathematics algorithms and tools. These tools must integrate into application software to enable extremely large concurrency; address fault tolerance and resiliency; high-degree of heterogeneity in memory access, operations, and I/O; large costs of moving data, load balancing, etc. To harness exascale compute power, the key tools challenge may lie in providing runtime support for scalable pre-processing of performance data vertically through the software and hardware stack.

8.5 Middleware Tools and Services

GridFTP can be effective for much of the needed data movement between computing resources as the CNMS, NERSC, and NCCS. Online data analysis and reduction are also important. However, as computing speed increases towards exaflops, reliable and resilient data generation, delivery, and storage will become an issue. There is also immediate need to consider utilizing collaboration tools such as skype, remote desktop or some form of screen sharing, and remote application sharing, as many projects involve large teams that are not co-located.

8.6 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• Electronic structure calculations, quantum many-body simulations, molecular dynamics (atomistic and mesoscale), mean field approaches, statistical simulations on large computer clusters and capability computing systems • Quantum transport in nanostructures and virtual STM imaging • Self-assembly of nanostructured materials • Statistical physics in non-equilibrium systems • Emergent behavior in strongly correlated electronic systems	• Users and staff generate output from calculations or experiments. Results based on electronic structure, Monte Carlo, mean field theory, and dynamics calculations can be a significant fraction of the memory available on the computer platform. • Users will generally desire to transfer data files back to their home institutions for future reference, analysis, etc. • Memory ranges from several hundreds of GB on clusters to tens of TB on capability machines.	• Simulations can generate several hundred GB of data • Restart files, production data/results, (10-100) depending on the type of calculation • **Case 1**: CCSDT/aug-cc-pvtz for a heteroatom supramolecular complex of 60-100 atoms (100-500 GB); • **Case 2**: molecular dynamics for a million atom carbon system with a trajectory dump for the (**p**,**q**) every 10 fs over a time of 100 ps (100GB-1TB). • **Case 3**: plane wave periodic electronic structure calculations for heteroatom supercells of 100-300 atoms (10-50 GB) • **Case 4**: DCA++ calculations (5-10 GB) • **Case 5**: QMC for 100 + atom systems (10-100 GB) • **Case 6**: DMRG calculations on GMR (1-5 GB) • **Case 7**: experimental data from scanning probe microscopies, 5-30 GB • All seven of the above case studies will increase in the amount of data produced as faster hardware is deployed: for case 1 we will be able to treat a larger system and/or larger basis sets; case 2 longer time scales; case 3 larger supercell sizes; case 4 enhanced details for models of inhomogeneities, greater number of calculations; case 5 larger system sizes; case 6 larger system sizes; case 7, greater number of scans and including larger number of instruments	• 1-2 Gbps	• 1-2 Gbps

2-5 years				
• Same as above	• Same as above but a magnitude of 10-100 larger • Users also need to move experimental and simulation data/results between SNS, EFRCs, CNMS, computational resources (both capacity and capability), home institutions	• Same as above but the data set size will increase toward TB • Computing and data storage needs to be staged to allow seamless transfer of data and simulation capability between different major sites such as SNS/HFIR/CNMS/NCCS/NERSC/ EFRC's/… • Need improved latency	• 10-50 Gbps	• Same as for LAN
5+ years				
•	• By 2018 we expect to have exascale computing capability	•	• Beyond 100 Gbps: case study is to dump the entire system in 32 minutes (a projected 64 PB) ~33 Tbs • High quality of service for network reliability and **latency**	• About a factor of 10 less than WAN

9 Combustion Science at the Combustion Research Facility at SNL

9.1 Background

The goal of combustion science is to gain fundamental and predictive understanding of the complex multi-scale physico-chemical processes involved in combustion energy systems. This research spans atomistic scale chemistry and molecular properties, interfacial science of catalytic reactions and multiphase chemistry, through multiple scales of continuum scale chemical and fluid dynamics. The multi-scale nature of combustion science benefits from the construction of predictive models, such as those describing the interactions of reaction chemistry and microscale mixing, that can be assembled into full predictive models of device scale combustion processes. As such, the community relies on interdisciplinary communication and integration of research results for model development and validation by experiment and direct computational simulation. The community is taking advantage of advanced laser diagnostics and imaging detectors, along with leadership scale computational resources that are producing information at unprecedented rates, and are challenging many assumptions about how the collaborative scientific process can efficiently operate. These changing paradigms often require advanced networking capabilities and new software tools to use them.

The combustion research community is diverse and geographically distributed, involving many research institutions around the world. One of the focal points for the combustion science community is Sandia's Combustion Research Facility (CRF), a special-purpose DOE BES collaborative research facility. It is distinguished by its multidisciplinary integration of research activities that range from exemplary fundamental research in combustion chemistry, reacting flows, and laser-based diagnostics, to applied research focused on high-impact combustion systems such as internal combustion engines, coal and biomass combustion, industrial burners for process heat, high-temperature materials processing and manufacturing, and related environmental and defense applications. The CRF as a whole averages about 100 collaborators (those engaged in a research project with one or more staff) per year and about 900 visitors (those wanting to see the labs, attending technical workshops, or wishing to discuss future collaborations) per year. The core basic research typically involves the broadest collaborative interactions. While most of these activities are experimental, research in theory and computation is an important and growing aspect of CRF collaborative research. Recently, BES/EERE have sponsored the construction of a Computational Combustion Science Facility adjoining the existing office space and laboratory buildings at the CRF. Petascale direct numerical simulations (DNS) and large eddy simulation of combustion currently offer the most challenging requirements to ESnet.

In an effort to anticipate the networking requirements of the combustion science community, the relevant characteristics of examples of CRF experimental and simulation research are summarized below.

9.2 Key Local Science Drivers

9.2.1 Instruments and Facilities

Combustion Experiments

The advent of groundbreaking advanced laser diagnostics, imaging detectors, and computer data acquisition at the Combustion Research Facility has revolutionized experimental investigation of the structure and dynamics of turbulent combustion in the past twenty years. The CRF, and in particular the reacting flow program is configured to make multi-scalar point and line measurements in addition to planar laser-induced fluorescence imaging of select flame marker species (e.g. OH and CH) to obtain the instantaneous spatial flame structure and mixing field. In addition particle-based velocity measurements are used to characterize the flow field. More recently, the efforts are focused on time-resolved imaging of transient flame structure undergoing extinction and reignition in turbulent jet flames. The resulting unique, detailed experimental benchmark data serve as the basis for evaluation and development of turbulent combustion models throughout the community. In particular, these efforts plus those of other researchers throughout the world are coordinated and leveraged through the CRF led International Workshop on Measurement and Computation of Turbulent Non-Premixed Flames. Within this construct, researchers meet to exchange data, validate and develop statistical models against the benchmark data, and plan new experiments on a biennial basis. Through constant web and email collaborations this consortium of researchers has made progress well beyond that possible by individual researchers. Current and projected data rates are of the order of 100's of GB of raw data from a given experiment.

Combustion Simulations

The goal of the simulations is to advance the state of the art in the understanding and predictive modeling of reacting flows through the use of detailed computational studies. The work in this area involves investigations of fundamental turbulence-chemistry interactions using complimentary high-fidelity numerical simulation approaches, direct numerical simulation (DNS) and large-eddy simulation (LES), and the analysis and validation of chemical models in the context of low Mach number laminar flame-flow interaction.

Direct Numerical Simulation (DNS) is a first-principles description of chemically reacting flows, i.e. a description based on continuum-mechanics statements for conservation of mass, momentum and energy. DNS provides unique high-fidelity descriptions of turbulent convective transport, molecular diffusion transport and chemical kinetics, ideally suited to studying chemistry-turbulence interactions in flames, in which all relevant physical and chemical scales are resolved, both in space and time. Because of its stringent spatial and temporal resolution requirements, DNS is a computationally intensive approach that requires massively parallel computing power and its domain of application is limited to fundamental studies in canonical configurations. With the continual and fast-paced growth in scientific computing hardware and software technologies, DNS is the natural companion of detailed experimental laboratory-scale studies of ignition and combustion at moderate Reynolds numbers. These simulations are being used to address issues ranging from reactive scalar mixing, extinction and

reignition, flame propagation and structure in uniform and stratified mixtures, flame stabilization in autoignitive flows, autoignition under homogeneous charge compression ignition (HCCI) environments, and differential transport of soot in turbulent jet flames. In addition to the new understanding provided by these simulations, the resultant data are being used in apriori and a posteriori validation and improvement of engineering subgrid models in both RANS and LES. The modeling formalisms have evolved towards regime independent models that can capture different simultaneous modes of combustion, and modes that depend heavily on turbulence-chemistry interactions. Therefore, models are expected to differentiate fuel and kinetic effects and their coupling with turbulent mixing processes.

Large Eddy Simulation (LES) is a technique that employs filtering to account for a larger range of scales in practical flows. The large energetic scales are resolved directly, whereas the sub-filter scales are modeled. Primary efforts in the area of LES are to establish a set of high-fidelity, three-dimensional, computational benchmarks that identically match the geometry (i.e., experimental test section and burner) and operating conditions of selected experimental target flames and to establish a scientific foundation for advanced model development. The goal is to provide direct one-to-one correspondence between measured and modeled results at conditions (high Reynolds number) unattainable using DNS by performing a series of detailed simulations that progressively incorporate the fully-coupled dynamic behavior of reacting flows with detailed chemistry and realistic levels of turbulence. The LES resolves the large-scale structures and rely on subgrid models for turbulent mixing and reaction, and therefore are complementary to the DNS efforts that focus on small-scale mixing and reaction. The focal point is the series of flames that have been studied as part of the experimental reacting flow research program at Sandia's Combustion Research Facility. Information from these simulations combined with detailed laser-based experiments of carefully designed benchmark flames and ignition problems present new opportunities for understanding of turbulence-chemistry interactions and for the development of predictive models for turbulent combustion in practical devices. State-of-the-art petascale DNS and LES solvers are routinely running at DOE Leadership-Class facilities at ORNL through INCITE allocations and at NERSC through ERCAP allocations.

9.2.2 Process of Science

The scientific process for petascale simulation involves several stages categorized as:

Production run preparation and software readiness

• Determine and implement optimum programming model for multi-core processors through performance monitoring and use of OpenMP and MPI, OpenCL and CUDA for hybrid gpu-based architectures.

• Implement and test collective I/O and standardized I/O formats (pnetcdf or hdf5) scalable to peta and exascale architectures using ADIOS.

• Perform preparatory runs – i.e. coarse mesh runs to determine spatial resolution requirements and refine selection of numerical and physical parameters for production runs.

Perform production run

• Submit 10 million cpu-hr job on Cray XT5 at ORNL on the 2 Petaflop machine running on between 100,000-200,000 cores that will run for 7-10 days. This may happen 3-4 times in a year (involves ~5 people).

• Write restart files out each hour (0.6 TByte per file using collective I/O). Over the course of the run an aggregate of about 500 TB field data and 50 TB particle data is written to scratch.

• Monitor the health of the run daily to send diagnostics (x-y plots, isocontour plots back to SNL Livermore, CA) or using the Dashboard coupled with Kepler automated workflow to monitor data and manage data movement.

Postprocessing

• Morph data to N processor-domains for analysis and visualization.

• Archive restart files and morphed data to HPSS at ORNL or NERSC. The data needs to be archived for 5-10 years since it will be revisited multiple times by the modeling community.

• Move morphed data from scratch disk to analysis machine (Beowulf cluster or SGI Origin at ORNL).

• Move morphed data from scratch disk to cluster at SNL, Livermore for analysis and parallel volume rendering. (protocols used include parallel streams using bbcp or rsync with 4 streams.)

• Perform parallel analysis and visualization on data. This is a highly iterative process with portions of the analysis requiring new analysis and other portions relying on existing analysis approaches. The first stage of analysis (typically up to 1 year after data is generated) is performed by the data creators. Subsequent analysis of the data is by collaborators at universities and labs in the U.S. and abroad.

Current collaborators are at University of California at Davis, Iowa State University, Lawrence Livermore National Labs, Stanford University, Princeton University, University of Illinois, University of Utah, and Cambridge University.

9.3 Key Remote Science Drivers

9.3.1 Instruments and Facilities

The CRF is a collaborative research facility that houses a wide array of mid-scale computer clusters, which support both production calculations and staging of larger calculations to DOE capability class platforms. Collaboration is central throughout the facility and with our external partners in both academia and industry. A critical component of the success of the overall program is the intentional overlap and interaction between the individual investigators, both internally and externally. We work to nurture and encourage an atmosphere of close collaboration in order to assure that our overall program is far more than the sum of the individual research components. In this regard,

the exchange of data between the CRF, various DOE capacity and capability computational facilities, and between our partner institutions is a critical driver.

9.3.2 Process of Science

A key to the success of the CRF continues to be the interdisciplinary, collaborative nature of the center. Because researchers within the CRF explore the full range of combustion science, the strength of the program is enhanced by the close proximity of staff working on different research topics using a variety of techniques. As one example, Rob Barlow, Jonathan Frank and Joe Oefelein have established close collaborations that combine the complementary merits of experiments and numerical simulations. Joint research using advanced laser diagnostics in combination with LES has provided new insights in the area of turbulent reacting flows. Combining simulations and experiments, with emphasis placed on common experimental configurations such as those associated with the Measurement and Computation of Turbulent Nonpremixed Flames (TNF) workshop, continue to provide more detailed pictures of the complex turbulence-chemistry interactions that occur in devices such as aircraft and automobile engines. Recent imagings of the rate at which fuel and air mix in a reacting system, for example, have revealed that the structure of turbulent mixing consists of elongated and convoluted filaments. These images provide significant insights for development of turbulent combustion models and the validity of various modeling assumptions. Using these data, a set of high-fidelity LES calculations can first be validated for accuracy, and then used to extract information not available from the experimental measurements. The companion calculations have the potential to provide a quantitative picture of the corresponding three-dimensional mixing dynamics, which is an essential ingredient toward development of predictive combustion models. All of these interactions require efficient transfer of both experimental and LES data between various collaborative institutions and DOE computer centers such as LBNL, NERSC, and ORNL NCCS.

As a second example, Dr. John Dec in the Engine Combustion department has regular conversations with Dr. Tom Settersten, a developer of laser diagnostics. John pointed out the importance of H2O2 as an intermediate in HCCI ignition and asked if it could be imaged. This question initiated a novel effort in the laser diagnostics lab to develop a diagnostic based on photo-fragmentation of H2O2 combined with fluorescence detection of the OH that is formed. In one scenario, the OH that is formed directly in the A-state provides a fluorescence signature of the OH originating directly from the fragmentation process itself. While much remains to be done before such a measurement could be applied to an engine, proof of concept measurements indicate that techniques could be used in controlled ignition experiments, which in turn would be used for comparison to DNS results from Dr. Jackie Chen in the reacting flows program. Routine transfer of DNS data (both raw and reduced) from ORNL NCCS is an imperative requirement to facilitate such collaborations.

9.4 Local Science Drivers – the next 2-5 years

9.4.1 Instruments and Facilities

Combustion Experiments

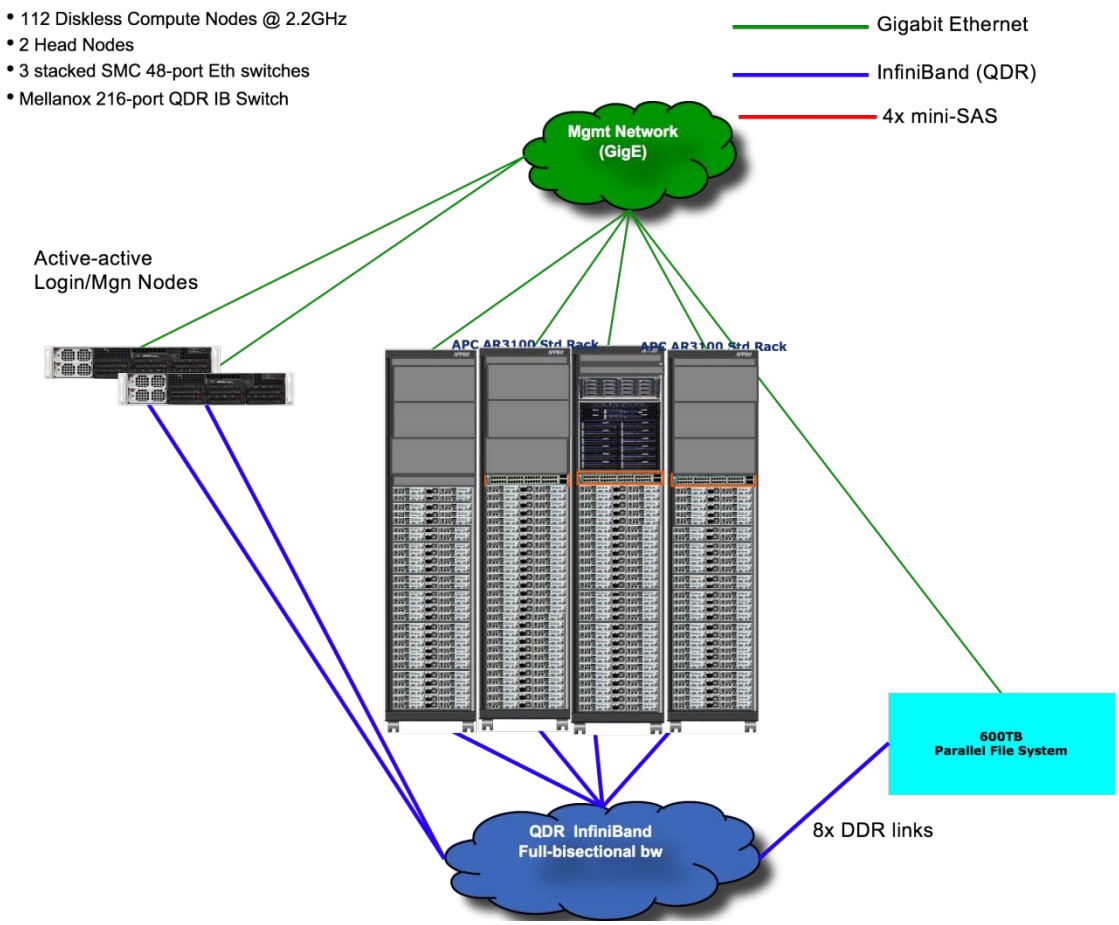
The volume of imaging data will steadily increase as we make significant capital investments – on the order of $1 million in the Advanced Imaging Laboratory (AIL) – in new high-speed imaging systems that will be unique worldwide. In one recent visit of an AIL collaborator, we acquired approximately 1TB of data. Currently, the AIL has its own server with a 13TB raid stack, of which we are using approximately 4TB to date. It's also very helpful to have a multi-processor capability for analysis. The AIL currently has an 8-core system. We store and analyze data on this system. Collaborators remotely access the data on the server after they leave Sandia. Thus far, it has been easy for us to share results of the analysis and for the AIL principal investigator to examine the results of an analysis that others are doing when a problem arises. There are definite advantages to having a server dedicated to the AIL. For example, we have complete control over it and can fix or modify it quickly without requiring a staff of computer support people. Furthermore, since the computer usage is limited to those who are collaborating with us, we have ready availability of CPU time and storage space.

Combustion Simulations

Early in FY11 the CRF will complete the installation of a new, $1.75-million computing cluster from Appro. This new facility will contain 112 compute nodes with 12-core 2.2GHz AMD Magny-Cours processors in a quad-socket configuration, complemented by dual-socket Magny-Cours servers as redundant head nodes. The configuration includes a non-blocking QDR InfiniBand fabric made up of a single Mellanox IS5200 216-QDR-port director-class switch to provide the required Full Bisectional Bandwidth (FBB) across the entire cluster for all of the nodes. An additional set of Gigabit Ethernet switches provides connectivity to all of the nodes for performing the management functionality. The 112 compute nodes are spread over four racks.

A 600 TB parallel file system is included. Bright Computing cluster management software is integrated with the cluster hardware, offersing integrated cluster & workload management, support for redundant head nodes, and diskless compute node provisioning capabilities.

Two networks tie the cluster together. The first network is the InfiniBand network and all nodes have a single connection to this network including the compute nodes and the active-active and redundant head node pair. The IB switch is Mellanox 216-port QDR full bisection bandwidth (FBB) switch, including the fabric management software FabricIT from Mellanox. The second network is the Gigabit (GigE) Ethernet Management Network. The cluster is supported by three (3) 48-port GigE switches with connectivity to all of the node types. The cluster network diagram is shown in the figure below.



CRFC Cluster High Level Network Connection

The cluster has an aggregate peak capability of 40.1 TF. Each node has 96GB of memory that is supported by twenty-four 4GB DDR3-1333 memory modules, for a total of 96GB per node. The memory is dual-rank which can operate at true 1333MHz system speed, without any down-clocking by the processor and chipset. The memory is ECC with chipkill that can detect single- and multi-bit ECC memory errors.

The compute nodes are interconnected by a non-blocking, full-bisectional bandwidth, fat-tree InfiniBand network, operating at QDR speed. A summary of the combined systems overall capability and capacity is as follows:

- Peak Capability for a compute node – 422 GFlops at 2.2GHz CPU speed
- Aggregated Peak Capability for all compute nodes – 47.3TF
- Main Memory Capacity per compute node CPU core – 2GB; or 96GB
- Main Memory Capacity for the CRFC cluster – 9.12 TB

The peak capability ranks this cluster roughly in the top quartile of the Top 500 open supercomputer platforms worldwide.

9.4.2 Process of Science

Production runs, in-situ analysis and visualization, and data sharing

By 2011, DNS simulations on 2 Petaflop production runs will produce 1 Pbyte of data per year at NCCS/ORNL, and by 2012 20 PF runs will produce 10 PB of data per checkpoint file (i.e. 10 billion grid * 80 variables *8 bytes/variable) = 6.4 TB. If 200 checkpoint files are written out at approximately 1 file per hour in a 7-10 day run, the data generation rate ~ 20Gbps. If we move data elsewhere at the rate it is generated then we need a network to move data at ~20Gbps. The network will need an even higher bandwidth to account for overheads due to protocol, metadata, contention, etc.

The goal is to stream data to an analysis and rendering machine as it is produced rather than waiting until the run is complete. In this manner, known analysis tools can be applied to the data as it is generated to get a first glimpse understanding of the underlying physics, and subsequent iterative analysis can be performed off-line. Automated workflow scripts using Kepler will facilitate the data streaming, morphing, archival, and analysis. Data will need to be moved to platforms and archival storage within a supercomputing center as well as to SNL Livermore, CA to two open network clusters for analysis and rendering.

DNS simulations will be instrumented with in-situ feature detection, segmentation and tracking to enable data reduction and querying on-the-fly, thereby reducing the amount of data for further analysis and enabling steering of adaptive I/O.

Web-based portal developed for sharing simulated benchmark data with modeling community of ~50-100 international collaborators at universities, national labs, and industry. A scalable, extensible framework will be developed for analyzing large data and comparing data with experiments. The framework will include capability for standardized formats, translators, graphics, parallel library of combustion analysis software, parallel feature detection/tracking library, inference software for automatic model generation, and query tools that can operate on portions of the data at the supercomputing facilities where the data resides. Reduced data and remote visualization results will be sent back to institutions via ESnet.

9.5 Remote Science Drivers – the next 2-5 years

The key remote science driver is the extensive list of collaborators who will need access to the data, which resides at NCCS/ORNL, NERSC, and SNL-CRF. The list of collaborators who will access the DNS data includes:

- Professor Chung K. Law, Princeton University
- Dr. Tianfeng Lu, U. Connecticut
- Professor Heinz Pitsch, Stanford University
- Dr. David Cook, Bosch Corporation
- Professor J. Y. Chen, University of California at Berkeley
- Fabrizzio Bisetti, University of California at Berkeley
- Professor Phil J. Smith, University of Utah
- Professor James Sutherland, University of Utah
- Dr. Chunsang Yoo, UNIST, S. Korea

- Dr. Evatt R. Hawkes, University of New South Wales, Sydney, AU
- Dr. Ramanan Sankaran, Oak Ridge National Laboratories, Tennessee
- Dr. Edward Richardson, University of Southampton, UK
- Dr. Ray Grout, National Renewable Energy Lab, Golden, CO
- Dr. Scott Klasky, Oak Ridge National Laboratories
- Professor Kwan-Liu Ma, University of California at Davis, CA
- Dr. Chaoli Wang, Ohio State University, Columbus, Ohio
- Professor Valerio Pascucci, U. Utah, Salt Lake City, Utah
- Professor Stewart Cant, Cambridge University, UK
- Dr. Andrea Gruber, SINTEF and U. Trondheim, Norway
- Dr. David Lignell, Brigham Young University, Utah
- Dr. Yuxuan Xin, Princeton University, New Jersey (Ph.D. student)
- Professor Stephen Pope, Cornell University, Ithaca NY
- Dr. Benedicte Cuenot, CERFACS, Toulouse, France
- Drs. Christian Angelberger and Cecile Pera, Institut for Petroleum (IFP), Paris, France
- Dr. Ed Knudsen and Heinz Pitsch, Stanford University, CA
- Professor Heinz Pitsch, Aachen, Germany
- Dr. Andrew Wandel, Queensland University, Queensland, AU
- Professor Suresh Menon, Georgia Tech U., Atlanta, Georgia
- Dr. Santosh Hemandra, University of Aachen, Germany

Those who will need access to the LES data include:

- Professor A. Dreizler, Technical University of Darmstadt, Germany from CRF
- Professor D. Haworth, The Pennsylvania State University from CRF
- Professor A. Kempf, Imperial College London, UK from CRF
- Professor S. Menon, Georgia Institute of Technology from CRF
- Professor C. Merkle, Purdue University from CRF
- Professor C. Rutland, University of Wisconsin, Madison from CRF
- Professor V. Sick, University of Michigan to/from CRF
- Dr. J. Bell, Lawrence Berkeley National Laboratory to/from LBNL and CRF
- Dr. T.-W. Kuo, General Motors R&D Center from CRF
- Dr. K. Tucker, NASA Marshall Space Flight Center from CRF
- Dr. D. Talley, Air Force Research Laboratory, EAFB, CA from CRF

We also need to transfer data transfer between the CRF and the DOE computer centers at ORNL, ANL, and LBNL to support staff and our in-house collaborators

9.6 Beyond 5 years – future needs and scientific direction

We envisage as machines progress towards exascale, it will be possible in combustion to consider hybrid multi-scale simulations of turbulent combustion in the gas phase interacting with soot particle formation in a fully coupled simulation. Similarly, it will become possible to consider multi-phase turbulent combustion (dense and dilute sprays) through atomistic simulations of the condensed phase and to handle their coupling with the continuum gas-phase turbulent combustion. A third example would be to consider

the deflagration to detonation transition in simulations that handle both flame propagation as well as embedded shocks – again, a hybrid simulation involving solution of a large system of PDE's using method of lines together with Monte Carlo or Molecular Dynamics approach for treating internal shock structure. Therefore, the analysis and visualization software framework will need to handle heterogeneous data types to an even greater extent than before. In addition to multi-scale simulations it may be possible to embed uncertainty quantification within DNS to understand uncertainties in kinetic and transport properties and their propagation to macroscopic combustion quantities.

We also anticipate there will be more emphasis placed on in-situ segmentation, tracking and query-driven analysis and model inference to glean insight from such large data. Hence, there may be a shift in paradigm from moving large amounts of data over the network to greater emphasis on remote intelligent and versatile data reduction and visualization.

9.7 Middleware Tools and Services

A key role of scientific data management middleware will be to realize a software infrastructure for exploiting today's petascale and tomorrow's exascale architecture's ability to overlap immediate data analysis with I/O. This overlap will connect the algorithmic innovations at the higher levels with designs and utilization of the deep and heterogeneous memory hierarchies to come. We will build on our earlier work to develop robust techniques for future exascale applications and machines. In particular, a key issue will be to ensure reliable and resilient data generation, delivery, and storage at extreme scale. We propose to use the computational power of processors to reduce the need for data output by building I/O systems that enhance massively parallel methods for data movement (e.g., many network links or many OSTs) with extensive processing resources to manipulate data as it is being moved. Our approach to implementing such enhancements is a `staging approach' to high performance I/O in which staging resources comprised of many processors with local storage are tightly coupled with the petascale machine, and these resources can be used for immediate data analysis and visualization. Specially, we will leverage the existing ADIOS I/O middleware, which is already used by several of the combustion simulations, to immediately focus on data resiliency and quality. Two important aspects of this system that make this transition feasible are that it: 1) reduces output volumes through immediate and `in transit' processing of data, and achieves improved resilience through strategic redundancy and run-time error checking; and 2) makes no changes in the simulation codes, but uses new methods to incrementally use more powerful elements of the proposed I/O infrastructure to improve the ways in which insights are gained from data produced by the applications. Within the data management general infrastructure we will develop in situ data reorganization and processing components with minimal overhead to the running simulation. Building on the ADIOS componentized SDM solution approach provides multiple advantages:

(1) Scientists can construct desired output pipelines using both commonly available and custom components;
(2) Scientists can manage and adapt these pipelines at runtime, for instance, to focus on particularly interesting data;

(3) Resources allocated to output pipelines can be dynamically adapted to meet current science needs, for example, to conserve resources for applications whenever possible;
(4) Output pipelines can be monitored via web-based interfaces, which gives users immediate and interactive access to their codes' data; and
(5) Data can be moved to remote sites for inspection and analysis.

The solutions developed will also incorporate specific support for code coupling services, realized in ways that make it easy for end users to experiment with various coupling methods and science-appropriate algorithms. Finally, there will be well-defined interfaces for using both custom data visualization methods and for interacting with rich visualization toolkits.

9.8 Summary Table

Key Science Drivers			**Anticipated Network Needs**	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• 2 pflop DNS/LES simulations running for 1 week on Jaguar at ORNL • INCITE DNS/LES running for 1-2 weeks at a time, 3-4 heroic runs per year at ORNL.	• Data generated on petascale machines at ORNL/ANL and NERSC • Some analysis performed at ORNL on ewok, lens, or in situ on Jaguar and other analysis/viz and data reduction is performed at SNL or NERSC. Some analysis and data queries are performed via client/server arrangement between ORNL/SNL. • Data sharing with geographically distributed collaborators (US, Europe, Australia)	• Write restart files out each hour (0.6 Tbyte per file using collective I/O). • Over the course of a run an aggregate of about 500 TB field data and 50 TB particle data is written to scratch.	•	• 20Gbps WAN bandwidth need • Key sites: ANL, NERSC, ORNL, LLNL, SNL, UC Davis, U. Utah, Stanford U., Princeton U. • In situ real time remote analysis and rendering, client/server arrangement

2-5 years				
• 20 pflop DNS/LES simulations running for 10-12 days per run	• 50-100 collaborators around the world • International Workshop on modelvalidation by comparing DNS/LES and experimental data against models.	• Write restart files out each hour 6 Tbyte per file using collective I/O) • Over the course of a run an aggregate of about 5 PB field data and 500 TB particle data is written to scratch	•	• 100Gbps WAN bandwidth need • Connections to universities, labs, industry • Domestic and International
5+ years				
• 200 pflop DNS/LES simulations performed with embedded UQ and in situ analysis/viz	• Extensive community-wide collaboration and data sharing through web portal or Combustion Hub	• Write restart files out each hour 60 Tbyte per file using collective I/O • Over the course of a run an aggregate of about 50 PB field data and 5 PB particle data is written to scratch	•	• In situ remote visualization, feature extraction/tracking, downstream analysis on PB sized data running on hybrid machines

10 Computational Chemistry at Ames Laboratory

10.1 Background

Our team's focus is on using high quality electronic structure theory, statistical mechanical methods, and massively parallel computers to address the prediction of bulk properties of water systems which require high fidelity. These molecular scale problems are of critical importance to national priority scientific issues such as global warming, the environment and alternative fuels. Many important problems in the chemical sciences are so computationally demanding that the only perceived option for addressing such problems is to employ low-level methods whose reliability is suspect. A primary motivation for our computational work is to ensure that reliable levels of electronic structure theory (e.g. many body perturbation theory and coupled cluster theory) can realistically be employed to solve challenging problems of national interest. We are currently examining three systems: i) molecular level dynamics of water, ii) the formation of aerosols important in cloud formation, and iii) the interactions of dendrimers with ligands of environmental importance. Simulations are run at the Argonne Leadership Computing Facility (ALCF).

10.2 Key Local Science Drivers

10.2.1 Instruments and Facilities

We use a variety of Mac and PC workstations to receive simulation data from ALCF for further processing. The LAN is 100Mbps Ethernet.

10.2.2 Process of Science

The transfer of input data files required to start the simulation is on the order of several MB. A single simulation will generate 10 to 100GB of data, depending on the size and length of the simulation. Most of the post-processing analysis and visualization is done at ALCF. Smaller subsets of data (~10GB) are transferred from ACLF to Ames Lab and/or Iowa State University by scp/sftp for further post-processing.

10.3 Key Remote Science Drivers

10.3.1 Instruments and Facilities

We remotely connect (ssh) to the supercomputer at ALCF to launch the simulations.

10.3.2 Process of Science

Molecular dynamics simulations are run at the Argonne Leadership Computing Facility (ALCF). Data sets range from 100GB to 10TB, depending on the size and length of the simulation. Because of the data set size, most of the post-processing (analysis and visualization) is conducted at ALCF. Subsets of data (1 - 10GB) are transferred to Iowa State University by scp/sftp for analysis.

10.4 Local Science Drivers – the next 2-5 years

10.4.1 Instruments and Facilities

Some campus buildings (Iowa State University) will have the LAN upgraded to a fiber optic-based network in the next year.

10.4.2 Process of Science

Our simulations will move to petascale supercomputers and we will generate significantly more data. We will continue to transfer subsets of data to Ames Lab/Iowa State University and these subsets will be larger (100GB) due to increased system sizes and length of simulation trajectories.

10.5 Remote Science Drivers – the next 2-5 years

10.5.1 Instruments and Facilities

We will continue to run simulations remotely on DOE supercomputers. We anticipate doing more post-processing of the data sets at the remote facilities, including remote visualization.

10.5.2 Process of Science

Some of this data (e.g. high fidelity molecular dynamics of water) will be in demand by the scientific community. We will need solutions for sharing TB data sets with the science and engineering community.

10.6 Beyond 5 years – future needs and scientific direction

Our team's vision is to create a portal for the data generated by all the petascale simulations conducted by our team so that others in the community can interrogate the data to answer scientific questions of interest.

10.7 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• ALCF	• Transfer subsets (10GB) of data to local institution	• 10GB/day • Single file archive	• 1-2 hours	• 1-2 hours • From ALCF to local institution
2-5 years				
• Continued use of DOE supercomputing facilities	• Transfer subsets (100GB) of data to local institution • More emphasis on remote analysis and visualization	• 100GB/day • Single file archive	• 1-2 hours	• 1-2 hours • From supercomputing facility to local institution
5+ years				
• Continued use of DOE supercomputing facilities	• Access portal for petascale simulation data sets	•	• 1-2 hours	• 1-2 hours • From supercomputing facility to local institution

11 Computational Materials Science

11.1 Background

The following section is based on experience with three projects/facilities: user projects from the Center for Nanophase Materials Sciences (CNMS, http://www.cnms.ornl.gov/), and the theory components of two Energy Frontier Research Centers (EFRCs), the Center for Defect Physics (CDP, http://www.ms.ornl.gov/cdp/index.shtml) and the Fluid Interface Reactions, Structures, and Transport Center (FIRST, http://www.ornl.gov/sci/first/index.shtml). The networking requirements of the CNMS are also discussed in an earlier section of this report.

Computational materials science – a subject that overlaps with computational physics, chemistry, nanoscience, and engineering – has highly varied simulation requirements owing to the broad range of lengthscale, timescale, and accuracy requirements for research level property prediction and explanation. The range of data sizes, degree of interactivity, and associated networking requirements are consequently extremely varied. New methods are also expected to be invented and deployed during the next 5 years.

To make significant advances in the areas of interest to DOE, it is proving crucial to increase the realism of the models employed and to extend previously static models to incorporate dynamics, reactions, and phase transformations. Today, for all but the simplest of these systems, very few of these problem dimensions can be considered well converged. Future networking requirements will be driven by the improvements in computational methods, algorithms and architectures that enable these model dimensions to be systematically explored. Past trends indicate that there will be a commoditization of today's exotic methods – such as automated reaction path discovery – and the continued development of entirely new methods that extend the complexity of systems and properties that can be modeled, analyzed, and understood.

The materials science problems investigated by computation inevitably contain a variety of parameters of interest, e.g. temperature or dopant concentration in an alloy. Thus although select problems in computational materials science are able to utilize the largest supercomputers, more typical investigations consist of dozens, perhaps hundreds, of individual calculations at a lesser scale. Management of data and data provenance is an increasingly common issue. For example, in computational materials design, it is now common to combinatorially calculate the properties of tens of thousands of compounds. Subsequent analysis is usually performed at a home institution, inevitably requiring transfer of all data back to the home site.

11.2 Instruments and Facilities

Calculations in computational materials science are typically performed on institutional computer clusters and at National level supercomputing facilities (e.g. ALCF, NERSC, OLCF). Although these investigations are now often paired with simultaneous experimental investigations, each branch of investigation is usually performed independently owing to the incompatible scheduling of the very different calendar scales required for simulation and experiment.

11.3 Process of Science

First, typically, a range of input parameters or models are constructed, either by hand or automatically. For example, in a large-scale atomistic simulation of the properties of metal alloys, a variety of defects, stacking faults, and impurity levels would be constructed. Second, molecular dynamics would then be performed at a supercomputer site, after transferring these input data. Trajectories would be periodically saved. This implicit data reduction is already necessary to keep data sizes manageable. Finally, a subset of trajectories would be retained and transferred back for analysis to users home workstation or computer cluster. In the most challenging cases (millions of atoms for millions of timesteps), analysis might be performed at the supercomputer site to avoid data movements, or be performed on the fly.

Grid approaches to the above are not widely used despite some suitability, presumably due to the maturity of the approach. A large investment of effort is required to setup a robust grid computing structure that can e.g. recover from most calculation failures without human intervention.

11.4 Key Remote Science Drivers

11.4.1 Instruments and Facilities

Access to large-scale computational resources and significant cloud-based computational resources are the major driver for remote science. Each of the computational methods employed (finite element, classical molecular dynamics, ab initio simulation) has the potential to generate very large quantities of data (TB to PB). Since the exact nature of the analysis that will be required is often not known in advance, long-term remote storage of this data is important. On the fly analysis is only possible where, e.g. the fundamental mechanisms and processes are well understood.

11.4.2 Process of Science

Same as above.

11.5 Local Science Drivers – the next 2-5 years

11.5.1 Instruments and Facilities

The next 2-5 years is likely to see a significant shift in the science performed locally, owing to the architectural shifts underway in computing: the (re)emergence of compute accelerators such as graphics processing units (GPUs) and the increasing core count of modern processors. These developments (“hybrid multicore”) are likely to democratize the levels of computation available, particularly facilitating use of methods that have limited parallelization. An example of would be the availability of GPU implementations of classical molecular dynamics methods that offer 1-2 orders of magnitude improvement in time to solution for small problems. The use of these architectures for computational materials science at scale will depend on the availability of new suitable implementations of simulation methods that can successfully hide the communications costs.

The networking transfer requirements will typically scale linearly-quadratically with the increase in computational resource made available to researchers. Many of the data sets utilized in computational materials science scale primarily linearly (e.g. with simulated timescale), but some data sets scale quadratically (e.g. the extended plane –wave wavefunctions of condensed matter density functional implementations scale quadratically with system size).

A second shift in network requirements will result if the concept of high throughput computation is more widely adopted. In this eventuality, it will be more common for researchers to investigate whole classes of materials by rapidly computing their properties. This approach is analogous to drug candidate screening that has been adopted in the pharmaceutical industry. The few high profile successful examples of this approach in materials, e.g. Prof. Ceder's battery research at MIT, have utilized departmental-level compute clusters. However, in future, these calculations are likely to be run at remote supercomputing sites.

11.5.2 Process of Science

Same as above.

11.6 Remote Science Drivers – the next 2-5 years

11.6.1 Instruments and Facilities

Same as above.

11.6.2 Process of Science

Same as above.

11.7 Beyond 5 years – future needs and scientific direction

The next 5 years will see the emergence of numerous multi-petaflop computational installations and a few installations reaching towards the exascale. This increase in computational capability will come at the expense of architectural shifts: different forms of compute elements, and a lower availability of memory per compute element than is the norm today. Two orders of magnitude improvement in time to solution has already been achieved through the use of GPUs for some limited problems and computational methods: conceivably this could result in a two orders of magnitude increase in the data produced, stored, and transferred. However, it is also probable that this increase in performance will be used to increase model complexity (e.g. size) or accuracy, so that the increase in data requirements is more modest.

Due to the diversity of approaches within computational materials science combined with the architectural uncertainties, it is likely that many methods will be able to use many-petaflop scale resources, but it is too early to predict which ones. The new architectures may enable the use of new methods or stimulate the development of new ones. The uncertainly in the methods that will be used in 5 years generates significant uncertainty in the networking requirements. It is certain that some investigations will be able to utilize a higher "level of theory", for more accurate simulations. However, ***it is already the case***

that huge and unmanageable quantities of data can be generated. In either classical or ab initio molecular dynamics, the millions of generated frames (timesteps) can easily generate petabytes of data. Similarly, in combinatoric investigations of materials, a few tens of megabytes of output data per configuration can easily result in PB sized data sets.

The overall scientific direction in five years is likely to be one of increasing complexity, with the increased computational power enabling the universality of different types of material behavior to be elucidated, contrasting with today's calculations that often focus on the specifics of individual systems.

11.8 Middleware Tools and Services

Computational material science investigations often require a range of size of calculation, with a significant number being small enough to be amenable to a grid-computing environment. Although in principle these environments can be utilized today, the startup effort required remains significant. The largest "hero" runs inevitably require tuning and careful setup at a specific supercomputing resource, but all other computations could in theory be performed entirely within a grid or cloud environment.

Science is increasingly being performed by diverse and geographically dispersed teams. E.g. The Energy Frontier Research Centers involve teams of 20-50 students, postdocs, professors, and staff scientists. Ad-hoc video conferencing and screen sharing for meetings remains challenging owing to the bandwidth and quality requirements, requirement to involve many diverse sites (DOE, University), as well as local networking and security requirements.

11.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• Ensemble of TFLOP to PFLOP calculations run at NERSC, OLCF, ALCF • Small scale testing and computation on new architectures	• Data generated and partially reduced, primarily at remote site • Analysis performed at home institution	• MB-TB/day • Split over a few large files, e.g. trajectories, wavefunctions	• O(hours)	• O(hours) return to home institution, e.g. ORNL.
2-5 years				
• Larger runs at ALCF, NERSC, OLCF	• Where practical, process unchanged • Analysis performed on the fly or at remote supercomputer site where data transfer and storage impractical. • Increasing use of automation to explore parameter space and perform analysis	• Increase linear-quadratically with computational resources • Highly dependent on methodological advances and spectrum of methods employed	• O(hours)	• O(hours)
5+ years				
• Larger runs, exploring longer timescales using more accurate methods, e.g. quantum instead of classical • Computation of new properties, e.g. optical properties at finite temperature, wear rates under realistic conditions	• Where practical, process unchanged • Analysis performed on the fly or at remote supercomputer site where data transfer and storage impractical. • Increasing use of automation to explore parameter space and perform analysis	• Increase linear-quadratically with computational resources • Highly dependent on methodological advances and spectrum of methods employed	• O(hours)	• O(hours)

12 Materials Simulations at Ames Laboratory Using NERSC Supercomputers

12.1 Background

Computational condensed matter physics and computational materials science is an important component of materials research program at Ames Laboratory. First principles quantum mechanics calculations and atomistic simulations (molecular dynamics and Monte Carlo) are the commonly used computational tools to study the structural, electronic and dynamical behavior of materials under various conditions. While the small to mid scale calculations are done using the local PC-clusters at the Ames Laboratory, most of the large-scale computational studies are performed at the supercomputers at the DOE's centers such as NERSC. There is a need for data transfer between the local computers and the computers or storage devices (e.g., HPSS) at NERSC for analysis and visualizations. Examples of such data are the wave functions (or electronic charge densities) from first-principles quantum mechanic calculations (e.g., VASP) and atomic trajectories from long-time and large-scale molecular dynamics simulations. The sizes of such data range from tens of GB to several TB. Sometime remote code debugging is needed in order to optimize the performance of the codes on the NERSC's supercomputers, especially when new machines are put into use.

12.2 Key Local Science Drivers

12.2.1 Instruments and Facilities

Currently the Division of Materials Science and Engineering at Ames Laboratory has several PC-clusters with 32-64 computer nodes interconnected by gigabyte network such as InfiniBand technology. The newly installed cluster *Dense* has 7 TB of hard disk storage capacity. Most job submission and data analysis are done on local Linux or Windows based workstations. The clusters and workstations are locally interconnected by 100Mbps or 1Gbps Ethernet.

12.2.2 Process of Science

The computational algorithm and code developments are mostly done using the local computers and then optimized for large-scale computation on the supercomputers at NERSC. Small to mid scale scientific calculations and simulations are also done locally. Large-scale calculations are done at NERSC. Currently most of the data generated using the supercomputers at NERSC (such as wave functions from first principles quantum mechanics calculations and atomic trajectories from large scale molecular dynamics simulations) are transferred back to the local clusters or workstations for further analysis and graphic visualizations. Most of users login to NERSC computers by SSH and upload/download data by SCP. It would be more efficient if the graphic visualization can be done at NERSC to reduce the amount of data transfer. However, using current tools, network performance is too low to perform remote graphic visualization at NERSC from the terminals at Ames Laboratory. Remote debugging at NERSC is difficult due to the performance issues related to the tools in current use.

12.3 Key Remote Science Drivers

12.3.1 Instruments and Facilities

Franklin, Carver, and Hopper at NERSC are the most frequently used supercomputers by our group at Ames Laboratory. We also use HPSS for data storage at NERSC. Some useful X-windows based visualization software such as IDL, VMD, and XCrysDen are also available at NERSC.

12.3.2 Process of Science

We use the supercomputers at NERSC to perform large-scale first-principles quantum mechanical calculations and molecular dynamics simulations to study the structures, electronic and dynamic behaviors of materials under various conditions. Currently, the data generated from such calculations and simulations are transferred back to the local computers at Ames Laboratory for further analysis and graphic visualizations. Some data are also stored at NERSC's HPSS for future use. Some data are too large to be visualized by the local computers at Ames Laboratory. It would be desirable to have the graphic visualization for such big data files using the computers and software provided by NERSC. However, the current network is too slow to satisfy such usage.

12.4 Local Science Drivers – the next 2-5 years

12.4.1 Instruments and Facilities

In the next 2-5 years, we expect new PC-clusters with large disk capacity and computing power will be installed to replace the old ones. Some workstations used by the scientists or students at Ames Laboratory will be upgraded to have better graphic capabilities. Internal network will probably still be 100Mbps Ethernet.

12.4.2 Process of Science

We will develop novel parallel algorithms and codes for computational material discovery. Remote parallel program debugging and graphic visualization on supercomputers at NERSC will be useful. However, full use the software for visualization provided on remote computers at NERSC is limited by the speed of the current network. We hope this situation can be improved in the next 2-5 years.

12.5 Remote Science Drivers – the next 2-5 years

12.5.1 Instruments and Facilities

Better network for fast data transfer to allow remote graphic visualization and on-the-fly analysis of simulation results is desired. Facilities that have web-conference capabilities are also desirable in the next 2-5 years.

12.5.2 Process of Science

For example, the DOE-BES Computational Materials and Chemistry Science Network (CMCSN) projects involved scientists from different institutions (DOE Labs and Non-

DOE Labs or Universities). Large-scale calculations and simulations of the CMCSN projects will be done at NERSC. It is desirable to have the simulation results visualized simultaneously by the scientists at different locations using the graphic software and web-conferencing facilities at NERSC. Because some of these data will be too large to be transferred to different locations, a central visualization of the results through NERSC will promote the discussion and collaboration among the scientists more efficiently.

12.6 Beyond 5 years – future needs and scientific direction

The rapid development of computer technology will enable us to perform calculations and simulations with large number of electrons and atoms that more close to the reality in the next 5 years and beyond. This means an even larger amount of data on the materials properties and behaviors will be generated which will be very difficult to transfer from location to location for analysis or visualizations. For example, a 100 nanosecond simulation of a system of 10 million atoms could potentially generate 4 petabytes of data. A first principles calculation of 2000 atoms will generate wave functions with size in an order of hundred GB. With such huge quantities of data, it is clear that tools and methodologies need to be developed to analyze data “on the fly” or by remote to reduce the volume of the data to be transferred or stored. This will present a big challenge to the speed or bandwidth of the network that allows the science to be performed efficiently.

12.7 Middleware Tools and Services

Web-based audio/video conferencing services that allow the scientists to visualize and discuss simulation results while sitting in front of their desk computer in their office.

12.8 Outstanding Issues

Computational materials science has become a very important branch of materials research. A lot of data about material properties on the electronic and atomistic levels have been generated through computer simulations. It is expected that more and more data will be generated in the years to come. The issue is how to select and store the “reliable” data and make it easily accessible (includes remote visualization) to the materials research community. The Science Gateway Project at NERSC is moving in this direction. More well organized and centralized data collecting, storing, and sharing in terms of “materials genome” will be very useful.

12.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• Large scale *ab initio* calculations and atomistic simulations at NERSC supercomputers that generate huge data sets	• Analysis of electronic density distributions in materials; simulated STM images using the electronic wave functions; study the atomic motions using the trajectories from MD simulations	• Data volume 10GB to 1TB/day	• Most of the data will be transferred within 10 min. Some can be transferred overnight	• Some files (less than 10GB) may need within 5 minutes; some larger files can be transferred overnight. • Data will be transferred from (to) NERSC to (from) Ames Laboratory
2-5 years				
• Large scale *ab initio* calculations and atomistic simulations at NERSC supercomputers that generate huge data sets	• Analysis of electronic density distributions in materials; simulated STM images using the electronic wave functions; study the atomic motions using the trajectories from MD simulations.	• Data volume 1-5TB/day	• Most of the data will be transferred within 10 min. Some can be transferred overnight	• Some files (less than 10GB) may need within 5 minutes; some larger files can be transferred overnight.
5+ years				
• Large scale *ab initio* calculations and atomistic simulations at NERSC supercomputers that generate huge data sets	• Analysis of electronic density distributions in materials; simulated STM images using the electronic wave functions; study the atomic motions using the trajectories from MD simulations.	• Data volume 5-10TB/day	• Most of the data will be transferred within 10 min. Some can be transferred overnight	• Some files (less than 10GB) may need within 5 minutes; some larger files can be transferred overnight.

13 Molecular Foundry Theory Facility, Molecular Foundry, LBNL

13.1 Background

The Molecular Foundry at Lawrence Berkeley National Laboratory (LBNL) is a user facility charged with providing support for research in nanoscience to academic, government and industrial laboratories around the world. User proposals consist of multidisciplinary projects in nanoscience submitted by scientists seeking to enhance their studies of the synthesis, characterization and theory of nanoscale materials with the help of Foundry scientists and facilities. Proposals are reviewed twice a year, and research begins immediately on approval. Only proposals with the highest level of scientific merit are accepted.

Foundry staff members are also charged with developing their own research program in nanoscience. Collectively, these programs comprise the Foundry's multidisciplinary nanoscience program that focuses on the understanding of "soft" (biological and polymeric) and "hard" (inorganic and micro-fabricated) nanostructured building blocks, and the integration of those building blocks into complex functional assemblies.

This report focuses on the network needs of the Foundry's Theory of Nanostructured Materials Facility. Theory Facility staff members Jeff Neaton, David Prendergast and Steve Whitelam have three main areas of expertise: electronic structure, X-ray spectroscopy, and soft matter self-assembly. We run computer simulations on our in-house, 240-node (1920-processor) cluster Vulcan, on Berkeley Lab's Lawrencium cluster (1584 processors), and on NERSC's Franklin cluster (38,288 processors). We have three main network-related needs. First, we transmit files between these clusters in order to do simulations. Second, we need to visualize data housed on these clusters. And third, we need to share these data with national and international users of the Facility.

13.2 Key Local and Remote Science Drivers

13.2.1 Instruments and Facilities

We run computer simulations on our in-house, 240-node (1920-processor) cluster Vulcan (successor to Nano, a 600-processor machine which we continue to operate), on Berkeley Lab's Lawrencium cluster (1584 processors), and on NERSC's Franklin cluster (38,288 processors). Vulcan's 8-processor nodes have 24GB of memory collectively (3GB memory per processor), and are linked by a high-speed interconnect. Lawrencium's and Franklin's processors each have 2GB of memory.

We currently have 10TB bulk storage on local Foundry filesystems, and about 30TB in principle if we add up capacity of all terminals (though this would be hard to use in "bulk").

We shuttle data between these clusters, and from these clusters to the Theory Facility. In general, the average rate of data transfer, rather than the peak rate, is important for our work. NERSC is linked to LBLnet via a 10Gbps connection. LBNL's internal network connects to the Foundry via a 1Gbps link, which is also the connection rate between the

Foundry and Vulcan and Lawrencium. Most 'choke points' for data transfer are local, and are related to the rate that our hosts can write data to disc. For example, on Friday 10th Sep, around noon, we wrote data from Franklin to a Theory Facility terminal's disc at an average rate of 400MB/sec (theoretical network rate 1Gbps). We also tested data transfer rate from the LBNL cluster environment to Franklin. This should be greater than the rate to local terminal disc, but that day was only 40MB/sec (suggesting a connection fault).

We also transmit data from the Foundry to external users, both nationally and overseas, and hold videoconferences (Skype or ichat) with these users.

13.2.2 Process of Science

Our science has important network needs, explained in more detail below. We transmit (often large) files between clusters in order to do multi-stage simulations. We need to visualize data residing on these clusters. And we need to share our data with national and international users of the Facility.

Computing: In the course of our work we run simulations on the compute clusters described above, and analyze the resulting data. We do both parallel and serial simulation. Parallel simulations for e.g. calculating the electronic structure of 100 atoms might use 3000 Franklin processors in parallel for 30 min, and produce 10GB of data. Serial simulations of e.g. protein crystallization might require 100 simulations run simultaneously (for many different values of protein interaction parameters) for 300 CPU hours each, and produce 10GB of data.

Cluster-to-cluster transport: Often, we use the output from one type of simulation as input for another. Massively parallel jobs of e.g. 3000 processors cannot be done on our in-house cluster, so we run these on Franklin. However, since time on Franklin is charged, and there are strict limits on how long simulations can run, we often use the output of a massively-parallel simulation on Franklin (e.g. a multi-TB quantum mechanical wavefunction) as an input file for a set of simulations on Vulcan. Data transfer of this nature often takes about a working day. Cluster-to-cluster transfer within LBNL goes through a single 1Gbps node.

Visualizing data: We usually need to visualize large data sets produced on NERSC or local computer clusters. For example, we calculate the X-ray spectra of e.g. 100 biomolecules in a simulation box. For each biomolecule in a specific conformation we do an electronic structure calculation on Franklin to determine its spectrum from wave functions 10s of GB in size. These spectra are then ensemble-averaged to produce a spectrum than can be compared with experiment. However, in order to understand physically the contribution of each molecule to the averaged spectrum, it is necessary to visualize that molecule's wavefunction. To do so requires downloading 10s of GB to a local workstation. Doing this multiple times would be very revealing scientifically, but is not realistic. Similarly, visualizing these wavefunctions on the cluster cannot be done rapidly enough to be worthwhile. Our capacity for scientific discovery is in this case (and many others) strongly limited by our ability to visualize data, which in turn is difficult with current network and local hardware capabilities.

Sharing data with users: We often need to share the output of e.g. a self-assembly simulation with remote users at e.g. Bath University in the UK. This output might be in the form of a 1-10GB movie file (e.g. a large postscript or pdf document). Direct data transfer is limited by the speed of the direct line between institutions, and bottlenecks reading and writing from and to discs. We also mount such movies on webpages hosted on local servers, so that the user can access them via browser connection. In addition, we hold regular Skype or ichat conferences with remote users.

13.3 Local Science Drivers – the next 2-5 years

13.3.1 Instruments and Facilities

In 2 years we anticipate increasing available Foundry data storage to ~50 TB bulk (from 10TB currently). We anticipate increasing Vulcan's compute capability by 20%.

Boosting the speed of the Foundry's connection to LBNL's core network to 10Gbps can be achieved for ~$10K. Although this boosting was discussed two years ago, it is not currently planned. In part, this is because 'write-to-disc' speed is a key limit on data transfer rate, and boosting network connection speed would not address this problem.

13.3.2 Process of Science

We already generate data files in the TB range that tax existing network capabilities. Our projection is that typical simulation output file sizes will increase by a several multiples in the coming 5 years, but that major increases in speed of scientific discovery will not come about *unless* we find ways of transporting and especially visualizing these data faster.

13.4 Remote Science Drivers – the next 2-5 years

13.4.1 Instruments and Facilities

Last year the Foundry theory Facility's NERSC allocation was about 2 million hours. This year our combined request may top 5 million hours. In 2-5 years we may approach the lower limit for application to DOE's 'leadership class' supercomputers. We anticipate that our data production rate could increase 10-fold in 5 years.

13.4.2 Process of Science

Much of the code we use scales well to large numbers of processors. As interconnects between processors become faster, we will generate correspondingly more data.

13.5 Beyond 5 years – future needs and scientific direction

We anticipate that visualizing large data sets will remain a key need beyond 5 years. We plan to investigate GPU computing for e.g. molecular dynamics simulation and cloud computing for e.g. theoretical spectroscopy.

13.6 Outstanding Issues

Our key needs are for increases in the rate at which we can transfer and visualize data housed in local clusters and on NERSC. A dedicated transfer node or specialized visualization software are possible solutions to this problem.

13.7 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• In-house 1920-processor compute cluster	• Simulation to produce data; visualization of those data (transfer from NERSC to LBNL cluster environment, and from both sources to local workstations).	• E.g. 100 sets of 10 GB wavefunctions	• Current rates 100s of Gbps (theoretical max 1 Gbps)	• Rates from NERSC to Foundry ~500 Mbps (10 Gbps main connection, 1 Gbps local lines)
2-5 years				
• Anticipated 20-50% boost in capacity of in-house cluster capabilities in 2 years. Possible migration to larger DOE machines (e.g. Jaguar).	• Increased demand from User community for fast visualization of simulation data.	• 2-10 times current rate of data production	• Need 2-10 times current rate to cope with anticipated	• Need 2-10 times current feed rate.
5+ years				
•	•	• In excess of 10 times current rate of data production	• Need in excess of 10 times current rate	• Need in excess of 10 times current rate

14 Theoretical Chemical Physics at PNNL

14.1 Background

We conduct molecular science using supercomputers located within the DOE complex (*e.g.* Jaguar at ORNL, Franklin and Hopper at NERSC, and the BlueGene/P at ANL). We utilize the network for fast data transfer between facilities and from the facilities to my home institution at PNNL.

14.2 Key Local Science Drivers

14.2.1 Instruments and Facilities

ESnet is used at PNNL to connect to supercomputers throughout the DOE complex.

14.2.2 Process of Science

Supercomputers are used to solve equations of motion for electrons and atoms. The positions of the atoms are then analyzed and compared to available experimental data or used to predict properties. The electron wavefunctions can also be stored and analyzed in variety of different ways to make contact with experiment.

14.3 Key Remote Science Drivers

14.3.1 Instruments and Facilities

My worldwide collaborators use the Internet to connect to supercomputers throughout the DOE complex. We share disk space at the DOE facilities in order to facilitate our communication and analyze data. Data is transferred between US and international institutions and the shared disk space at DOE supercomputer facilities.

14.3.2 Process of Science

Supercomputers are used to solve equations of motion for electrons and atoms. The positions of the atoms (< 1 GB per run) are then analyzed and compared to available experimental data or used to predict properties. The electron wavefunctions (< 1 GB per wavefunction) can also be stored and analyzed in variety of different ways to make contact with experiment.

14.4 Local Science Drivers – the next 2-5 years

14.4.1 Instruments and Facilities

Continued connection to the current and next-generation supercomputers will be necessary to conduct more sophisticated simulations.

14.4.2 Process of Science

We will continue to perform more sophisticated simulations of condensed phase process (*e.g.* reactions in the condensed phase and interfaces relevant to the DOE mission of controlling matter at the atomic scale).

14.5 Remote Science Drivers – the next 2-5 years

14.5.1 Instruments and Facilities

Same as above.

14.5.2 Process of Science

Compute simulations of larger more complicated systems and processes on the DOE supercomputers

14.6 Beyond 5 years – future needs and scientific direction

Continued access to the supercomputers available within the DOE complex to perform molecular science utilizing first-principles (or *ab initio*) approaches.

The ability to transfer large quantities of data (say 100s of GB) over the Internet will become necessary, if not imperative.

Molecular simulation using first-principles methods is not usually I/O limited. System sizes range from 100-1000s of atoms, generating partial data sets of < 1GB. Storage of wavefunctions (in terms of standard cube files) can be large (e.g. 1GB per snapshot). Thus, care is usually taken to analyze these files at the local computing centers, and are generally not transferred over the WAN.

14.7 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• Supercomputers within the DOE complex	• Calculations performed remotely at supercomputer facilities. • Data analyzed either at supercomputer site or transferred to local computers.	• Data volume 1-10 GB/day • 1000 files 10MB-1000 GB	• 1 hour	• 1 hour. • Data are transferred from supercomputers to local machines
2-5 years				
• New supercomputers within the DOE complex	• None anticipated	• 1-10 GB/day • 1000 files 10MB-1000 GB	• 1 hour	• 1 hour. • Data are transferred from supercomputers to local machines
5+ years				
• Supercomputers within the DOE complex	• None anticipated	• 10-100 GB/day • 10000 files 100MB-10 GB	• 1 hour	• 1 hour. • Data are transferred from supercomputers to local machines

15 Findings

General Findings:

- The current paradigm for data processing for a Light Source is about to change. In the current model, data is copied to a portable disk and taken back to the scientists' home institution where it is analyzed. This method will not work in the future, as the data set sizes will be much to big for portable drives, and the processing requirements will be much larger. Because of this, the NSLS is building their own data center, and the other facilities will likely need to do the same thing unless the DOE decides to support general purpose mid-range compute facilities that can be shared
- Several groups reported that they found tools like GridFTP and bbcp too difficult to install and use. More documentation and training is needed, especially about the use of GridFTP using ssh keys. Documentation is available on http://fasterdata.es.net currently, but more awareness of these web pages is needed.
- Network support for remote instrument control at the BES facilities was a recurring theme at the workshop. Interdomain virtual circuits (such as OSCARS circuits provisioned over ESnet's Science Data Network) have a role to play here, since the bandwidth and service guarantees offered by virtual circuits provide the capabilities needed for remote control. However, some work will need to be done with the light sources to establish a workable operational model that includes both the facility and the end site. The NSLS is a likely collaborator with ESnet in developing this operational model.
- There is a clear need for ad-hoc video conferencing for many BES projects. Skype is currently what most groups use, and some use Caltech's EVO. Most attendees did not use and/or were not aware of ESnet's ECS service.
- Many projects still have issues with firewalls and site security policies. Remote control in particular is quite difficult to make work through a firewall. High performance data transfer tools also run into problems with firewalls. Secure Shell (ssh) and Secure Copy (scp) are installed by default on most systems, since ssh is typically used for access to systems by users. However, ssh and scp (as well as rsync over ssh) perform poorly for long-distance transfers due to inherent protocol limitations. Therefore, users find themselves caught between a poor tool that is installed by default (and that has the support of site security policies), and a high-performance tool such as GridFTP that can move data well, but is not installed by default and is not supported by site security policy.
- Many projects would benefit from a community portal. One possible solution is the NERSC Science Gateway Service (http://www.nersc.gov/nusers/services/SG/sg.php)
- Most simulation groups generate data at one of the LCF facilities, and transfer only a small subset back to their home institute.
- Several research groups are content with SCP today, but SCP will become inadequate in the future as data set sizes scale up.

- Several groups are looking into cloud computing, including using Amazon cloud services.

Findings specific to particular facilities:

- Single data transfers for NSLS-II will only be around 1Gbps, but there will be many simultaneous flows, totaling 6Gbps to 12Gbps for the facility. Most current NLSL users are from the northeast part of the US, but with NSLS-II users will come from all over the US, so it will be even more important for NSLS data servers to be tuned for wide-area access. 6-12Gbps of data throughput means at least10-20Gbps of deployed capacity.
- The Combustion Research Facility at Sandia is currently building a new 50 TF data center for smaller scale simulations. This data center is expected to serve data to many sites. This data center would benefit from a DTN.
- The peak data rate at the LCLS will be up to 1GB/sec in 5 years. Also, there will be an increase in the number of research groups transferring data from the LCLS to their home institutions over the network.
- Several ALS experiments will need greater than 10Gbps of capacity to the supercomputer centers in less than 5 years.
- The neutron scattering science community is having problems navigating the challenging technical and policy issues of international user identification and authentication. There are currently no methods for (easily) integrating user authentication systems across multiple facilities – either within the US or internationally. In the short-term, the neutron science facilities can also benefit from DTNs deployed at different facilities, which would enable faster data transfers between the facilities and help researchers conduct cross-cutting science (e.g., between neutron scattering and X-ray facilities.) For example, there is already the need to move APS data to SNS.
- There are network performance issues between Ames Lab and NERSC that are impeding scientific productivity (see section 12). Trying out new tools such as VNC, VISIT, and scp with the hpn-ssh patch may help.

16 Action Items

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

- Begin looking into need for Livermore Valley Open Campus (LVOC) site connection (this process has begun).
- Work with Ames Lab and NERSC to improve network performance.
- Work with LCLS personnel at SLAC to improve data transfer performance to DESY in Germany. Also, help LCLS personnel set up a GridFTP server for data transfers.
- Work with the NSLS to look into federation technologies such as those used by the Open Science Grid.
- Work with the Molecular Foundry at LBNL to improve network performance at the facility. One solution may be to deploy a dedicated Data Transfer Node.
- ESnet will continue to develop and update the fasterdata.es.net site as a resource for the community
- ESnet will continue to assist sites with perfSONAR deployments and will continue to assist sites with network and system performance tuning
- Work with SNS (and a collaborating X-ray facility) to deploy a DTN for use by the neutron science community for cross-cutting research

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

17 Glossary

GB/sec: Gigabytes per second – a measure of network bandwidth or data throughput

Gbps: Gigabits per second – a measure of network bandwidth or data throughput

MB/sec: Megabytes per second – a measure of network bandwidth or data throughput

Mbps: Megabits per second – a measure of network bandwidth or data throughput

PB/sec: Petabytes per second – a measure of network bandwidth or data throughput

Pbps: Petabits per second – a measure of network bandwidth or data throughput

PKI: Public Key Infrastructure

TB/sec: Terabytes per second – a measure of network bandwidth or data throughput

Tbps: Terabits per second – a measure of network bandwidth or data throughput

18 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 BES program office for their help in organizing the workshop and providing insight into the facilities supported by the BES 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, Facilities Division, and the Office of Basic Energy Sciences.

This is LBNL report LBNL-4363E