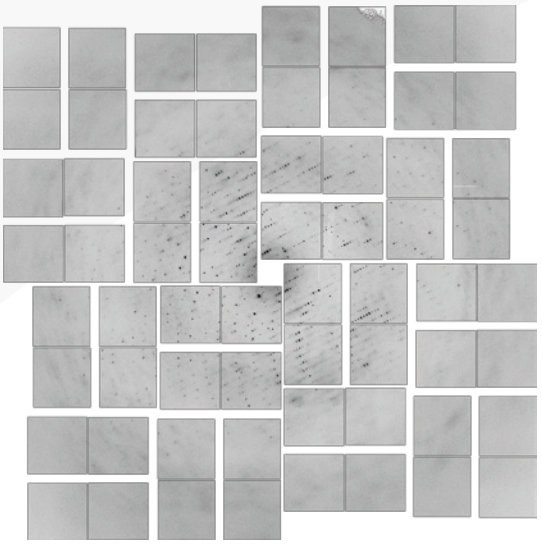
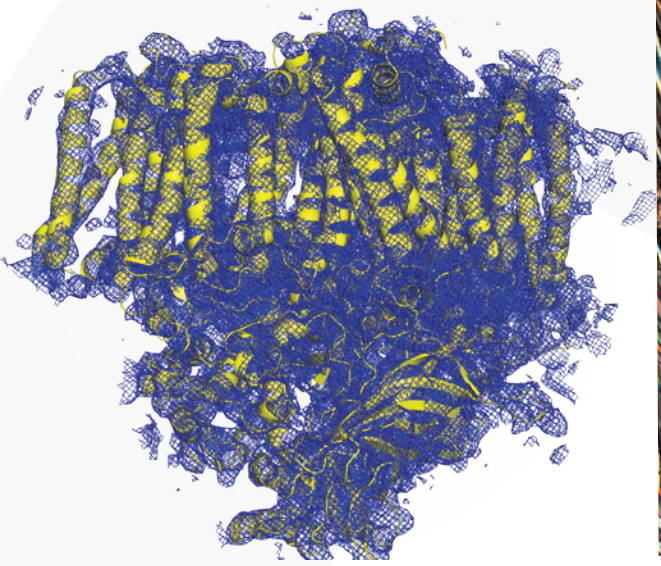
ESnet
ENERGY SCIENCES NETWORK

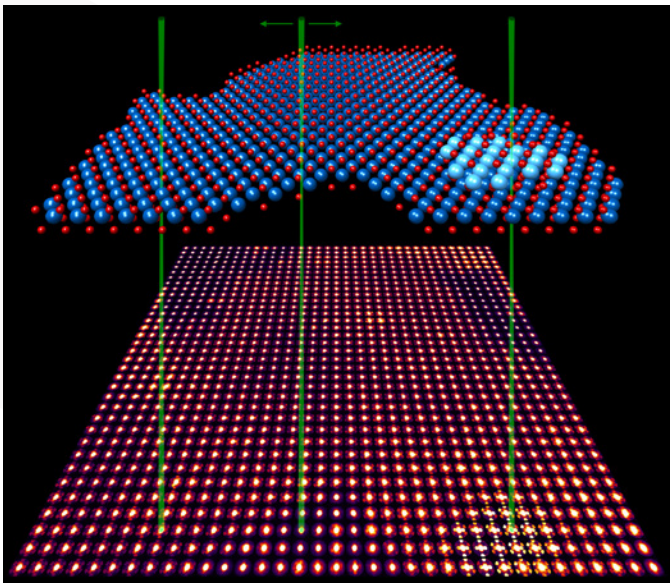
# Basic Energy Sciences Network Requirements Review

Final Report

September 9-10, 2014







BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

# Disclaimer

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

# Basic Energy Sciences Network Requirements Review Final Report

Office of Basic Energy Sciences, DOE Office of Science
Energy Sciences Network (ESnet)
Germantown, Maryland
September 9–10, 2014

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

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

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

This is LBNL report LBNL-6998E.

# Contents

# Participants and Contributors

Greg Bell, ESnet (ESnet Director)
Amber Boehnlein, SLAC (LCLS, SSRL)
Robert Dalesio, BNL (NSLS-II)
Eli Dart, ESnet (Science Engagement, Review Chair)
Vince Dattoria, DOE/SC/ASCR (ESnet Program Manager)
Jim Davenport, DOE BES Program Office (Materials Science and Engineering)
Jim Freericks, Georgetown University (Theory)
Mike Griffin, NIH (Networking)
Mary Hester, ESnet (Science Engagement)
Christopher Jacobsen, ANL (APS Facility)
Steve Lee, DOE/SC/ASCR (CAMERA Program Manager)
Eliane Lessner, DOE BES Program Office (Accelerator and Detector Research)
Raj Kettimuthu, ANL (Globus)
Thomas Ndousse-Fetter, DOE/SC/ASCR (Network Research)
Joe Oefelein, SNL (Chemistry, Combustion)
Colin Ophus, LBNL (NCEM)
Dula Parkinson, LBNL (ALS)
Mark Pederson, DOE BES Program Office (Computational and Theoretical Chemistry)
Kristen Persson, LBNL (Materials Project)
Don Preuss, NIH (NLM/NCBI)
Thomas Proffen, ORNL (SNS)
Lauren Rotman, (Science Engagement Group Lead)
James Sethian, LBNL (CAMERA)
Bobby Sumpter, ORNL (CNMS)
Craig Tull, LBNL (Spot Suite)
Priya Vashishta, USC (Materials Modeling)
Jason Zurawski, ESnet (Science Engagement)

## Report Editors

Eli Dart, ESnet: dart@es.net
Mary Hester, ESnet: mchester@es.net
Jason Zurawski, ESnet: zurawski@es.net

# Executive Summary

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the US Department of Energy (DOE) Office of Science (SC), the single largest supporter of basic research in the physical sciences in the United States. In support of 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 25 years.

In September 2014, ESnet and the Office of Basic Energy Sciences (BES), of the DOE Office of Science, organized a review to characterize the networking requirements of the programs funded by the BES program office.

Several key findings resulted from the review. Among them:

1. Workflow tools to integrate experimental facilities with computational facilities have been deployed at a number of facilities, and others are engaged in trials of the tools. The resultant "superfacilities" which are composed of an experimental facility, a computational facility, the network to interconnect them, and the workflow tools to tie it all together are expected to become increasingly important to BES-funded science programs.
2. Many collaborations continue to use portable storage media (e.g., removable hard disks) for data transport and archival storage. The most important limitation of this antiquated approach to data management is that data stored on portable media is generally not available to other research groups, which greatly limits its usefulness. This "dark data" is typically used only for a single publication by the researcher who collected the data, and represents a significant untapped resource. A mechanism for sharing this data, subject to appropriate access policies, could allow many more scientists to make effective use of experimental data. In addition, a common mechanism for sharing would offer significant performance benefits when compared to physical transport of portable media, as demonstrated by emerging superfacility workflows.
3. Many of the micrographs collected in electron microscope facilities are not published, and are not searchable. An example of "dark data," these unpublished micrographs might hold the data underlying additional discoveries, if there were a way to analyze them (e.g., for particular types of diffraction patterns such as those characteristic of quasicrystals). A data storage mechanism that allowed many micrographs from many electron microscopy facilities to be analyzed together could provide the field of materials science with significant scientific leverage and new discoveries.
4. There are two distinct workflows involved in the use of any experimental facility: a data acquisition workflow which involves running the experiment, collecting the experimental data, and near-real-time data analysis to guide the experiment; and a deep analysis workflow which involves detailed scientific analysis of the collected experiment data leading toward discovery. These two workflows are equally important, but operate on different time scales and have different requirements.
5. At many facilities, there is a mismatch between the resources spent on instruments which generate data during experiments and on tools and staff for analyzing the experimental data produced by the experiments. Several attendees felt that it would be wise to devote additional resources to software programmers or the development of analytics tools that could be used by multiple experiments.

This report expands on these points, and addresses others as well. The report contains a findings section in addition to the text of the case studies discussed during the review.

# Findings

Below are the findings for the BES and ESnet Network Requirements Review held in Germantown, MD on September 9–10, 2014. These points summarize important information gathered during the review.

- Workflow tools to integrate experimental facilities with computational facilities have been deployed at a number of facilities, and others are engaged in trials of these tools. The resulting "superfacilities" (which are composed of an experimental facility, a computational facility, the network to interconnect them, and the workflow software and tools to tie it all together) are expected to become increasingly important to BES-funded science programs.
- Many collaborations continue to use portable storage media (e.g., removable hard disks) for data transport and archival storage. The most important limitation of this antiquated approach to data management is that data stored on portable media is generally not available to other research groups, which greatly limits its usefulness. This "dark data" is typically used only for a single publication by the researcher who collected the data, and represents a significant untapped resource. A mechanism for sharing this data, subject to appropriate access policies, could allow many more scientists to make effective use of experimental data. In addition, a common mechanism for sharing would offer significant performance benefits when compared to physical transport of portable media, as demonstrated by emerging superfacility workflows.
- Many of the micrographs collected in electron microscope facilities are not published, and are not searchable. An example of "dark data," these unpublished micrographs might hold the data underlying additional discoveries, if there were a way to analyze them (e.g., for particular types of diffraction patterns such as those characteristic of quasicrystals). A data storage mechanism that allowed many micrographs from many electron microscopy facilities to be analyzed together (e.g., in statistically significant ways, in contrast to the current practice of publishing a single "representative micrograph") could provide the field of materials science with significant scientific leverage and new discoveries.
- There are two distinct workflows involved in the use of any experimental facility: a data acquisition workflow which involves running the experiment, collecting the experimental data, and near-real-time data analysis to guide the experiment; and a deep analysis workflow which involves detailed scientific analysis of the collected experiment data leading toward discovery. These two workflows are equally important, but operate on different time scales and have different requirements:
  - The data acquisition workflow is tightly coupled to the operation of the experiment, and the goal is to maximize the productivity of the available beam time. Automation is key, as time pressure makes all manual processes expensive. The collection of data provenance, data storage, transfer to real-time analysis capabilities, and analysis feedback for guiding the experiment are all important candidates for automation. Authentication and security systems, workflow management tools, and so forth cause problems if they stand in the way of automation or compromise the ease of use of the experiment environment. The data analysis that occurs during the data acquisition workflow must directly support the operation of the experiment, and provide actionable information to guide the experiment such that the best data sets are obtained. Fast turnaround for data analysis is key, because time is of the essence in the data acquisition workflow—coarse analysis sufficient to guide the experiment is far more

important than detailed analysis that takes a long time.

  - The deep analysis workflow relies heavily on the quality of the data collected during the data acquisition workflow—the beam time is over, the experiments have been conducted, and the data sets (along with the collected metadata and other provenance information) are all that remain. In contrast to the data acquisition workflow, time pressure is not the dominating factor. Rather, the best tools are used for meticulous analysis of the experimental data, and the goal is a discovery fit for publication. The deep analysis workflow relies on getting the data to the places where the right analysis tools can be run—if the computation can be moved to the data then this is done, and if the data must be moved to the computation, that is done.

- At many facilities, there is a mismatch between the resources spent on instruments which generate data during experiments and on tools and staff for analyzing the experimental data produced by the experiments. Several attendees felt that it would be wise to devote additional resources to software programmers or the development of analytics tools that could be used by multiple experiments.
- There are many aspects of an experiment at a light source that contribute to the data rates (both peak and average) that result from that experiment. The detector capabilities are one component, but there are many others—sample change time and exposure time often limit the amount of data that can be collected, and the local networking infrastructrure (both the LAN and the path to the wide area) can be a factor as well. In many cases, increasing the network interface speed will not have any effect on data rates or on scientific productivity, because the limiting factor is elsewhere in the experiment workflow. However, there are several cases where straightforward workflow improvements will result in a significant increase in the volume of data produced and the rate at which the data are produced. In many cases, these improvements would increase data rates by a factor of 5 or more.
- Multiple facilities and research groups mentioned improvements in performance and scientific productivity after deploying modern data management capabilities such as Data Transfer Nodes and Globus.
- Several attendees discussed the advantages of common tools for the same analysis processes at multiple facilities. A suggestion was made for multiple facilities to each donate a person to an effort focused on building tools for common tasks that could be deployed and supported at multiple facilities. These might be integrated into “superfacility” workflows.
- There was significant discussion at the review about data policies, and the uncertainty surrounding the long-term responsibility for data archival and the costs associated with it. The DOE program managers indicated a preference for the facilities to self-organize around a set of solutions for data rather than wait for an edict from headquarters.
- The use of Globus rather than physical transport of portable media for data transfers at ALS beamline 8.3.2 has significantly reduced the difficulties associated with data transfer for the users that are able to use Globus. These benefits accrue both to the users and to the beamline support staff.
- Data Transfer Nodes and Globus have been very helpful for data management at NCEM. It is expected that this would be helpful for other microscopy centers as well, especially for handling the dramatically larger data sets produced by state-of-the-art cameras (e.g. 16TB in 15 minutes at NCEM). In the coming years, NCEM plans to move much of its data management and data analysis workload to large HPC facilities (e.g. NERSC).
- A significant amount of data is transferred from the LCLS to a group at DESY in Germany. Historically, data transfer rates had been too slow, but the deployment of Globus has improved things (they are now able to transfer 8TB in 2 days). While these data rates are not as fast as they could be, it appears that data transfers are no longer a matter of concern for this group.
- Representatives from several facilities expressed an interest in collaborating together and with ESnet on the development of materials to give users a background in data transfer tools. This “network on-boarding” process has the potential to significantly improve the use of beam time by training users in the use of modern data transfer and data management tools.

- SLAC and NERSC are collaborating on the use of large-scale computing at NERSC in support of the experiments done at the LCLS. This is likely to drive increased network usage between these facilities in the future.
- Long-term archival storage is a concern for some facilities. Several options are being explored, including the use of cloud storage (e.g., Amazon Glacier).
- Federated identification for facility users would solve several problems at multiple facilities, including data access, access management, and the reduction of IT costs associated with supporting users.
- Software tools (e.g., workflow or analysis packages) that provide "information rather than data" to scientists are of significant value. However, these software tools often face challenges for sustainable maintenance after the development phase. This is currently an unsolved problem.
- At the Combustion Research Facility, data transfers scale according to the resources involved—large data sets are transferred from the national computing facilities to the local clusters at the CRF, and smaller data sets are transferred to home institutions or laptops for further analysis.
- There is a synergistic relationship between combustion simulations and experiments. Experiments can be used to extend the capabilities of a model (e.g., provide experimental results for an area in which a model is weak or unstable), and the resultant improved model then shows where the next experiments are needed.
- There are common users between the light and neutron sources—for example, 7% of APS users also use the SNS. Many scientists find it useful to get both an X-ray and a neutron characterization of a sample material. NOMAD proposals have a reciprocity agreement with the APS for those cases where X-ray work would provide additional benefit to a neutron experiment.
- The data analysis code or experiment support code required for a particular experiment may not run well (or run at all) on the computing resource available at the experimental facility. In these circumstances, it is necessary to move the data to the computation because the computation cannot be effectively moved to the data.
- It is very useful to have both the simulations and experiments associated with a particular scientific project produce and consume the same file formats. The Mantid framework does this, and some tools use the Nexus file format for both simulation and experiment.
- The neutron sources are unlikely to see the same level of increase in data rates and volumes as the light sources are expected to see. A factor of two or three increase is expected, but not an order of magnitude or two.

# Action Items

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

- ESnet will work with the Reactive Molecular Dynamics Simulations of Materials group at the University of Southern California to get a Globus data transfer node deployed there.
- ESnet will work with the Combustion Research Facility on data transfers with collaborating institutions, including the exploration of Globus.
- ESnet will work with the APS and other light sources to help identify major users from institutions that have recieved CC-NIE or CC*IIE grants from the NSF.
- ESnet will include the SNS in the development of “data onboarding” materials for facility users.
- ESnet will work with the NSLS-II facility to organize a site visit by ESnet staff for further discussion of data and science engagement topics.
- ESnet will work with scientists at Georgetown University on data transfer performance to national and international collaborators.
- ESnet will continue to develop and update the http://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.
- ESnet will continue to support the development and deployment of perfSONAR.
- ESnet will continue to support the development and deployment of the ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS) to support virtual circuit services on the ESnet network.

# ESnet SC Requirements Review Background and Structure

Funded by the Office of Advanced Scientific Computing Research (ASCR) Facilities Division, ESnet's mission is to operate and maintain a network dedicated to accelerating science discovery. ESnet's mission covers three areas:

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

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

*(a)* A review of major stakeholders' plans and processes, including the data characteristics of scientific instruments and facilities, in order to investigate what data will be generated by instruments and supercomputers coming online over the next 5–10 years. In addition, the future process of science must be examined: How and where will the new data be analyzed and used? How will the process of doing science change over the next 5–10 years?

*(b)* Observing current and historical network traffic patterns to determine how trends in network patterns predict future network needs.

The primary mechanism for accomplishing (a) is the SC Network Requirements Reviews, sponsored by ASCR and organized by the SC Program Offices. SC conducts two requirements reviews per year, in a cycle that repeats every three years:

The primary mechanism to accomplish (a) is through SC Network Requirements Reviews, which are organized by ASCR in collaboration with the SC Program Offices. SC conducts two requirements reviews per year, in a cycle that assesses requirements for each of the six program offices every three years.

The review reports are published at http://www.es.net/requirements/. The other role of requirements reviews is to help ensure that ESnet and ASCR have a common understanding of the issues that face ESnet and the solutions that it undertakes.

In August 2014, ESnet organized a review in collaboration with the BES Program Office to characterize the networking requirements of science programs funded by Basic Energy Sciences.

Participants were asked to codify their requirements in a case study format that included a network-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the network services needed; and how the network is used. Participants considered three timescales in their case studies: the near-term (immediately and up to two years in the future); the medium-term (two to five years in the future); and the long-term (greater than five years in the future).

The information in each narrative was distilled into a summary table, with rows for each timescale and columns for network bandwidth and services requirements. The case study documents are included in this report.

Participants were asked to codify their requirements in a case-study format that included a network-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the network services needed; and how the network is used. Participants considered three timescales in their case studies: the near-term (immediately and up to two years in the future); the medium-term (two to five years in the future); and the long-term (greater than five years in the future).

More specifically, the structure of a case study is as follows:

- Background—an overview description of the site, facility, or collaboration described in the case study
- Collaborators—a list or description of key collaborators for the science described in the case study (the list need not be exhaustive)
- Near-term Local Science Drivers—a discussion of science drivers that have network implications for the local infrastructure at the primary site or institution of the case study. The time period for "near-term" is 0-2 years. A good way to think about the "near-term" time scale is as the current set of circumstances, along with the current two-year budget time horizon. This section has two subsections—instruments and facilities, and process of science.
  - Instruments and Facilities—this describes the "hardware" of the science of the case study. Instruments and facilities might include detectors, microscopes, supercomputers, telescopes, fusion reactors, or particle accelerators. The instruments and facilities view of the case study provides the information about data rates, data volumes, location of data, origin of data, and so forth.
  - Process of Science—this section describes the ways in which scientists use the instruments and facilities for knowledge discovery. The process of science section captures aspects of data flow, instrument duty cycle, data analysis, workflow, and so forth.
  - Software Infrastructure—this section describes the software used to manage the daily activities of the scientific process in the local environment. It also includes tools that are used to locally manage data resources, facilitate the transfer of data sets from or to remote collaborators, or process the raw results into final and intermediate formats.
- Near-term Remote Science Drivers—a discussion of science drivers that are not local to the primary site or institution of the case study. These typicall involve the use of the wide area network for some purpose (data transfer, remote control, remote access, etc). The time period for "near-term" is 0-2 years, as above. This section has two subsections—instruments and facilities, and process of science.
  - Instruments and Facilities—this describes the "hardware" of the science of the case study as above, except from a non-local perspective. Examples might include the use of remote HPC resources, or a particle accelerator at another facility.
  - Process of Science—this section describes the process of science as it pertains to the use of remote instruments and facilities.
  - Software Infrastructure—this section describes the software used to manage the daily activities of the scientific process in the wide area environment. It includes tools that are used to manage data resources for the collaboration as a whole, facilitate the transfer of data sets to or from remote collaborators, or process the raw results into final and intermediate formats. The objective is to capture the software tools that move data over the network.
- Medium-term Local Science Drivers—similar to near-term local science drivers, but with a time horizon of 2–5 years in the future. A good way to think of this is that the medium-term view in-

corporates the current technological paradigm or facility environment, rather than just the current budget horizon.

  - Instruments and Facilities—local instruments and facilities, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
  - Process of Science—local process of science, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
  - Software Infrastructure—local software infrastructure, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?

- Medium-term Remote Science Drivers—similar to near-term remote science drivers, but with a time horizon of 2–5 years in the future.
  - Instruments and Facilities—remote instruments and facilities, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
  - Process of Science—remote process of science, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
  - Software Infrastructure—remote software infrastructure, with a 2–5 year time horizon. Specifically, what will change in the next 2–5 years?
- Beyond Five Years—this describes the strategic planning horizon, including new facilities, major technological changes, changes in collaboration structure or composition, etc.
- Network and Data Architecture—this section describes the use of specific networking resources (e.g., a Science DMZ) and how those resources are structured in the context of the science. In addition, if changes would significantly impact the science, they can be captured here.
- Data, Workflow, Middleware Tools and Services—anything not captured in the software infrastructure section can be captured here.
- Outstanding Issues—if there are current problems that should be brought to ESnet's attention, they are captured here.

The information in each narrative was distilled into a summary table, with rows for each timescale and columns for network bandwidth and services requirements. The case study documents are included in this report.

# Office of Basic Energy Sciences Overview

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 nanomaterials 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, multiinvestigator 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.

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 were reported in a final report found on the ESnet website: http://es.net/assets/Uploads/BES-Net-Req-Workshop-2010-Final-Report.pdf. Since then, and

as highlighted throughout the following pages, the data sets generated and used by BES scientists continued to increase in size and Internet access to these data sets continued to be a vital requirement that is fulfilled by ESnet. In order to assess the increasing challenges and needs, a new workshop was organized ands the results and studies are presented here. 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.

# Case Studies

# Case Study 1

# Advanced Light Source

## 1.1 Background

Operated since 1993, the Advanced Light Source (ALS) is a third-generation synchrotron and National User Facility located at the Lawrence Berkeley National Laboratory (LBNL). The ALS operates about 40 beamlines simultaneously with an approximate 60% operational duty cycle, operating over 5000 hours per year. It produces some of the brightest ultraviolet and soft X-ray beams in the world, as well as intense hard X-rays and infrared (IR) beams. It has about 200 staff, an annual operating budged of around $60 million, and an output of more than 800 refereed publications per year.

## 1.2 Collaborators

The ALS hosts over 2,000 distinct users annually from around the world and from diverse fields of science (see Figure 1.1).

The ALS itself runs the majority of its beamlines, but a number of other organizations either run beamlines or are significant participants as users or leaders. These include:

- Berkeley Center for Structural Biology
- Center for X-ray Optics
- LBNL Chemical Sciences Division

## 1.3 Near-term Local Science Drivers

### 1.3.1 Instruments and Facilities

At each beamline, X-rays, UV, or IR light are directed at a sample, and some kind of signal is detected. Signals include transmission, reflectance, fluorescence, scattering, etc. Detectors can be either a single element or array detectors.

Detectors are connected to beamline control computers. A majority of control computers at the ALS are Windows systems. In some cases, the beamline control computer both triggers data acquisition by the detector and then receives the image or signal data from the detector. In other cases, the control computer triggers data acquisition and a separate, dedicated computer or server receives the detector data. In either case, data is usually transferred to another storage server or device, most often located

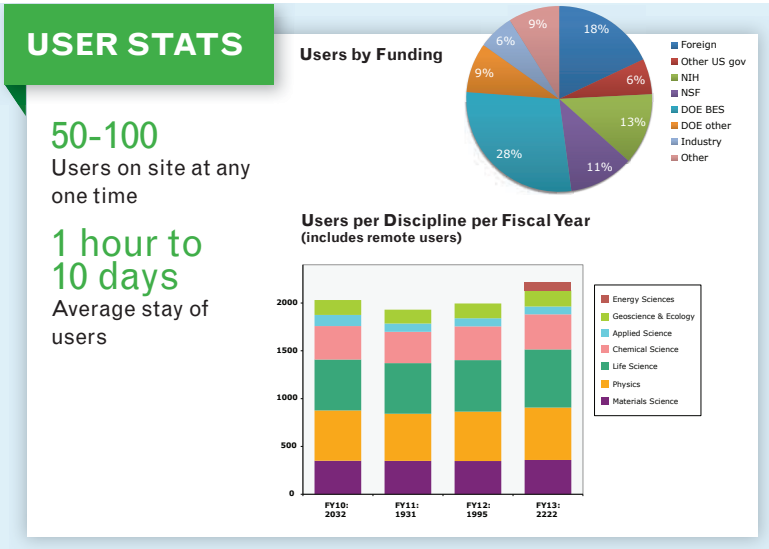



*Figure 1.1: ALS user statistics.*

directly at the beamline. In the new ALS User Support Building completed in 2010, which is adjacent to (and connected to) the ALS, there is a server room. Some beamlines have moved storage and processing servers to that room. In general, storage servers are not shared across multiple beamlines—each beamline has its own storage system. Users of a beamline usually copy their data from the storage server to portable media, usually USB media (though in some cases beamlines do not have a separate storage server and users access data directly from the control computer).

Most beamlines have some kind of data processing capability either at their beamline or in the ALS User Support Building server room. In some cases the processing is done at workstations at the beamline with access to the beamline's storage server.

The ALS also has a new Visualization and Analysis Laboratory that has six high-end desktop workstations and has been connected to the storage servers of multiple beamlines. In some cases, beamlines have small CPU and/or GPU computing clusters at the beamline or in the server room. These clusters are beamline specific rather than shared across beamlines. The protein crystallography beamlines each have compute clusters, as do the microdiffraction, small-angle X-ray scattering (SAXS) to wide-angle X-ray scattering (WAXS), and ptychography beamlines.

Results from a survey given to ALS beamline scientists were compiled into an interactive data workflow model as depicted in Figure 1.2.[1]

In this figure, a column represents each step in the data flow, with the height of the column being proportional to the square root of the transfer rate. By compiling results from a survey into this tool, the website creates a model for data production at the ALS, allowing characterization of the limiting factors in the data rates at the various beamlines. It also enables forecasting of data rates under different upgrade scenarios.

A number of simplifying assumptions were made in constructing this data flow model. One important assumption is that all raw data is immediately transferred a single time from one step to the next; data processing that either reduces or enlarges the amount of data is ignored, and buffering of data between steps is not considered. Of course, most beamlines also run in multiple modes, and except for a few cases, this model does not capture those different modes and just uses one representative mode. The assumptions for this model are meant to create a simple model for insights into ALS beamline data flows.

[1] An interactive version of this display is available at: http://bl832web.lbl.gov/esnet where a CSV version of the underlying data is also available. A google doc with the underlying data is also available at: https://docs.google.com/spreadsheets/d/1tJgtz2PgbOL_PRf0LSHEiUIvOPGBuF9cAVRX-Y05dDY/edit?usp=sharing.

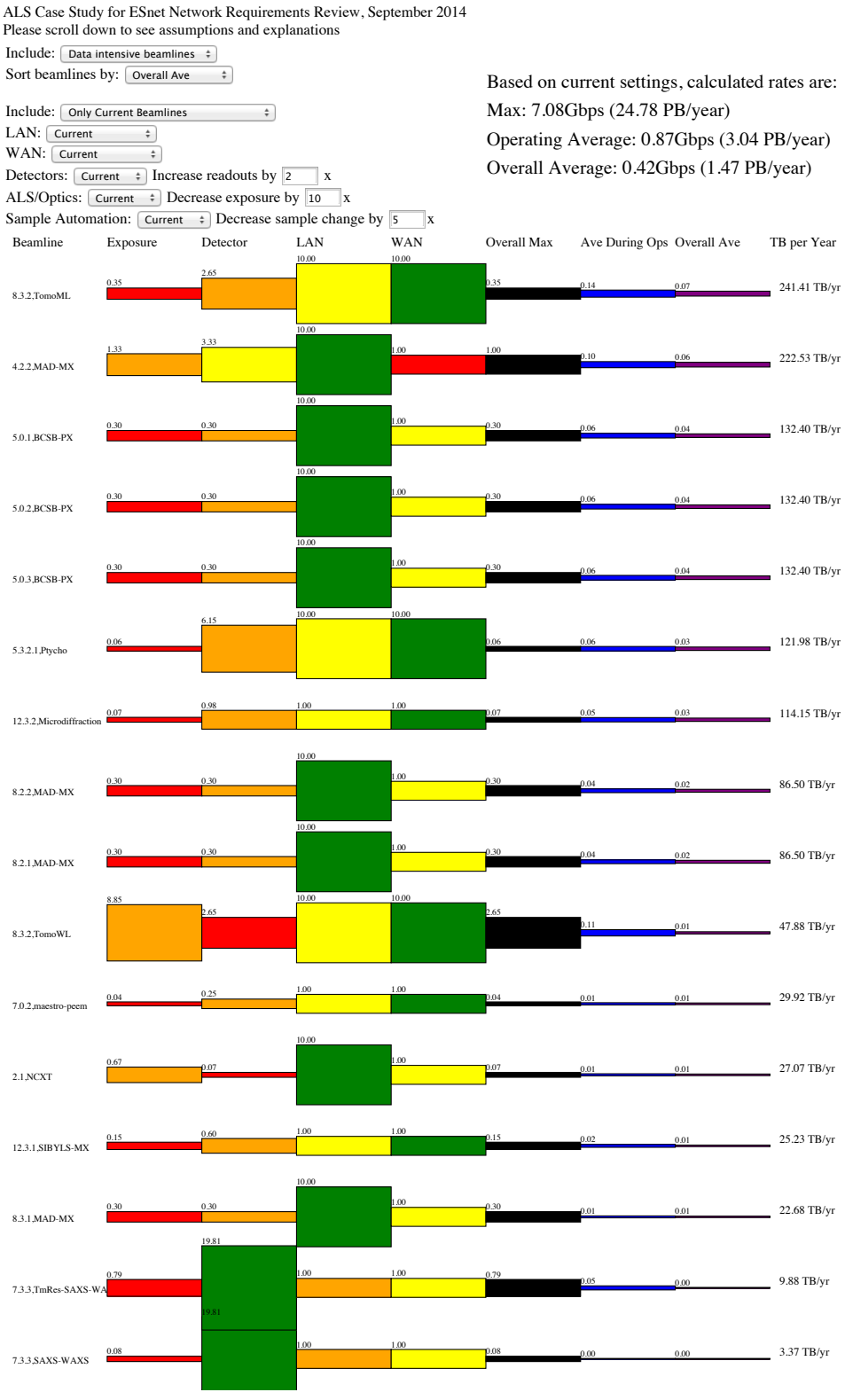



*Figure 1.2: Screen shot of survey results displayed at http://bl832web.lbl.gov/esnet. Interface shows the ALS beamline name, exposure, dector, LAN, WAN, overall maximum, average during operations, overall average, and total data rate per year. Color coding of the first four columns highlight the limiting factors for data rates at each beamline: from highest to lowest data rates the colors are green, yellow, orange, and red.*

The columns in Figure 1.2 from left to right represent:

- **Exposure**: data rate as limited by "exposure time." To determine this column, we assume the current detector (in number of pixels and bits/pixel) but with infinite readout speed. Most beamline take scans of a single sample, either moving the sample or changing some beamline or sample parameter during a scan. In many cases, actual exposure time (time during which the detector accumulates counts) is the main contributor to limiting this data rate. In other cases, a beamline also has significant contributions in this column from the time required to move beamline or sample motors between successive shots during a scan. This column does NOT take into account time to change between samples.
- **Detector**: data rate as limited by detector readout to the acquisition computer or camera server, assuming the shortest possible exposure time allowed by the detector.
- **Local-area network (LAN)**: data rate as limited by connection from the acquisition computer or camera server to the storage server.
- **Wide-area network (WAN)**: data rate as limited by connection from the storage server to the WAN.
- **Overall Max**: maximum data rate as limited by the smallest of the preceding items.
- **Ave During Ops**: average data rate during operation, given by the Overall Max multiplied by a beamline-specific duty cycle for beamline setup and sample changing.
- **Overall Ave**: average data rate, given by the Ave During Ops multiplied by duty cycles for shared beamlines (variable) and for ALS light available ( 60%).

From the model, it can be deduced that the factor that most limits the overall maximum data rate of a beamline is shown in red. For the majority of beamlines though, the Exposure column is the limiting factor, and the Detector is the second-most common limiting factor.

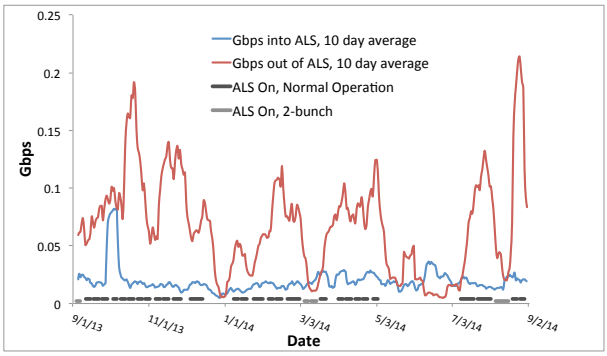
In collecting data for this chart, beamline scientists were asked about their annual data production rate, and values for sample scanning, changing, and beamline setup were adjusted to yield that value within a factor of 3. In the top right hand corner of the chart, the ALS-wide maximum and average data rates are given. This indicates that the current rate of data production at the ALS (given all the assumptions listed above, and to within a factor of 3) is 1.5 PB/year. Note that this number is significantly higher than the expected ALS outbound network traffic. See Section 1.4 for further discussion.

### 1.3.2 Software Infrastructure

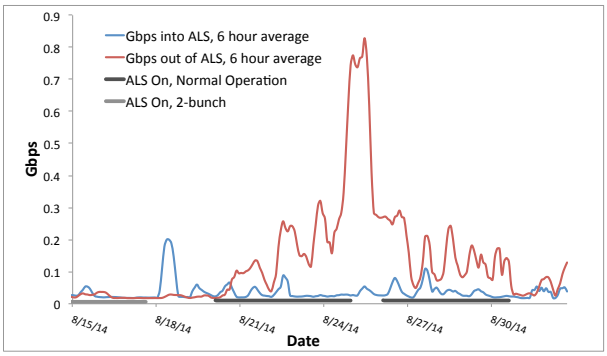
A majority of the beamline control computers at the ALS run a LabVIEW-based control system. The ALS Controls group writes and maintains the core software, and they then customize it for each beamline. In many cases, the customization is done in collaboration with the staff of each beamline. In terms of the software interface for data acquisition from the detector, this is sometimes done directly through LabVIEW, but a number of beamlines also use commercial software that accompanies their detector for data collection.

Most beamlines have data processing capabilities on workstations located at the beamline or available in the ALS Visualization and Analysis Lab. Users run a huge variety of software even at one beamline, making standardization and automation a challenge. In most cases, users manually run each processing step on their data. The protein crystallography beamlines have the most established and advanced capabilities in the area of automated data processing. The ptychography beamline is developing this capability, with automated software that runs on a GPU cluster in the ALS server room and provides real-time image feedback as data is collected. The microdiffraction beamline automatically runs a cluster version of the X-ray Microdiffraction Analysis Software (XMAS) that has been developed at the ALS.

Very few beamlines provide any software for data management; in many cases "data management" just means leaving files on a server until it is full then manually purging old data.




*(a)*




*(b)*

*Figure 1.3: The figures display network traffic in and out of the ALS, as well as whether the ALS is on or off ("on" in this case refers to user operations mode during which beamlines are collecting data), and if on, whether it is running under normal operation or in "2-bunch" mode. (a) shows network traffic to and from the ALS over the past year. (b) shows network traffic over a smaller date range, in August 2014. An overlay shows when the ALS is in normal operation, 2 bunch, or unavailable for user operations.*

A new framework called SPOT Suite, developed as part of a collaboration between the ALS, the LBNL Computational Research Division, NERSC, and ESnet, has been deployed at three beamlines (SAXS-WAXS, microdiffraction, and tomography). SPOT Suite provides both data management, data transfer, and data processing capabilities. Under this framework, data is automatically packaged with meta-data and sent to a National Energy Research Scientific Computing Center (NERSC) supercomputer for archiving and processing. (See Section 1.4 for further discussion of SPOT SUITE.)

### 1.3.3 Process of Science

The local and remote aspects of the process of science at the ALS are described in Section 1.4.3.

## 1.4 Near-term Remote Science Drivers

### 1.4.1 Instruments and Facilities

Most beamlines are connected to the wide area network (WAN) through 1 Gbps routers, which in turn are connected though the ALS 10 Gbps WAN connection. A few beamlines are directly connected to the 10 Gbps WAN (see online version of Figure 1.2).

ESnet has tracked overall network traffic in and out of the ALS since August 2013. Figure 1.3a shows this traffic over the year from 1 September 2013 to 1 September 2014, and Figure 1.3b shows traffic

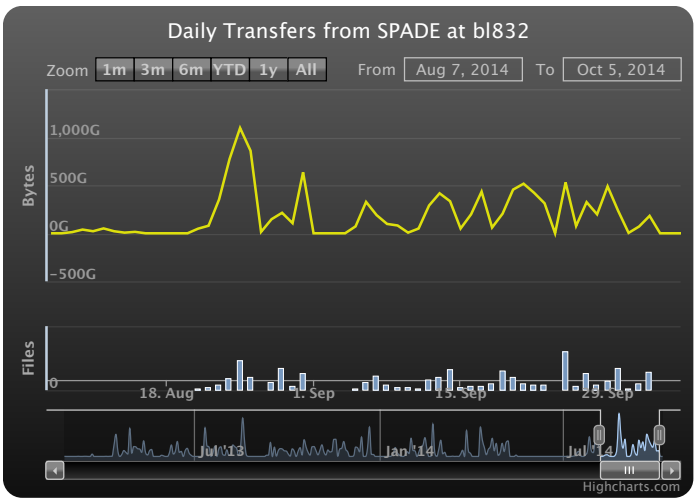



*Figure 1.4: Data transfers from the Beamline 8.3.2 (hard x-ray microtomography) Data Transfer Node over the same time period as displayed in Figure 1.2. Chart taken from spot.nersc.gov, the SPOT Suite data portal hosted at NERSC.*

during a recent two-week period in 2014.

The network traffic out of the ALS can be attributed to transfer of data collected at the ALS. Although a majority of users still copy data using portable devices, an increasing number of users are transferring data to their home institutions over the WAN. This conclusion is supported by Figure 1.3a, which shows a decrease in network traffic outbound during windows of time when the ALS is shutdown. Note that the traffic in this figure is shown with a five-day average, so that one- or two-day shutdowns do not correlate with large changes in network activity.

Figure 1.3b shows a surprising increase in network traffic during a two-day shutdown (around August 25). This can be explained by turning to the data transfer record for Beamline 8.3.2 (hard X-ray microtomography), which is shown in Figure 4 for the same time period noted in Figure 1.3b. Normally, as data sets are collected at that data-intensive beamline, they are packaged by SPOT Suite (see Section 1.3.2) on the beamline's data transfer node and sent to NERSC; this transfer is an important contributor to the WAN traffic out of the ALS. The data transfer node had a problem between August 20 and 23 that caused a number of collected data sets to not be transferred. This problem was fixed on August 25, and a backlog of data sets was then transferred on a day in which the ALS was off.

In general, NERSC is the largest recipient of outbound ALS network traffic, though certainly not the only recipient. For example, the ESnet Arbor system, [2] in Figure 1.5, shows the bulk of the network traffic out of the ALS on the August 25 was to NERSC.

From Figure 1.3, the network traffic out of the ALS does not appear to have increased in a measureable way over the past year. On the other hand, historical data about network traffic out of the ALS dating back to 2008 is available through LBLnet's RRDTool, which shows that the network traffic has increased. The increase in ALS outbound traffic seems to be in line with the overall increase in ESnet traffic, which increases by a factor of 10 about every 4 years.

Staff at the ALS envision some possible scenarios under which there could be a large increase in network traffic, but in the short term, the facility is expected to continue to produce the current, more modest trend of increase in outbound network traffic.

[2] Results from the ESnet Arbor system are available from the My ESnet Portal at https://my.es.net/facility/als/#flow/t=30d&s=site.

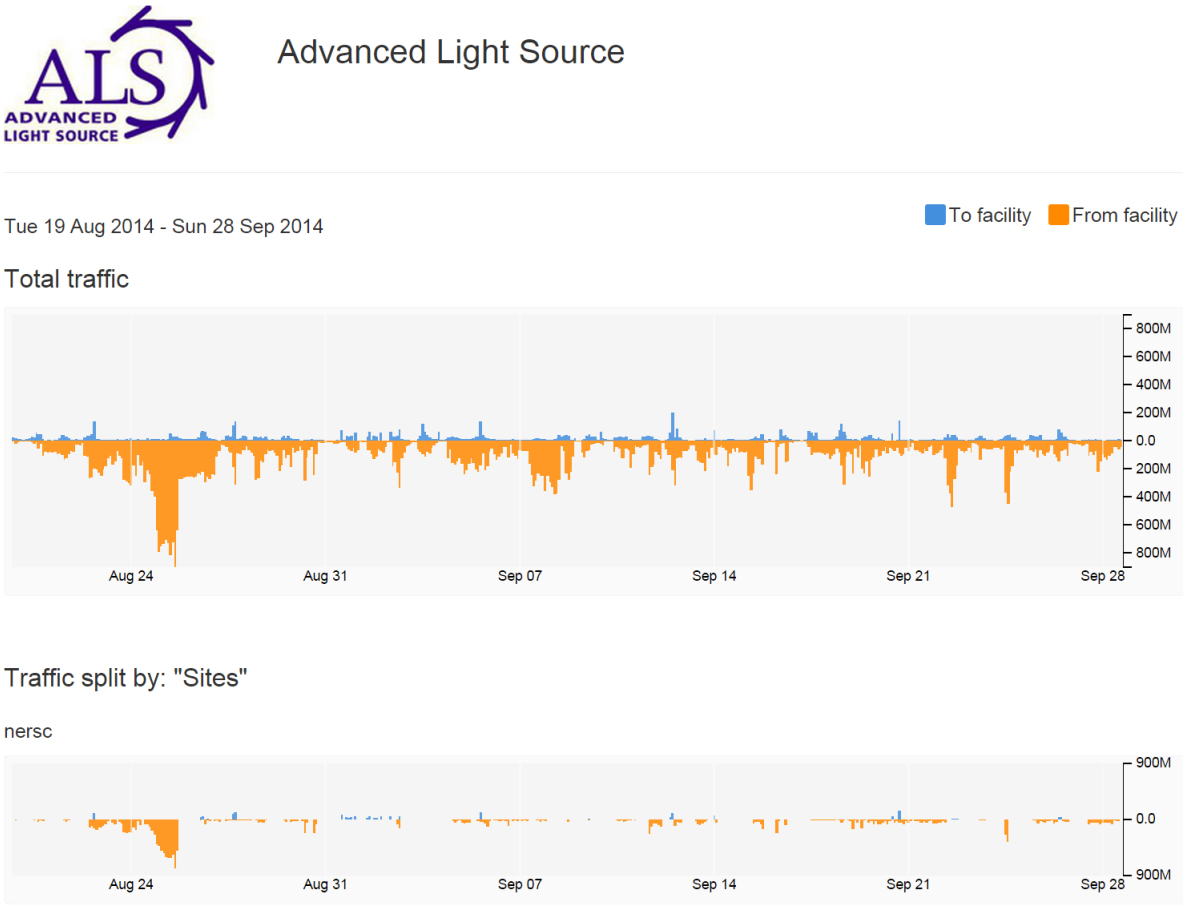



*Figure 1.5: Screenshot from https://my.es.net/facility/als/#flow/t=30d&s=site which gives visualizations of data collected by the ESnet Arbor system.*

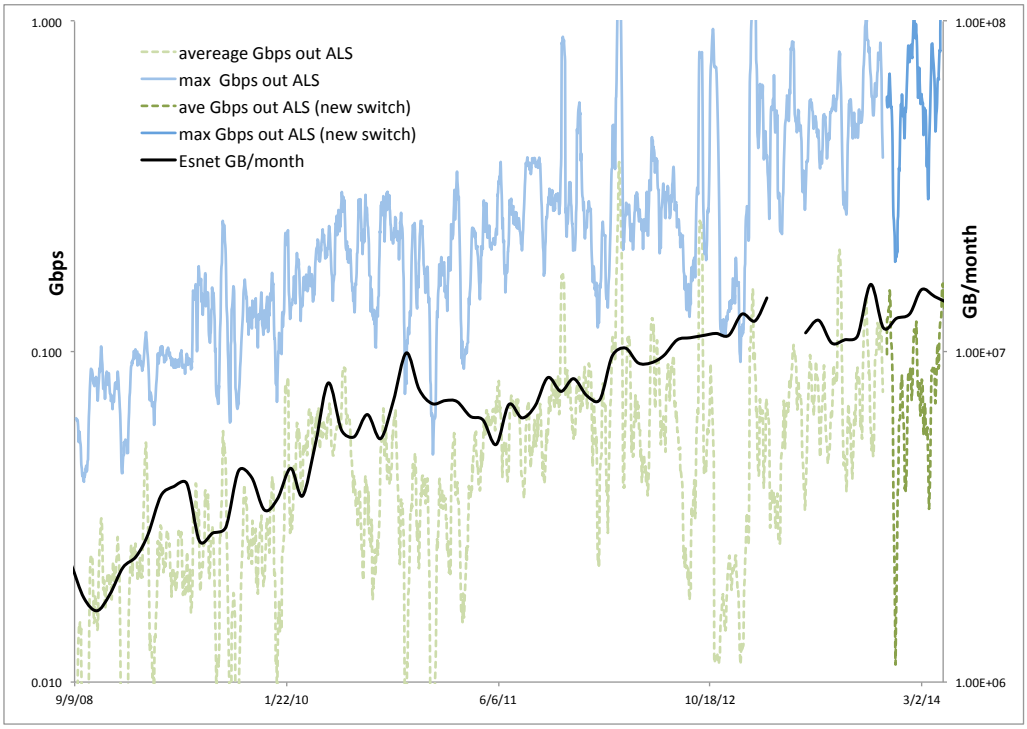



*Figure 1.6: Network traffic out of the ALS, overlayed with overall ESnet traffic.*

### 1.4.2 Software Infrastructure

Raw images are captured at the ALS along with beamline and user-generated metadata that define experimental parameters. Data transfer nodes (DTNs) at the ALS, package the raw image files (typically in tif format) into HDF5 format files where each image is stored as a 2D dataset. The user and system metadata are attached to the HDF5 file as attributes at the top level, as well as at the dataset level for image specific metadata. The packaged HDF5 is automatically transferred to NERSC DTNs via ESnet using the SPADE (South Pole Archival and Data Exchange) data transfer package and GridFTP (a high-performance, secure, reliable data transfer protocol optimized for high-bandwidth wide-area networks). The files are archived to the HPSS (High Performance Storage System) tape system and staged onto NERSC's GPFS-based (General Parallel File System) global project file system which is accessible to both data transfer nodes and HPC compute machines.

The HDF5 metadata and hierarchy are registered in a MongoDB database (an open-source document database, the term Mongo derives from "huMONGOus") at NERSC. There is a one-to-one correspondence between the HDF5 metadata and the contents of the Mongo database. Analysis on NERSC HPC computer resources is then automatically initiated. The analysis is broken down into as many as 30 individual steps; concurrency and dependency is managed in a custom workflow engine described below. Computationally intensive analysis steps are typically run in a data-parallel mode on up to 100 cores. Raw and analyzed data are available to users remotely via a web-portal built on top of NERSC's NEWT API (Nice Easy Web Toolkit, Application Programming Interface) as well as custom data APIs and interfaces built with MongoDB, Django, PHP, JQuery, Bootstrap, HTML5 and WebGL. Users can access, download and visualize their data without requiring advanced software expertise.

Data managed and transferred by SPOT is now a major contributor to the outbound traffic of the ALS. The SPOT framework is a prototype software solution for both LAN and WAN data management and processing that could be deployed across most beamlines at the ALS. If SPOT Suite spreads to more beamlines, or to the entire ALS, there could be an enormous increase in network traffic out of the ALS, since significant amounts of data that have not previously been transferred over WAN will begin to be transferred that way.

While many beamlines still do not make their data available over the WAN, an increasing number of beamlines are transitioning to this model. The majority of beamlines that make their data available over the WAN use software such as ftp or sftp. One beamline (8.3.2, a hard x-ray tomography beamline) has its storage server set up as a Globus endpoint.

Visualization of large data sets, and in particular 3D data sets, is an enormous challenge in general. Implementing remote visualization capabilities will be an important development that lets users without other access to good computational resources take advantage of remote resources. But it is not clear that this will contribute significantly to network traffic.

A number of beamlines are now allowing remote beamline operation access. In general, this is accomplished using NX, VNC, or Remote Desktop, which require relatively low bandwidth. There are no special video conferencing needs beyond those of other organizations.

To take advantage of the increasing speed of data collection, there is an increasingly urgent need for tools to manage data workflows.

### 1.4.3 Process of Science for Local and Remote Science

As noted in the discussion of Figure 1.2, data transers play a large part in the local and remote processes of sceince at the LAS. And, related to our simple model, there are three more questions underlying the assumptions from the simple data flow model:

- Where does data processing happen?
- Which data is saved and/or transferred from one location to another?
- When does data processing happen?

Here, by "data processing" we simply mean any operation on the data. For example, this could be image compression, it could be peak finding, or it could be tomographic reconstruction.

*Where*

Data processing can happen at the detector, on the beamline control computer, on processing clusters at the beamline or ALS server room, or offsite. In some cases, processing greatly reduces the quantity of data that is carried to the next step. This is the case, for example, in protein crystallography, microdiffraction, and ptychography. In others cases, such as tomography, data processing produces results that are even larger than the original data sets.

*Which*

In cases where a data processing pipeline has been refined, users do not necessarily save all raw data, meaning that only the results are stored and transferred over the network. But in most cases at the ALS, data processing pipelines are still in development, and users prefer to keep both raw and processed data in case a new and improved algorithm becomes available allowing additional extraction of information from their data.

*When*

Data processing is sometimes required in "real time," meaning that results of the processing must be produced and presented to the user during their beamtime so that they can make decisions about how to proceed with their experiment. "Real-time" for the ALS is usually on the time scale of minutes, since that is the time between scans for many beamlines. In some cases, such as the new ptychography beamline, "real-time" results based on processing on a GPU cluster are expected within a matter of seconds to guide users during sample setup. Most often, however, only a minimal amount of processing is really required in real time to guide the experiment, while a significant amount of the processing can take place in the days or weeks following beamtime.

The location at which processing is done, which data are saved and transferred, and the time window in which results are expected all have an enormous impact on the ALS network requirements. The

*Table 1.1: Predicted data rates for different scenarios, under the assumptions that are described for the model in Figure 1.2.*

| Which Beamlines | Upgrade Scenario | Max Gbps | Operating Ave. Gbps | Overall Ave. Gbps |
|---|---|---|---|---|
| Current | Current | 7.24 | 0.91 | 0.43 |
| Current | 10G LAN/WAN | 7.57 | 0.92 | 0.44 |
| Current | Detectors 5x | 13.94 | 0.95 | 0.44 |
| Current | Exposure 5x | 10.37 | 0.95 | 0.44 |
| Current | Sample 5x | 7.24 | 1.59 | 0.68 |
| Current | All 5x+10G LAN/WAN | 34.56 | 2.98 | 1.31 |
| Current+New | Current | 25.84 | 3.22 | 1.81 |
| Current+New | All 5x + 10G LAN/WAN | 81.58 | 8.43 | 4.68 |

most network-intensive scenario is if all processing is done off-site, and results are needed in real time. In this scenario, sufficient bandwidth is required to handle the maximum data rate from all beamlines simultaneously. A more realistic scenario is that some data reduction is performed on-site; a subset of collected data is processed offsite; and only a subset of the processed results are needed for real-time feedback, meaning that the required bandwidth is closer to the average data rate of the ALS after accounting for sample change and ALS duty cycles.

We note that the process of science is not the same across beamlines, which are generally independent of each other and operate with different approaches.

## 1.5 Medium-term Local Science Drivers

### 1.5.1 Instruments and Facilities, Software Infrastructure, and Process of Science

There has been little increase in network traffic out of the ALS over the past few years. In contrast, based on interviews with beamline scientists, the amount of data being produced by the ALS is increasing rapidly. There is clearly a disconnect between the data production rate of the ALS and the network traffic out of the ALS.

SPOT Suite has the greatest potential to impact network traffic out of the ALS because it would suddenly couple ALS data production to network traffic. The web visualization referenced in Figure 1.2 allows different scenarios to be explored, within the context of the assumptions in the model—this would be a simplistic implementation of SPOT Suite across all beamlines at the ALS, where all data would be transferred to NERSC for processing.

The current maximum, operating average, and overall average data rates are 7.24, 0.91, and 0.43 Gbps respectively. Although most of the ALS is on 1Gbps connections, upgrading all LAN and WAN connections in the ALS to 10Gbps would have only a tiny effect on network traffic because these are not the limiting factors. Upgrading all detectors to have better readout speeds by a factor of 5 at each beamline would increase the rates to 13.94, 0.95, and 0.44 Gbps. In other words, the maximum possible data rate would increase significantly, but due to other factors, the overall data rate would remain approximately the same. Similarly, upgrading the ALS source and/or all beamline optics to reduce exposure times by a factor of 5 would result in data rates of 10.37, 1.26, and 0.63 Gbps. And upgrading or automating sample handling so as to decrease sample change times by a factor of 5 results in data rates of 7.24, 1.59, and 0.68 Gbps. More significant increases come if all of the upgrades mentioned are combined, giving rates of 34.56, 2.98, and 1.31 Gbps. Table 1.1 shows a list of these data rates for different scenarios, and Figure 1.7 shows a few of these scenarios plotted on the historic data rate graph from Figure 1.6.

New beamlines that are expected to come on line in the next two years will have an even larger impact than the upgrades mentioned so far. Two of them, Cosmic and IR, are each expected to have an enormous impact on ALS data rates. Importantly, both of these beamlines are in discussion with the

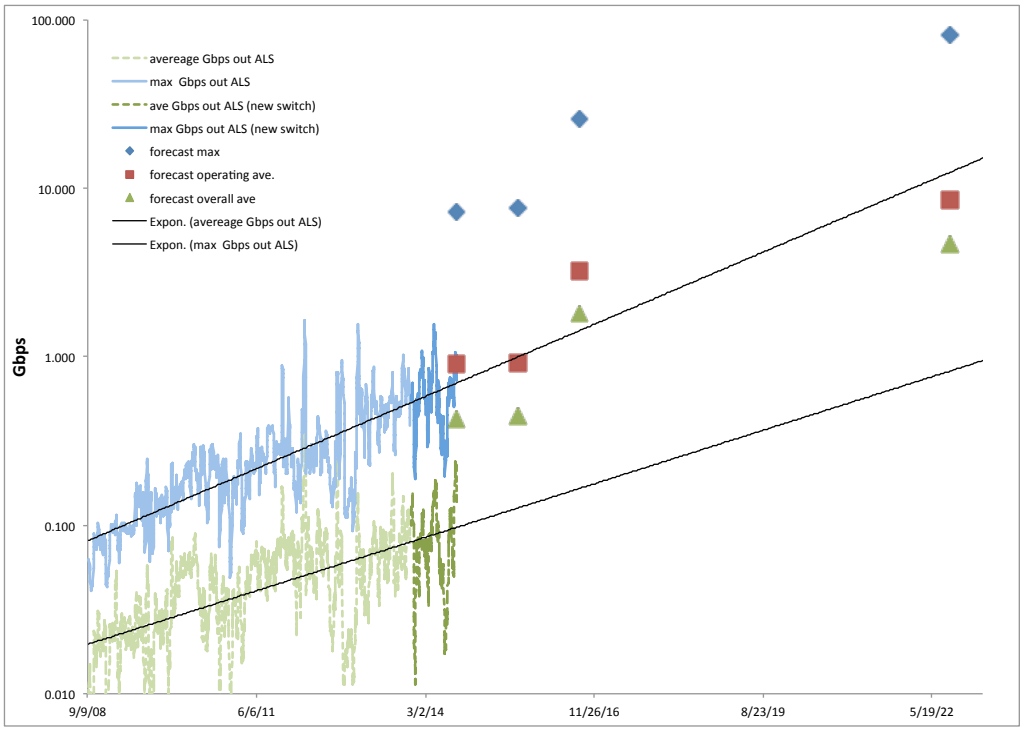



*Figure 1.7: Plot of data rates with some projected scenarios.*

SPOT Suite team, and it is likely that data from those beamlines will be transferred to NERSC under that system. If no upgrades are considered, but the additional beamlines are added as they are currently planned, data rates will increase to 25.84, 3.22, and 1.81 Gbps. If all upgrades are added on top of these new beamlines, data rates will increase to 81.58, 8.43, and 4.68 Gbps.

Given the assumptions in the model, these numbers should be seen as an upper bound on network traffic out of the ALS. Upgrading detectors, exposure times, sample changing, LAN, or WAN across the ALS would be expensive. On the other hand, it is likely that these upgrades will take place on a few of the beamlines which are already the most data intensive, making the numbers more realistic.

Of course, it is often difficult to predict the changes that will come as data rates continue to increase based on simple extrapolation. One protein crystallography beamline scientist pointed out an interesting trend at his beamline, which fits into this category. As that beamline upgraded their detector and sample handling in the last few years, they have greatly increased the rate at which data can be collected. Previously a user might not have collected data on every crystal, instead only investing the time for data collection in the best crystals. Now, the time for data collection has been reduced to the point that many users prefer to collect data on every crystal, and use new algorithms that are being developed to extract partial information and combine that with partial information from other crystals. This means that not only is the maximum possible data rate from the detector much higher than before due to the detector and sample change upgrades, but the average data rate has increased even more due to an accompanying change in the data collection and processing approach.

By a combination of locating some computing resources for real-time processing on site at the beamlines or in the ALS User Support Building server room, data reduction, and selective data saving, the number most relevant for ESnet planning is likely the overall average traffic rather than the maximum or operating average. On the other hand, the computational needs of the new beamlines are expected to be quite high. There is some chance that scientists would want computational resources beyond what is possible on site at the ALS, and there is some chance that they would want the results of these computations in “real time.” In this scenario, the operating average would be relevant for ESnet planning.

# Case Study 2

# Advanced Photon Source

## 2.1 Executive Overview

The Advanced Photon Source (APS) located at Argonne National Laboratory (ANL) produces large volumes of data on the order of 1.5–2 PB/year. New detector technology will increase this as much as an order of magnitude in a few years. Computer networking is a cornerstone technology for operation of the APS facility. Some key needs for that include: transfer of large datasets to user labs or computing centers for remote analysis in a reliable and timely manner; remote instrument operations, which require relatively modest data transfer rates, but minimal data latency and very high reliability; and remote interactive access for visualization and data analysis, which is more tolerant of latency but has more demanding data rates. All scientists would benefit from more advanced conferencing and data sharing platforms that would improve collaboration without a need for travel. To satisfy all these needs, a flexible high-capacity "smart" network is needed that can optimize and prioritize connectivity to match task demands.

## 2.2 Background

The APS at ANL is a synchrotron light source that serves as the nation's premier facility for scientific and engineering research using hard X-rays (Figure 2.1). In fact the facility hosts experiments spanning a wide range of photon energies from about 0.3 keV up to 130 keV, and even as other light sources are brought online, it will remain the brightest X-ray source in the Americas for experiments at energies of 15 keV and above.

The APS began operations around 1995, with a set of beamlines [1] that has grown to about 66 today (Figures 2.2 and 2.3). More than half of these are operated by the X-ray Science Division of the APS, and the rest are operated by Collaborative Access Teams or CATs involving partnerships between universities, research laboratories, and industry. The number of beamlines continues to grow, with the most recent being the Dynamic Compression Sector beamline built by a team led by Washington State University with funding from the National Nuclear Security Agency.

The APS is now in the detailed planning stages for a proposed Upgrade based on an August 2013 Basic Energy Sciences Advisory Committee (BESAC) report[2] which recommended that the United States pursue the development of "diffraction-limited storage rings with beamlines, optics, and detectors compatible with the $10^2$–$10^3$ increase in brightness afforded by upgraded storage rings." This will translate directly to new scientific capabilities. The Multi-Bend Achromat (MBA) lattice upgrade being planned

[1] The term used for each X-ray beam transport system and experimental endstation that can operate simultaneously

[2] The Basic Energy Sciences Advisory Committee (BESAC) report from August 2013 is available at http://science.energy.gov/~/media/bes/besac/pdf/Reports/Future_Light_Sources_report_BESAC_approved_72513.pdf.

*Figure 2.1: Aerial view of the APS at ANL. The Advanced Protein Characterization Facility (APCF) is at lower left, while the Center for Nanoscale Materials (CNM) is at middle left.*

would involve a shutdown for about one year to replace the storage itself, and to build several beamlines and upgrade others to take advantage of the improved source. At the time when this proposed upgrade is completed (possibly around 2021), the APS is expected to be the brightest storage ring X-ray facility worldwide at photon energies above about 3 keV.

Synchrotron light sources serve very diverse scientific communities. Many of the 66 beamlines are optimized for one particular X-ray technique or measurement type. For example, adjacent beamlines at the facility might have a detector development company testing a new detector, followed by researchers from General Electric studying the failure mechanisms of aircraft turbine blades, next to paleontologists studying the internal structure of a rare fossil, near to medical school researchers examining the localization of cancer drugs within cells. Most of these separate research uses involve specific X-ray techniques with their own data volumes and software tools for analysis. Even among similar techniques, there may be differences between beamlines operated by different APS or CAT groups. It is more efficient to develop common data formats, data management systems, and analysis tools where possible, and more beamlines are moving in this direction as certain key software packages emerge. This also helps drive scientific and technical cross-fertilization between beamlines. Still, the APS presents a highly heterogeneous environment in its scientific research, institutions, and computing.

## 2.3 Collaborators

The APS serves almost 6,000 users per year, carrying out almost 25,000 experiments per year (Figure 2.4). Some experiments require days or even weeks of dedicated time such as when one is developing a new method or recording exceptionally weak signals over a range of parameter space, while others take only minutes with robotic systems used to change samples for the next experiment. In the latter case, an increasing number of users are able to work in an offsite mode, where they ship the sample to be studied and then either operate the beamline remotely (for example, in macromolecular crystallography) or simply allow an automated data collection system to carry out the measurements (some powder diffraction experiments are carried out in this manner).

As noted, the scientific communities served by the APS are quite diverse. In Figure 2.5, we show both the scientific categorization of APS experiments, and also the agencies that fund the various research programs (nearly all of the main US research funding agencies). Industry users include those in petrochemistry (in particular, for studies of catalysts), in materials development (transportation materials as noted above, polymers such as at the DND-CAT beamline with Dow and DuPont as partners, or semiconductor materials), and in the pharmaceutical industry (Eli Lily operates its own beamline called LRL-CAT, and several other companies partner together in the operation of IMCA-CAT).

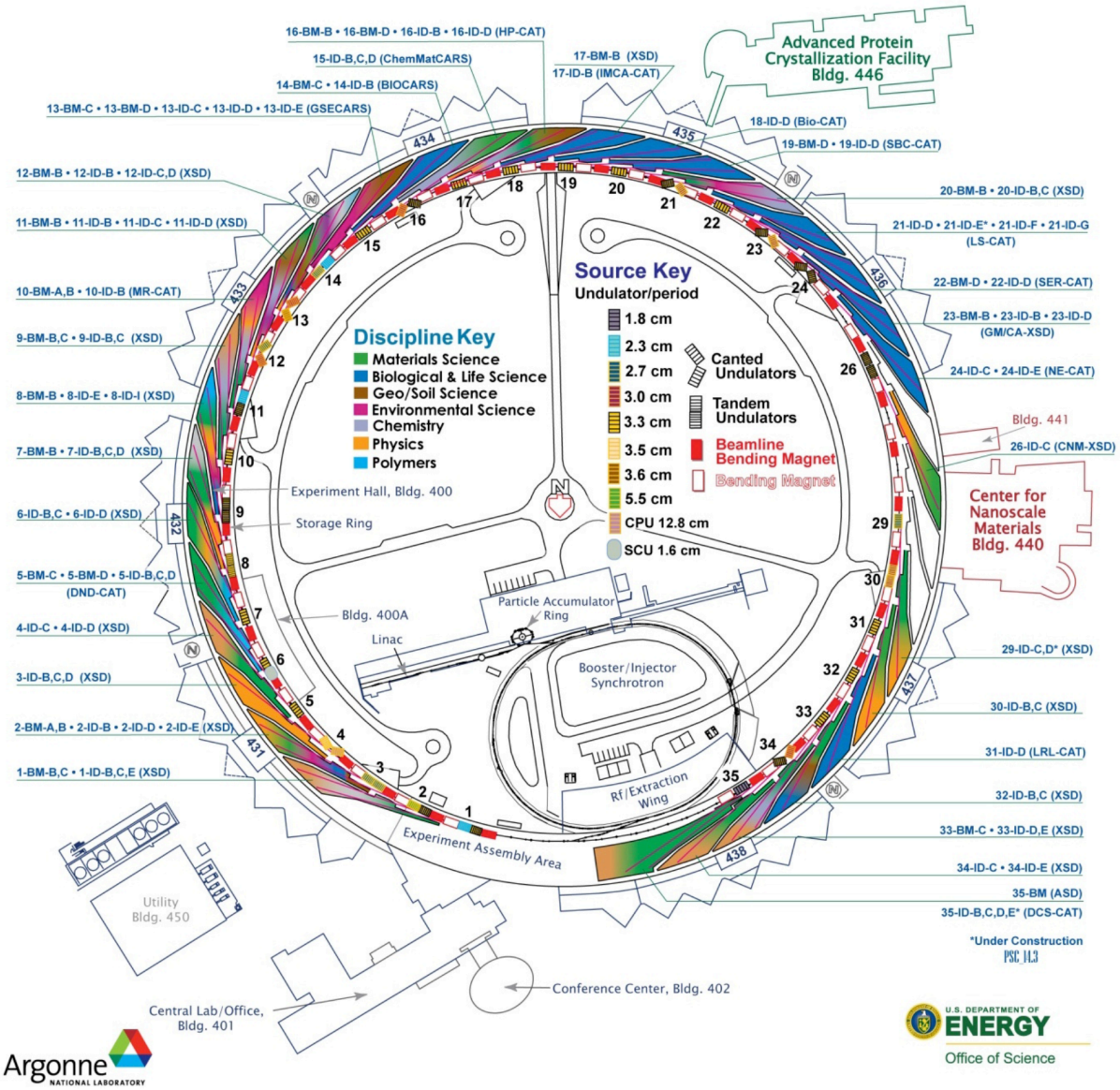



*Figure 2.2: Layout of the APS, showing beamlines in operation or in construction today. The aerial photo of Figure 2.1 was taken from the perspective of above this image, looking down.*

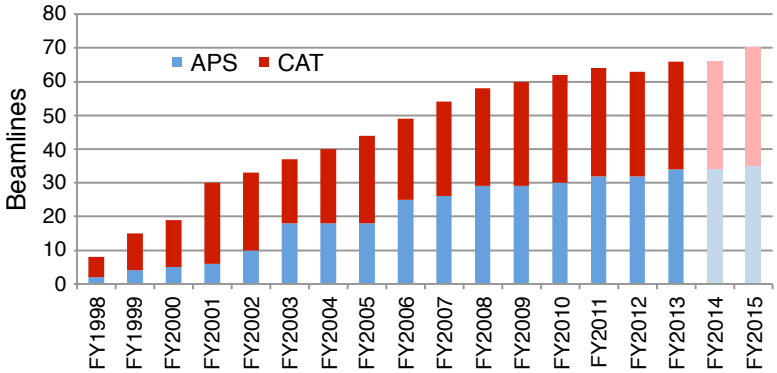
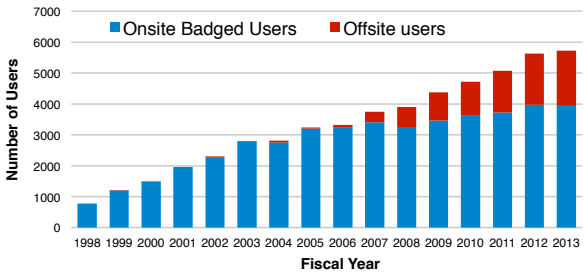
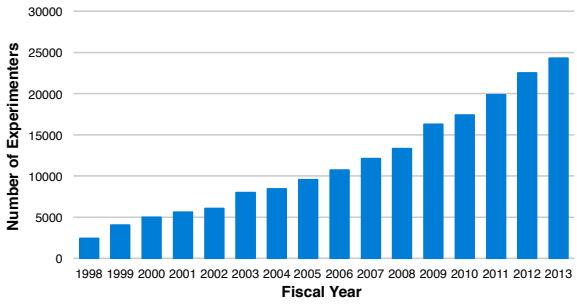



*Figure 2.3: Beamlines (independently-operable experimental endstations) at the APS as a function of time (fiscal year or FY). Some beamlines are operated by the APS, while others are operated by collaborative access teams (CATs). Over time, some former CAT beamlines have come to be operated by the APS, while at the same time additional CAT beamlines have been developed.*

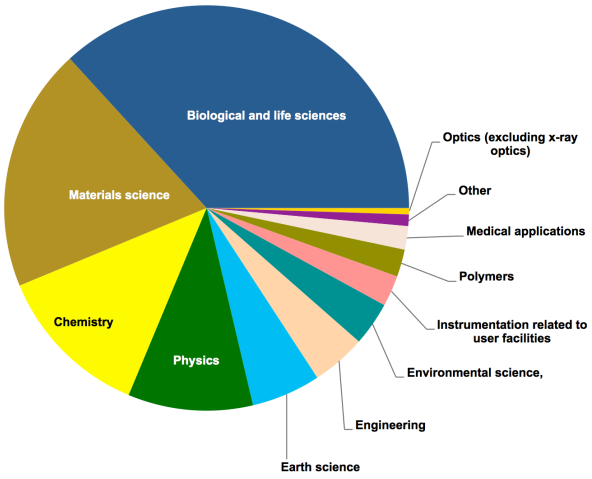



*(a)*

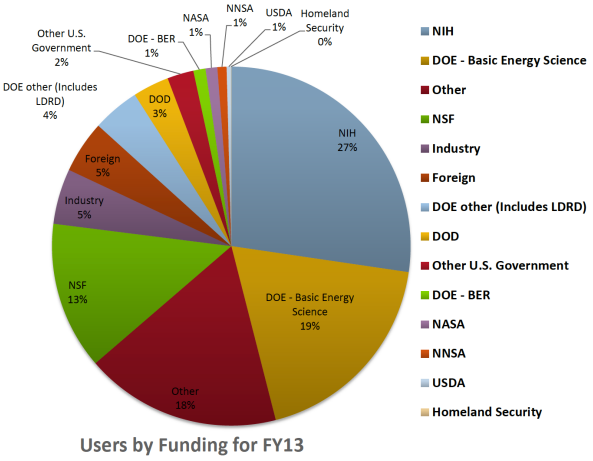



*(b)*

*Figure 2.4: (Left) Number of individual researchers making use of the APS by fiscal year. Most researchers travel to the APS for their experiments (onsite badged users), but an increasing number are able to work as offsite users where they ship a sample and carry out their measurements via network connections with local staff overseeing beamline operation. (Right) Number of experiments carried out at the APS by fiscal year.*




*(a)*




*(b)*

*Figure 2.5: APS researchers by scientific area (left) and by funding source (right). The APS serves a very diverse set of scientific and engineering communities.*

## 2.4 Near-term Local Science Drivers

### 2.4.1 Instruments and Facilities

The APS has a linear accelerator and booster synchrotron used to inject electron bunches into the main storage ring. These accelerators are all controlled using the EPICS protocol over Ethernet connections, as managed by the Accelerator Science Division (ASD) of the APS along with the Controls group of the APS Engineering Support Division (AES).

Beamlines often have dozens of motors, readout electronics, and x-ray detectors that are also controlled by EPICS, with VME crates commonly used to host particular electronic interfaces. Many x-ray detectors have specific manufacturer-supplied interfaces. In some cases this may be a specific hardware interface to a Windows or Linux workstation, but sometimes detector vendors supply a specific detector control computer, and a data buffer computer.[3]

APS-operated beamlines are each equipped with one distributed server or "dserv" computer located in the main APS computer room near the accelerator control room. These computers provide a local boot capability for beamline EPICS devices and workstations, and in some cases they are the primary data storage computer for the beamline. Most dserv computers are connected on a local 1 Gbps network, but some higher data-rate beamlines have a 10 Gbps connection to their dserv. Many beamlines also have several high-end workstations, and some use their own storage solutions such as Windows file servers. The total data acquired each month by each beamline is shown in Figure 2.6. We believe that the aggregate annual data volume of about 1.5–2 PB is as large as that at any DOE user facility in the United States.

The APS has two larger centralized computer systems. The "orthros" cluster consists of 45 compute nodes (three GPU-equipped) and 225 TB of Lustre disk storage (orthros has associated with it a smaller, five-node cluster "blacklab" which is used for testing purposes). The "tao" storage system consists of only 2 compute nodes, but 240 TB of Lustre disk storage (after RAID6 and hot spare redundancies are accounted for), and dedicated internal and external transfer nodes. Both systems have Globus endpoints.

The APS has two 10 Gbps connections to the Argonne central computing facility on one fiber, and another 10 Gbps connection on a separate physical path. The Argonne central computing facility in building 240 hosts the Argonne Leadership Computing Facility (ALCF) for which user proposals must be submitted for access to its four machines: Mira (768000 cores), Cetus (65000 cores), Vesta (33000 cores), or Tukey (1500 cores). ALCF has a total of about 40 PB of storage available. Also hosted in building 240 is the Laboratory Computing Resource Center with resources available to any Argonne employee; LCRC includes the machine Blues (5000 cores) and Fusion (2500 cores). Finally, ALCF and the Globus team have together configured disk storage from a previous generation ALCF machine (Intrepid) as the data storage system "petrel," with a Globus endpoint and 1.7 PB of storage available for DOE facility users, plus management software that permits users to request allocations and then manage access to data within their allocations.

### 2.4.2 Software Infrastructure

Because the majority of APS beamlines were built as CAT beamlines by a variety of institutions (Figure 2.3), and also because of the diversity of scientific communities served by the APS (Figure 2.5), the software infrastructure of the APS is very heterogeneous. All beamlines use EPICS to control the undulator source energy and to monitor some APS source parameters, most use EPICS for control of basic beamline hardware, and many use EPICS for control of their experimental endstations. There are fewer commonalities in data management, data transfer schemes, file formats, or analysis programs.

[3] An example is the large Dectris Pilatus3-X-6M system, where Dectris encourages the purchase of their Pilatus Processing Unit computer system in order to deal with data rates as high as 1.7 GB/s.

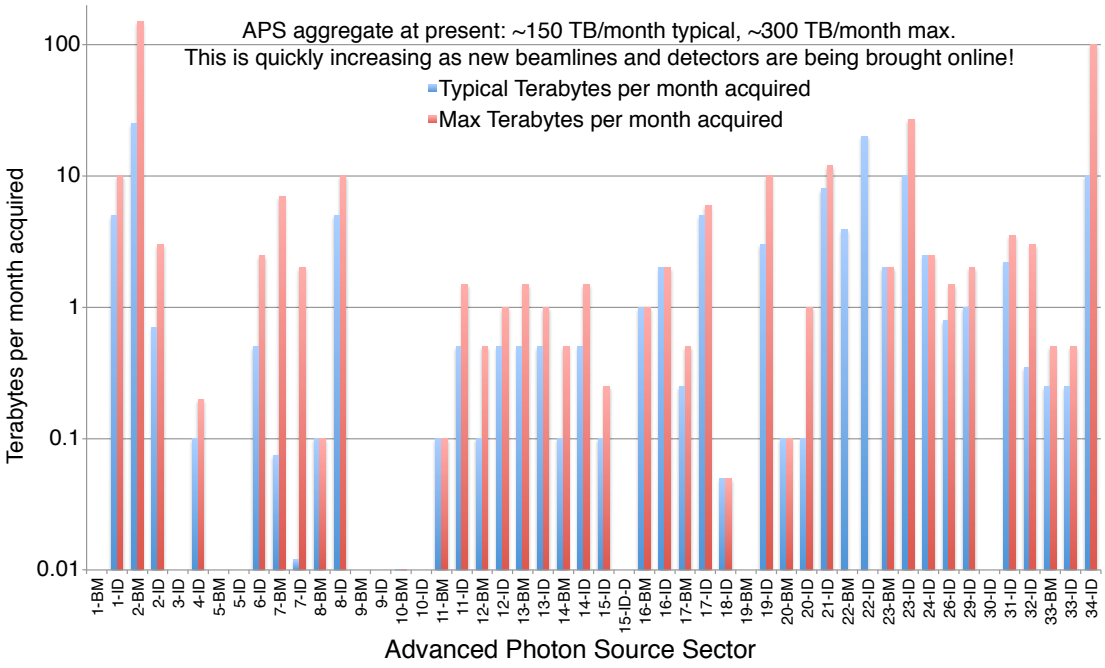



*Figure 2.6: Data acquired per month by APS beamline, as reported by each beamline's person responsible for computing in summer 2013.*

While we have available a separate document with specific case studies from several beamlines, we summarize here the trends:

- Many beamlines use locally-developed software programs for data acquisition and analysis. Data files are often saved as ASCII text for simple spectra, or Tagged Image File Format (TIFF) files if area detectors are used.
- The dominant data management solution is for beamline scientists to record metadata about experiments in paper logbooks, and to transfer data to users via portable hard drives.
- Macromolecular crystallography presents a uniquely unified community due to the near-identical nature of data collection and processing among all experiments. Efforts such as the Protein Data Bank,[4], dating back to 1971, and CCCP4,[5] dating back to 1979, have helped provide common data file formats and subroutine libraries for the standard sequence of operations used in obtaining molecular structures from x-ray diffraction data. Many macromolecular crystallography beamlines provide remote operations for experiments, and distribute data via their own ftp servers.
- A small but growing number of beamlines are using Globus services[6] for data distribution, synchronization, and/or sharing. Several are helping the Globus team test out additional tools such as the Globus Catalog, which is intended to allow for easy searchability of data files based on a database formed from experimental metadata.
- The majority of large data sets come from pixelated area detectors, and most of these use an EPICS-based framework called AreaDetector for their control and data acquisition.[7] The framework has been developed by a CAT team (CARS) lead by University of Chicago, and implementations for various detectors have been developed by CARS, the APS, and by other light sources such as Diamond in the UK.

[4] www.rcsb.org
[5] www.ccp4.ac.uk
[6] www.globus.org and Foster, *IEEE Internet Computing* (May/June), 70-73 (2011)
[7] See cars.uchicago.edu/software/epics/areaDetector.html.

- A small but growing number of beamlines are using Hierarchical Data Format 5 (HDF5)[8] as their basic file structure. Several beamlines are now using a schema called Scientific Data Exchange[9] for organizing their HDF5 files for both raw and analyzed data. With tomography in particular, this schema is spreading to other facilities. One user from the Materials Science Division and one APS beamline use the NeXus schema[10] for HDF5 file writing.

- An Argonne LDRD is supporting a pilot project for automatic data management. The goal of this project is to use the APS user database (APS badge numbers) to create Globus accounts, and the APS Experimental Safety Approval Forms (ESAFs) to create Unix groups into which data are placed. The pilot project is testing Spade[11] for automating the migration of data from beamline computers to Globus endpoints. The goal is to remove the burden of data management from the shoulders of beamline scientists, so that users will automatically have a Globus account where they can see only the data from experiments where their badge number was listed on the ESAF. A similar approach to automated data organization has been developed for the high-throughput "mail-in" program in powder diffraction at the APS beamline 11-BM.

- Argonne has also funded an LDRD as part of a *Grand Challenge on Data-Driven Science*, to provide an interactive framework that enables processing of large data sets (20–50 GB) and comparison to advanced simulations in real-time on the many beamlines that do not have well-defined analysis workflows. It can be challenging to develop optimized software when each measurement requires identical analysis steps, but it is even harder when the instrumental configuration is changed for each experiment and facility users have to tailor their analysis to each scientific problem. The raw data and harvested metadata are meant to be streamed to a remote server, such as petrel, and merged to form NeXus files, so that the experimental scientists can perform preliminary inspections and analysis of the remote data using a customizable Python-based GUI that is run locally on the beamline. The Swift Parallel Scripting system,[12] being developed at the University of Chicago, ensures that compute-intensive tasks can be performed remotely in real-time without losing the flexibility of a scripting interface to the data. This is being initially tested in the analysis of single crystal diffuse scattering on Sector 6 and 11. The work is being coordinated with a BES/ASCR-funded pilot project between Argonne and Oak Ridge National Laboratory to conduct the joint analysis of x-ray and neutron scattering data; a new diffuse scattering beamline, *Corelli*, is now being commissioned at the Spallation Neutron Source. There are also plans to deploy these developments in high energy diffraction microscopy and grazing incidence small-angle scattering (GISAXS).

- Several other beamlines have achieved, or are working towards, near-real-time analysis of data. These projects are described in further detail in a separate document with case studies. However, some highlights from the report include:

  - In x-ray photon correlation spectroscopy at beamline 8-ID, a Hadoop-based analysis system has been developed. This allows for experimenters to see when they have acquired sufficient statistics on short-time-scale phenomena, or sufficient correlation times for slower phenomena.

  - Beamlines 1-ID and 34-ID both carry out different forms of x-ray diffraction microscopy to map crystalline domains in materials. They have especially well-developed data handling pipelines using parallel computing for data reconstruction, and in the case of 1-ID their workflow has been developed in collaboration with researchers in Argonne's Mathematics and Computer Science Division to make use of Swift. The resulting capability allows beamline users to verify results during experiments.

  - Tomography reconstructions at beamlines 2-BM, 2-ID, and 32-ID are now being carried out using the TomoPy software package developed via Argonne LDRD support.[13] This package

[8] www.hdfgroup.org/HDF5/

[9] See www.aps.anl.gov/DataExchange/, github.com/data-exchange/, and De Carlo et al., *J. Sync. Rad.* **21**, 1224 (2014).

[10] www.nexusformat.org

[11] nest.lbl.gov/projects/spade/html

[12] swift-lang.org

[13] github.com/tomopy/, and Gursoy et al., *J. Sync. Rad.* **21**, 1188 (2014).

has recently been parallelized, allowing for a 700x speedup in processing based on algebraic reconstruction algorithms.

  - With joint support from Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) at the Department of Energy (DOE), a toolkit and operating program called PtychoLib has been developed to speed up ptychography reconstructions by a factor of about $10^4$ using Graphical Processing Units and MPI programming on 128 nodes of the tukey cluster at ALCF.[14] This work has allowed for reconstructed images to be viewed as the experiment proceeds.

- In the past two years, there have been an increasing number of connections with applied mathematicians and computer scientists in the Mathematics and Computer Science Division at Argonne to develop new approaches to analyze and understand data acquired at the APS. Examples include the use of Swift and the development of PtychoLib noted above, as well as a new ASCR-funded project on Robust Optimization and Modeling for Phase Retrieval (ROMPR) as well as new collaborations to simulate molecular configurations and compare them with x-ray scattering measurements (Knight and Bachir; Ferrier and Nealey). Other recent examples include the automated classification of x-ray fluorescence map data,[15] and the application of optimization methods to x-ray spectromicroscopy analysis.[16]

Again, the above represent only particular vignettes among a much larger sweep of activities at the APS. There are also a variety of sophisticated analysis programs developed by users of the APS for their own research; one challenge is that relatively few of these programs make it back to the facility in a way that they can be used by other researchers.

### 2.4.3 Process of Science

In spite of the large size of synchrotron light source facilities like the APS, most experiments are performed by small teams over short visits to the APS. These teams must submit user proposals to the APS about three months before the earliest possible experiment start time, and oftentimes their beamtime request cannot be accommodated until the following scheduling cycle four months later. There are mechanisms for rapid access for preliminary studies, but most physical science experiments involve this extended planning process, while almost two-thirds of the macromolecular crystallography beamtime requests are submitted as rapid access proposals.

In synchrotron studies, the science is often in the sample, and research groups may spend some time synthesizing a material, culturing and treating biological cells, or growing crystals. There is often extensive sample examination done in the home laboratory, including various spectroscopies and/or microscopies, but those data are almost never integrated together with data taken at the APS except through notations in the experimental team's logbook. In some cases, simulations might be carried out to design material synthesis or to test new algorithms that might be used for data analysis, and those can help guide experiments.

When the scheduled beamtime arrives, a subset of the research team arrives at the APS for their experiment. Depending on the complexity of on-site sample handling, they may arrive several days in advance, or simply show up the morning when their beamtime is scheduled to start. Depending on previous experience, some hours may be required to train the user in operation of the beamline, and sometimes work must be done to adapt the beamline's equipment to the particular needs of the experimenter. The user's beamtime might then run for several days of around-the-clock operations, which might be handled by having the team work in shifts or by having long acquisition runs go overnight.

During the experiment, researchers are eager to begin evaluating their data in order to guide the progress of the experiment. Here the experience varies greatly: in some cases one can obtain a preliminary analysis immediately (such as an electron density map in crystallography, an autocorrelation plot in photon

[14] Nashed et al., *Optics Express* **22**, 32082 (2014).
[15] Wang et al., *J. Sync. Rad.* **21**, 568 (2014).
[16] Mak et al., *Faraday Discussions* **171**, 357 (2014).

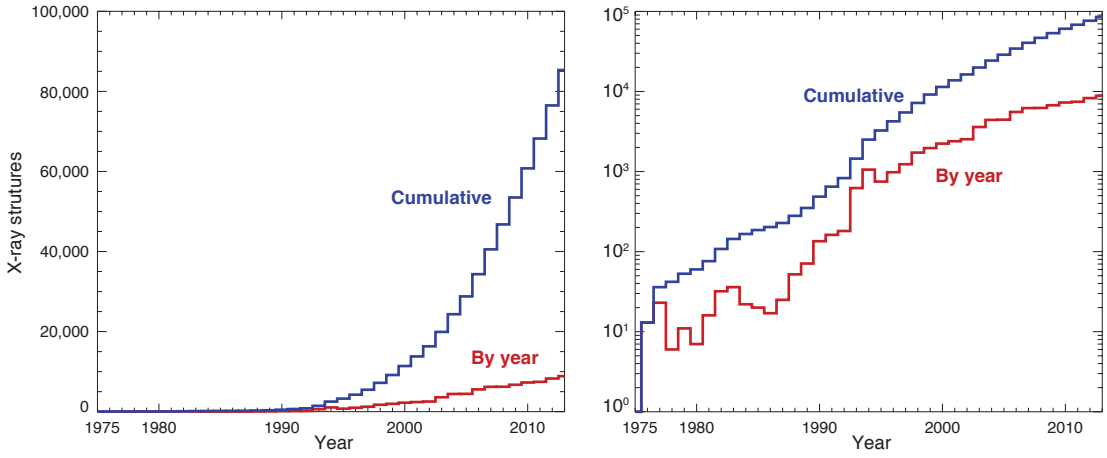



*Figure 2.7: X-ray crystal structure depositions in the Protein Data Bank (www.rcsb.org) by year. Starting around the mid 1980s, the vast majority of the structures have involved data taken at synchrotron light sources. The APS has been used for more than 20% of these, which is the most of any synchrotron light source.*

correlation spectroscopy, a comparison with a reference spectrum, or a volume rendering in tomography), while in other cases the data require considerable theoretical modeling for their interpretation (this is especially true in inelastic x-ray scattering studies). A more representative case is where one has an immediate view of the data for rough interpretation, while a truly quantitative analysis requires additional information and computation which is typically done after the beamtime has been completed.

At the conclusion of the scheduled experimental time, the research team often quickly copies data onto portable hard drives and perhaps also makes photocopies of key pages of the experimental logbook for sharing with beamline staff. They then return home with data in hand, and with some analysis tools that they have developed, have been provided by a beamline scientist, or that are available for download or purchase. Time is then spent going over the data, carrying out more complete analysis, and then preparing results for publication (usually with considerable involvement of the APS beamline scientist, especially for the large fraction of APS users who are experts in their own scientific domain but not in x-ray science, such as biologists).

This describes the reality for many experiments today, and as can be seen it involves relatively little in terms of network infrastructure. In Sections 2.5 to 2.8 below, we describe counter-examples to the above in which data networks play a significant role, both today and in the hoped-for future.

## 2.5 Near-term Remote Science Drivers

### 2.5.1 Instruments and Facilities

The best example of remote use of the APS today is in macromolecular crystallography. This field has been completely transformed in recent decades (Fig. 2.7) by improvements in crystallization methods, cryogenic crystal loading robots and data collection systems at synchrotron light sources, photon counting detectors, and analysis software. As a result, a significant fraction of the structures obtained are from groups in molecular and cell biology who are not specialists in crystallography. The APS hosts several regional, national, and industrial beamlines for macromolecular crystallography, and a new Advanced Protein Characterization Facility (APCF) has been built next to these beamlines at the APS (Figures 2.1 and 2.2).

To see one example of remote access crystallography, consider the General Medical Sciences and

Cancer Institutes Structural Biology Facility[17] (GM/CA) CAT beamline at the APS. Researchers new to crystallography often come in person for their first beamtime to understand the data-taking process, but then they often use remote access for follow-on experiments. Crystallographers have settled on just a few designs for sample mounting, and these standard mounting "pins" can be sent overnight in a cryogenic dry shipper to the APS. Pins are then loaded into a "puck," or multi-pin mounting cartridge, which is in turn placed in a liquid nitrogen dewar at the crystallography beamline. Data collection is then controlled by a program called JBluIce (see below).

The beamline hosts the usual set of beamline motors driven by EPICS, and a large-area detector. The newest detector installed at one of the GM/CA beamlines is the Pilatus3-X 6M, a detector with 2463×2527 photon-counting pixels with a data depth up to 20 bits, and a maximum frame rate of 100 Hz. This detector is connected directly to a control computer that streams data files to a Pilatus Processing Unit which consists of a custom-configured Linux computer with 32 Xeon cores, 768 GB of RAM, and 10 TB of RAID 50 disk. Supporting all of this is a computer network consisting of 10 Linux workstations, several of which have a high speed (80 MBps) connection to a 56 TB disk array running Global File System (GFS).

The most crucial networking requirement for remote access is the ability to use a remote desktop program like NxPlayer with sufficient throughput. Packet round-trip times of 20 msec, and overall transfer rates of 50 Mbps are sufficient for a successful experience for remote data access. It should be noted that this remote access mode has been used with great success even with very challenging small-crystal studies, such as that which led to the group of Brian Kobilka of Stanford University receiving the Nobel Prize in 2012!

### 2.5.2 Software Infrastructure

The GM/CA-developed program JBluIce handles all of the data acquisition. Crystallographic analysis is done by using any one of a number of programs in the "standard suite" developed within the crystallographic community.[18] In addition, automated data processing is initiated after a few frames have been collected, and two threads run in parallel: one is from scripts developed at the Diamond Light Source, and the other is GMCAproc, which was recently developed at GM/CA. The parallel processing increases the probability of determining a solution to the electron density.

### 2.5.3 Process of Science

Macromolecular crystallography begins with a research group in their home lab learning how to express a protein of interest (for example, using genetically-modified yeast), purify it, and crystallize it (similar process cycles exist for studies of other macromolecules such as viruses). This can take some time, and crystals are often screened with a laboratory x-ray source in the home lab to guide the optimization of crystallization conditions. Once crystals have been produced, they are mounted on standard pins and shipped to the APS for mounting in a crystal puck.

Data collection is controlled by a program called JBluIce (a locally customized, Java-based evolution of the BluIce software package developed earlier at Stanford Synchrotron Research Laboratory). The user selects a crystal pin from the puck, then an automated sample mounting robot grabs the pin and places it in the beam path. The crystal can be aligned to the beam using an on-axis visible light microscope with a viewing screen within JBluIce, and a series of regions on the crystal can be selected for data collection. The crystal is then scanned and rotated for the data collection series, before being returned by the robot to the puck. As described above, data processing is initiated automatically, but users can manually process data on a workstation at the beamline for online analysis during the day of the experiment, or a separate workstation for just one day after. Data can be shipped to the user via portable hard drive, or transferred over the network using Globus.

[17] www.gmca.anl.gov.

[18] A listing of those programs available on the GM/CA workstations is provided at www.gmca.anl.gov/computing/software_crystallography.html.

Users will often carry out further analysis of their crystallographic data in their home lab, especially if they have complications in their diffraction patterns such as crystal "twinning." The amino acid sequence of the protein is then fit to the electron density map obtained from the crystallographic data. The crystallographic community has long ago made it a requirement for journal publication that the structural data be submitted to the Protein Data Bank, with the option to delay release until one year after the paper describing the research has appeared.

## 2.6 Medium-term Local Science Drivers

### 2.6.1 Instruments and Facilities

As noted above, the APS is a very diverse facility with a large number of experimental techniques represented. Therefore consider two examples of studies where local presence is especially necessary because of the nature of the sample handling: dynamic tomography of materials failing under strain, and cryogenic x-ray ptychographic imaging of frozen hydrated biological cells. The instruments and facilities involved in these respective techniques are as follows:

- **Dynamic tomography**: by using broad-spectrum illumination at bending magnet beamline 2-BM, one can obtain sufficient photon flux to record an entire 1500-projection tomographic data sequence in under one second. This provides a unique opportunity to obtain 3D information on how lightweight structural materials (such as might be used in aircraft) fail under strain, and what material factors play a role. Because of the need to configure the strain-inducing mechanism, these are "local" or on-site experiments where the scientific team must travel to the APS to carry out their experiments. These experiments involve using a scintillator and lens system to deliver the x-ray projection image onto a visible light camera, and future enhancements might include the purchase of a camera such as the pco.dimax HS4 camera which can record 2000x2000 pixel images (12 bit depth) at a rate up to 2277 frames per second, with onboard storage for about 13,000 frames. Data can then be transferred from the camera to the beamline computer using a GigE interface (1 Gbps), where the data is stored prior to reconstruction and analysis as described below.
- **Cryo ptychography**: x-ray ptychography is a coherent diffraction imaging method where x-ray diffraction patterns are recorded as a coherent beam spot is moved to overlapping positions on a specimen. By phasing the diffraction patterns as assisted by their overlaps, one can obtain an amplitude and phase contrast image of a specimen beyond the resolution limit of the lens or beam-limiting aperture. This is especially useful in hard x-ray microscopy of biological specimens, where phase contrast dominates and where a spatial resolution in the 10–20 nm range is required to identify many organelles within cells.[19] Because of the need to work with cryogenically prepared samples so as to minimize radiation damage (using, for example, the Bionanoprobe instrument at 21-ID), cryo ptychography experiments will likely remain "local" experiments employing an NIH-supported cryo sample preparation laboratory established at the APS. Especially as one couples cryo pychography with x-ray fluorescence microscopy, datasets will evolve towards 1000x1000 recordings of diffraction patterns of size 512x512 pixels, each with a dynamic range of 16 bits or more, leading to a dataset size of about 0.5 TB for a single recording. Given that ptychographic tomography has already been demonstrated on room temperature specimens at the Swiss Light Source, we expect to move cryo pychography into 3D where one might acquire as many as several thousands of projection angles, leading to a data volume per sample approaching 1 PB (because of radiation dose limitations, these recordings will likely be very sparse so the data is likely to be highly compressible). Future experiments might hope to make use of detectors such as the EigerX-1M, which is capable of recording 1030x1065 pixel frames at 12 bit depth and at a rate of up to 3000 frames per second. This camera is ideally coupled with a Dectris Pilatus Processing Unit computer. Reconstruction of such large datasets requires the use of GPU-equipped cluster computers, as will be described below.

[19] J. Deng et al., *Proceedings of the National Academy of Sciences* (in press).

### 2.6.2 Software Infrastructure

**Dynamic tomography**: the first stage of data processing for dynamic tomography involves the reconstruction of 3D volumes from a set of projection images. This is done using the program TomoPy.[20] While the first version has been developed for use on single workstations, Doga Gursoy and Tekin Bicer have recently begun testing a cluster-parallelized version of TomoPy in which each slice along the length of the rotation axis is independently reconstructed on a single node. They have already demonstrated a 700-fold speedup in tomographic data processing by this approach.

**Cryo ptychography**: to obtain a reconstructed image from a ptychographic dataset, one must simultaneously "phase" all the diffraction patterns. An earlier version of software used to solve this problem ran for several weeks on a single workstation, which is clearly impractical for experiments where one would like to judge the quality of the specimen to see if the sample preparation conditions were good or if the right region was imaged. As was noted above, ASCR and BES have together supported the development of a toolkit and operating program called PtychoLib,[21] which has already achieved a 200× speedup in ptychographic image reconstructions by employing Graphical Processing Units (GPUs). This program is now being extended to work on a GPU-equipped cluster computer and additional 100× speed gains have already been obtained in simulations. With the advent of continuous scan schemes for higher throughput psychographic data collection,[22] the needs for rapid reconstruction only increase.

### 2.6.3 Process of Science

While the sample preparation requirements will likely make dynamic tomography and cryo ptychography experiments that continue to be done by scientists visiting the APS, the analysis of the data involves more than just initial volume or image reconstructions as described above. In the case of tomography, one might want to go from a series of reconstructed volumes to an analysis of just where the material failed and how that corresponds to inclusions, voids, or grain interfaces within the material (see for example Williams et al, *Acta Materialia* **58**, 6194 (2010)). There is no "standard" software now available for analyses of this sort, and the data volumes involved (a time series of around 10 GB volume reconstructions) are also challenging. In ptychography, the raw data volume is even larger and the computational work required to obtain an image is greater. Given that the scientific users of these capabilities might have consider expertise in materials synthesis (dynamic tomography) or cell biology (cryo ptychography), high performance computing might be too far beyond their skill set or the computer hardware capabilities they have immediately available.

One solution to these challenges is to develop a process where users of these techniques rely more on computational facilities at Argonne for the analysis of their data. A very interesting trend is the development of virtualized platforms in cloud computing, where one can have a specific analysis program be installed and preconfigured on a virtual host for on-demand access. Even the visualization of large volume data can be challenging, so one can imagine this process including tools such as ParaView where the data resides on higher-end compute facilities at Argonne and the user only requires a network connection with sufficient speed to drive their local display of 2D renderings from the 3D data as the sample is rotated, the data is zoomed into, and so on. This might require views of up to 2000×2000 pixels to be interactively transferred at rates of 60 frames per second, or a data streaming rate (in 24 bit color mode) of about 5 Gbps. Clearly this is well beyond standard network requirements, so ESnet has much to offer.

[20] Program developed with Argonne LDRD support, and Gursoy et al., *J. Sync. Rad.* **21** (2014).
[21] Nashed et al., *Optics Express* **22**, 32082 (2014)
[22] J. Deng et al., *Optics Express* (in press).

## 2.7 Medium-term Remote Science Drivers

### 2.7.1 Instruments and Facilities

Again we consider one example out of the many diverse research capabilities available at the APS. X-ray fluorescence microscopy provides a thousand-fold improvement in trace element detection relative to electron microscopes equipped with X-ray analysis systems, and this information is crucial for research topics such as understanding the role of trace elements in several diseases as well as the subcellular targeting of potential cancer drugs. While frozen hydrated samples imaged at cryogenic temperatures provide the best structure and chemical preservation and limited studies on them might be included as a cross-check, one can carry out much more rapid screening studies on populations of cells using freeze-dried samples in room temperature X-ray fluorescence microscopes such as those at beamlines 2-ID-D and 2-ID-E at the APS. At present these studies are carried out by users arriving in person for their beamtime, but one could imagine following the examples of crystallography and powder diffraction and going towards remote access for mail-in sample studies. One key instrument needed for X-ray fluorescence microscopy is a high quality visible light microscope with digital image capture and motorized sample stage for pre-identification of sample areas for x-ray imaging. Another is of course a scanning X-ray nanoprobe instrument, where X-rays are focused to a small spot and characteristic X-rays are recorded using an energy-dispersive detector as the specimen is scanned and (in the case of fluorescence tomography) rotated. With the recent introduction of continuous-motion or "fly" scans as opposed to move-stop-measure or "step" scans, X-ray fluorescence microscopy has moved to the acquisition of megapixel datasets with each pixel containing a full X-ray spectrum recording over about 1000 energy points. When combined with emerging capabilities in fluorescence tomography, one can imagine arriving at datasets with volumes approaching 1 TB before compression.

### 2.7.2 Software Infrastructure

At present users travel to the APS to carry out X-ray fluorescence microscopy experiments. They obtain an immediate online view of approximate elemental content using simple energy "windows" set around fluorescence lines. However, proper quantitation requires fitting of the full fluorescence spectrum, and this is done using the program MAPS (Vogt, *J. Physique IV* **104**, 635 (2003)) which is typically run as a background job with results available some hours after acquisition. For fluorescence tomography, the practice up until now has been to take the individual elemental concentration projections from MAPS and read them into NIH ImageJ for tomographic reconstruction, which is a very tedious manual process with poor algorithmic suitability for fluorescence tomography.

### 2.7.3 Process of Science

We describe here an improved workflow that would be required for the remote operation of fluorescence microscopy and tomography with freeze-dried speciments viewed at room temperature.

The first step is to record a large-area visible light micrograph of the specimen so as to identify the promising areas for imaging. There is an effort by the Software Services Group at the APS to develop such a mosaic imaging capability using a Leica microscope with digital camera, and to register this view on top of a large-area X-ray fluorescence map. Ultimately one would like to provide this capability for a variety of light microscopes in home labs, so that users could mail-in samples plus set their own pre-annotated light microscope mosaics to define scan regions.

The second step is to speed up MAPS processing so that quantitative results are available online. At present this is done by per-pixel fitting using a program written in IDL (Exelisvis Inc.), but a conversion to a Python-based matrix analysis code should speed this analysis up considerably.

The third step for fluorescence tomography is to take the MAPS output and feed it into TomoPy for volume reconstruction. A project to develop this workflow is already underway.

Finally, one should build upon early demonstrations of cell identification and elemental histogramming capabilies (Wang et al., *J. Sync. Rad.* **21**, 568 (2014)) to develop a workflow-usable set of analysis tools to extract the desired information from a more complex dataset.

With tools such as the above in place, one could develop X-ray fluorescence microscopy into a more automated workflow with more efficient use of APS beamline scientists and more efficient utilization of X-ray fluorescence microscope beamtime. This could allow users to work from their home institutions rather than travel to the APS for routine data collection. It would require similar developments in local data storage and analysis program availability as described in Section 2.7.

## 2.8 Beyond 5 years

In response to a national need described in an August 2013 report of the Basic Energy Sciences Advisory Committee (BESAC), the APS is developing plans for a major upgrade that would boost the brightness of delivered X-ray beams by a factor of 100 while providing additional signal gains due to improvements in optics and detectors. This upgrade, which could be completed as soon as about 2021, will increase the data and computation needs of experiments including x-ray nanoprobes (including X-ray fluorescence microscopy), X-ray ptychographic imaging, and x-ray photon correlation spectroscopy. When coupled with the growing capability (pixel count, frame rate) of commercial detectors used in other experiments, one can easily imagine that the aggregate volume of data collected by the APS will increase by a hundred-fold, reaching the level of many petabytes of data per month or even per week.

## 2.9 Network and data architecture

As the above use cases note, the APS has significant needs for high speed data networks. Some users will always wish to transfer data back to their home institution for storage and analysis. In other cases users may find it preferable to use cloud computing resources (whether from commercial services where competition is driving the cost down to very inexpensive levels; or using resources located at Argonne or other national laboratories, such as might be provided in a Federal Big Data initiative) for both data storage and computation. All users who have successful beamtime proposals at the APS have great expertise in their scientific domain, but many are not computationally savvy and for those users the best way to deal with their data may be to use Argonne computational resources with pre-installed analysis programs. These programs can then display results on the users' laptop or desktop computer.

## 2.10 Data, Workflow, Middleware Tools and Services

The unique characteristics of X-rays include penetration into thick materials, and minimal plural scattering compared to the use of electrons or visible light photons. This makes x-rays ideal for studying thick, complex, heterogeneous materials, and the complexity of such materials plus the many capabilities available at modern synchrotron light sources, combined to yield incredibly rich and large datasets. This places high demands on data storage and computation resources, and the networks needed to support the movement of data between these resources.

Because of the wide span of X-ray science techniques, it would be challenging to come up with "One Big Analysis Program" that could support the examination of all data at the APS. An alternative goal can be to find unified mechanisms for data management and file format schema, and basic toolkits for reading and writing data files, plus scalable subroutines for key mathematical operations, such as discrete Fourier transforms and matrix algebra operations.

While considering the advantages of storing data and providing computational resources at Argonne, one must also consider the cost competitiveness of the commercial marketplace. Competition between

several large vendors with the economies of scale are driving costs for data storage and computation down to very low levels. In an era of stiff competition for scarce research resources, this is a compelling argument in favor of the use of commercial services. Of course one popular scientific service (Globus) already uses commercial cloud services to provide on-demand compute resources for coordinating data transfers.

Finally, it is easy to fall into the pattern of thinking of data transfer and computation as involving only computer and network hardware. In many areas of X-ray science, what is even more valuable is the research and implementation expertise of applied mathematicians and computer scientists at places like Argonne's Mathematics and Computer Science (MCS) Division. As we have described above, there are recent and growing connections between the APS and MCS, and these connections have led to new algorithms and new analysis codes which have provided hundred-fold speedups in data analysis with no new computer or network hardware involved.

## 2.11 Outstanding Issues

While GridFTP and Globus simplify data distribution and sharing, and have improved achieved performance relative to prior approaches, high-speed network performance continues to be a problem in many settings due to various "last mile" problems. More effort needs to be devoted to diagnosing and optimizing end-to-end performance. Simply put, an ESnet that delivers 100 Gbps to Argonne is of little use to APS if it is not connected to APS and other resources at high speeds.

Growing data volumes and increasingly sophisticated data reconstruction and analysis methods imply increasing demand for computing resources. Widespread use of new methods such as quasi-real-time data evaluation and feedback, automated feature detection, and integration simulation and experiment will require on-demand access to computing on a scale not currently supported at Argonne. While not specifically a networking problem, this need has implications for networks in two respects: first, if on-demand computing is not provided at Argonne, researchers will inevitably want to move data over the network to remote locations for analysis, increasing network requirements. Second, regardless of where computing is performed, increasing ability to perform on-demand computing will increase demands for high-speed networking, especially when coupled with new beamline detectors.

*Table 2.1: The following table summarizes data needs and networking demands for a few key instrument types of the APS.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · Macromolecular crystallography.<br>· Instruments: Pilatus3-X-6M represents high-end detectors today.<br>· Software: suite of community developed tools. | · Crystal growth in the home lab, ship crystals to APA, remote beamline operation, automated first-pass data analysis. | | | · For remote beamline operation: 50 Mbps, <20 msec round trip packet time. |
| · X-ray Photon Correlation Spectroscopy.<br><br>· Instruments: LBL/ANL FastCCD2 at 1 kHz frame rate, Eiger 1M detector at 3 kHz frame rate.<br><br>· Computational facilities: 120 cores with 320 GB RAM dedicated to Hadoop-based analysis. | · Non-static sample prepared, acquire 1000-20,000 frames at ~kHz frame rate. | · 3-40 GB per sample, tens of samples per day.<br>· Going towards 30 TB/month. | · 40 GB in <1 minute for near-real-time analysis. | · Typically 5 TB data transfers per user group, every few days. |
| **2-5 years** | | | | |
| | | | | |
| **5+ years** | | | | |
| | | | | |

# Case Study 3

# Center for Nanophase Materials Sciences

## 3.1 Background

The Center for Nanophase Materials Sciences (CNMS) is one of five Nanoscale Science Research Centers (NSRCs) that was established as part of the Department of Energy's (DOE) Office of Science contribution to the U.S. Government National Nanotechnology Initiative (NNI). The CNMS provides a diverse user community with access to state-of-the-art nanoscience research capabilities, expertise, and equipment, and applies these resources to execute a cutting-edge science program with emphasis in theory and simulation, nanofab¬rication, macromolecular synthesis and characterization, and understanding of structure, dynamics and functionality in nanostructured materials using scanning probe microscopy, electron microcopies, neutron scattering, optical spectroscopy, helium ion microscopy, and atom probe tomography.

Motivating the research at the CNMS is the realization that the broadest range of energy applications relies on materials that are enhanced or even enabled by nanoscale phenomena. As described below, this applies in particular to batteries, supercapacitors, photovoltaics, catalysts, thermoelectrics, and many additional functional or even structural materials. Therefore, the central motivation for the work at the CNMS can be summarized as our desire to harness energy through nanoscience. To this end, the mission of the CNMS is twofold: to enable the external scientific community to carry out high-impact nanoscience research through an open, peer-reviewed user program, and to conduct in-house research to understand and control the complexity of electronic, ionic and molecular behavior at the nanoscale to enable the design of new functional nanomaterials. These two aspects of the CNMS's mission are closely linked and mutually benefit from each other. In particular, the partnering of key groups of users who bring outside expertise with the sustained scientific in-house effort allows the center to be a leading effort in the development of new tools and methods for nanoscience, including synthesis, theory/modeling, and characterization.

All researchers at the CNMS work with users and perform their own in-house research within the framework of the DOE-reviewed research project. The CNMS conducts in-house research *to understand and control the complexity of electronic, ionic and molecular behavior at the nanoscale to enable the design of new functional nanomaterials*, with the overall goal of harnessing energy through nanoscience. The scope of the in-house research (also referred to as staff science or theme science) is summarized by three research themes:

- The **Electronic and Ionic Functionality on the Nanoscale (EIFN)** theme *seeks to understand the link connecting the atomic scale physics of electronic and ionic transport with chemistry and electrochemistry*. In this theme, electronic and ionic material functionalities are examined on the atomic scale using advanced modalities of scanning probe microscopies and *in-situ/operando* scanning

transmission electron microscopy. The knowledge acquired is extended to the emergent behaviors at the scales of individual nanoparticles and defects and finally to the macroscale, where function can be translated into new technologies. A central thrust is to characterize and control fundamental mechanisms of coupling between electronic and ionic functionalities that underpin electrocatalysis and energy storage and conversion.

- The **Functional Polymer and Hybrid Architectures (FPHA)** theme focuses on *understanding, designing, and manipulating the multiscale self-assembly of macromolecular and hybrid materials to tailor electronic transport and response*. Understanding the chemical and physical mechanisms of the self-assembly of macromolecular and hybrid materials is essential for designing new materials with specifically designed functionalities for applications where materials with vastly improved functionalities are required to address needs for future energy technologies, such as improved photovoltaics, battery separators, etc. FPHA research includes the study of the role of non-covalent interactions in the self-assembly of energy-responsive macromolecular systems; investigation of the role of interfacial interactions between organic components and substrates in directing self-assembly and subsequent impact on optoelectronic properties; and finally the use of predictive theory and simulation to decipher how the chemical and physical information thus encoded at the nanoscale in our targeted polymer and hybrid architectures translates into structure, dynamics and function at meso- and macro-scales.
- The **Collective Phenomena in Nanophases (CPN)** theme *seeks to understand and control the collective behaviors of electrons, ions and molecules at the nanoscale to enable the design of new functional materials*. In the environments of future energy systems—from dimensionally confined semiconductor materials for photovoltaics to nanoporous supercapacitor and battery electrodes—this requires understanding the behaviors that emerge when confinement and crowding have forced correlations between the electrons, ions, and molecules that store, transport, and release energy. This work is carried out in two specific aims: (1) understanding how atomic scale structure, nanoscale confinement, and quantum mechanical effects impact electronic processes within nanostructures and across interfaces; and (2) understanding how correlations induced or enhanced by confinement and crowding lead to collective behaviors in chemical transport and reactivity. The first specific aim focuses on the need to develop a fundamental understanding of how the properties of 2D materials and interfaces are influenced by confinement, shape, defects, and composition by using theoretical modeling and simulations in concert with experimental synthesis and characterization. This aim is differentiated from the FPHA theme by its emphasis on the building up layers of materials to investigate how materials properties are modified by out-of-plane interactions during the transition from a single 2D layer toward a 3D bulk, one layer at a time. Additionally, considerable focus is on layered materials that intrinsically have strong electronic correlation, such as high-temperature (high-$T_c$) superconductors.

## 3.2 Collaborators

The CNMS benefits from an intrinsically strong interaction with Oak Ridge National Laboratory (ORNL) signature strengths in multiple areas and takes advantage of the distinctive capabilities of other DOE user facilities at ORNL, including the Oak Ridge Leadership Computing Facility (OLCF), the Spallation Neutron Source (SNS), and the High Flux Isotope Reactor (HFIR). In particular, the CNMS emphasizes a strong link to neutron sciences, providing an environment for researchers to integrate neutron studies into nanoscience efforts. The CNMS uses its expertise in materials sciences (including polymer synthesis) and computational sciences towards the incorporation and development of materials-by-design approaches, and seeks new advances in imaging sciences that build on ORNL's demonstrated leadership in scanning probes, scanning transmission electron microscopy, Helium ion microscopy, and atom-probe tomography.

Each year the CNMS supports approximately 600 unique users from more than 100 different institutions spanning academia to industry, and from around the world (see Figure 3.1) The CNMS user community is diverse, ranging from students who work closely with CNMS staff, learning unique skills from experts

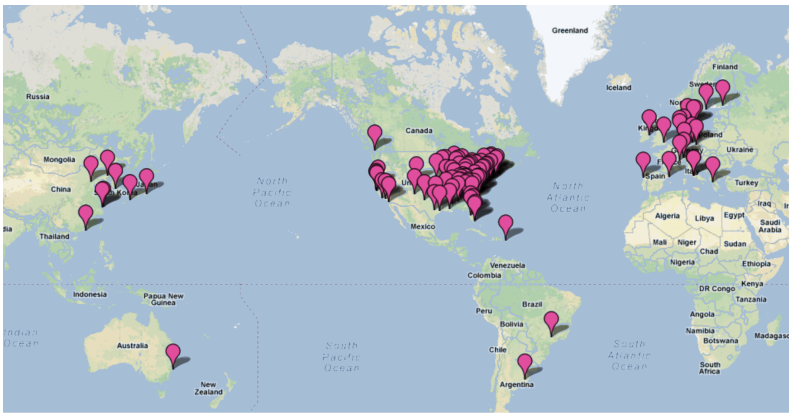



*Figure 3.1: Origins of the various user projects are noted with red markers. CNMS supports approximately 600 different users from over 100 different institutions, globally.*

and gaining access to cutting-edge instrumentation as they advance their research, to "partner users" who collaborate with staff to develop new capabilities and instruments that are then made available to the broad CNMS user community. About 10% of all users are theory users, who work with the staff of the CNMS Nanomaterials Theory Institute to gain access to expertise and computational resources. About 40% perform synthesis and/or nanofabrication and typically use a broad range of characterization tools to verify the quality of the synthesized materials or to investigate their novel properties. The remaining 50% of the users come to the CNMS specifically for characterization, using the broad range of tools including microscopy (electron, He-ion, scanning probe), atom-probe tomography, X-ray diffraction, optical spectroscopies, etc.

## 3.3 Near-term Local Science Drivers

### 3.3.1 Instruments and Facilities

The CNMS incorporates a substantial theory, modeling and simulation (TMS) effort in the form of the Nanomaterials Theory Institute (NTI) that is cross matrixed with the Computer Science and Mathematics Division. The NTI works primarily on atomistic simulations at the level of electronic structure and molecular dynamics, with some developments towards coarse-grained modeling for larger molecular-based systems. The calculations support and challenge experimental work in polymer synthesis, metal oxide semiconductor physics, superconductivity, carbon nano-architectures, electrochemistry, catalysis, photovoltaic, optoelectronic, and electronic device fabrication. This also includes modeling for the interpretation of neutron scattering experiments and hence provides a cross synergy with the needs of the SNS in this domain. The NTI also supports significant numbers of nanoscience user projects that require computational support.

Some of these calculations are utilizing tightly coupled codes (i.e., which do not scale well onto very large parallel formats) and ideally need to be carried out repeatedly over a wide range of parameter space (atomic or molecular constituency and arrangement, temperature, pressure, etc). Hence a key need is large throughput on medium-sized compute clusters for which the CNMS maintains. The independent data output from these calculations is usually not large but due to the need to run many simulations, it can become quite substantial (10-100 TB). Overall, it is what is learned from the highly condensed output (e.g., an electronic energy or a rate constant) that informs experiment that is the critical outcome.

**Computer Facilities**: ORNL operates the Oak Ridge Leadership Computing Facility (OLCF) which manages the computing program at ORNL for the Department of Energy. OLCF 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 OLCF staff provides contin-

uous 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:

- *Titan* is a hybrid-architecture Cray XK7 system with a theoretical peak performance exceeding 27,000 trillion calculations per second (27 petaflops). It contains both advanced 16-core AMD Opteron central processing units (CPUs) and unconventional NIVIDIA Kepler graphics processing units (GPUs). Titan features 18,688 compute nodes, a total system memory of 710 TB, and Cray's high-performance Gemini network. Its 299,008 CPU cores guide simulations while the accompanying GPUs that can handle hundreds of calculations simultaneously. The system provides decreased time to solution, increased complexity of models, and greater realism in simulations.
- *Eos* is a 744-node Cray XC30 cluster with a total of 47.6 TB of memory. The processor is the Intel® Xeon® E5-2670. Eos uses Cray's Aries interconnect in a network topology called Dragonfly. Aries provides a higher bandwidth and lower latency interconnect than Gemini. Support for I/O on Eos is provided by 16 I/O service nodes. The system has two external login nodes. The compute nodes are organized in blades. Each blade contains four nodes connected to a single Aries interconnect. Every node has 64 GB of DDR3 SDRAM and two sockets with eight physical cores each. Intel's Hyper-threading (HT) technology allows each physical core to work as two logical cores so each node can functions as if it has 32 cores. Each of the two logical cores can store a program state, but they share most of their execution resources. Each application should be tested to see how HT impacts performance before HT is used. The best candidates for a performance boost with HT are codes that are heavily memory-bound. The default setting on Eos is to execute without HT, so users must invoke HT with the -j2 option to aprun. In total, the Eos compute partition contains 11,904 traditional processor cores (23,808 logical cores with Intel Hyper-Threading enabled), and 47.6 TB of memory.
- *Rhea* is a 196-node commodity-type Linux cluster. The primary purpose of Rhea is to provide a conduit for large-scale scientific discovery via pre- and post-processing of simulation data generated on Titan. Users with accounts on INCITE- or ALCC-supported projects are automatically given an account on Rhea. Director's Discretion (DD) projects may request access to Rhea. Each of Rhea's nodes contain two 8-core 2.0 GHz Intel Xeon processors with Hyper-Threading and 64GB of main memory. Rhea is connected to the OLCF's 32PB high-performance Lustre filesystem "Atlas."

At ORNL there is also a substantial organization institutional cluster. The ORNL Institutional Cluster (OIC) has come together in five phases. Specifically for the CNMS, the 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.
- *Phase 3* consists of 2x42 nodes each with two Quad Core 2.26 GHz Intel Xeon E5520 (Nahelem-EP) processors with 24 GB 1066 MHz DDR3 memory per node and 500GB of local disk. These

units are an Intel/Linux/PBS/MPI/Infiniband system with 336 processor cores total, 3GB memory/core, delivering nearly 3 million CPU hours per year.

- *Phase 5* consists of 28 Compute Nodes: Each with two Interlagos Opterons 6274 with 16 cores 2.2GHz 16MB L3 Cache, 128GB RAM (DDR3 Registered, Dual Rank, 1333MHz, 1.5v), 1TB HDD. This system has a total of 896 cores, 3.5 TB of memory (4 GB/core of RAM) and approximately 7.89 teraflops (on the compute nodes) theoretical.

**Network Connectivity**. Primary external connectivity for ORNL is provided by ESnet. ORNL has redundant connections to ESnet, which maintains a hub at ORNL with primary connectivity via a 100G circuit to the ESnet Atlanta hub. Secondary connectivity for the ESnet hub located at ORNL is provided by ORNL on a site owned optical DWDM system and consists of dual 10G circuits to Nashville. ORNL also has 10G connectivity to the Southern Crossroads (SoX) in Atlanta, with commodity peering to Hurricane Electric. These connections into ORNL provide access to major research and education networks around the world.

The ORNL campus network 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 firewalls and/or access control lists and network intrusion detection. 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.

**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 1TB/sec of aggregate performance, scalability to over 26,000 file system clients, and storage capacity of over 32 PB, Spider is the world's largest scale Lustre file system. Spider consists of 48 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 TB of system memory. Metadata is stored on two LSI Engino 7900s (XBB2) and is served by three 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 three SL8500 tape libraries, holding up to 10,000 cartridges each and is in the process of deploying a fourth SL8500 this year. The libraries house a total of twenty-four T10K-A tape drives (500 GB cartridges, uncompressed) and thirty-two T-10K-B tape drives (1 TB 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.

**Compute and Data Environment for Science (CADES)**. ORNL hosts a number of signature facilities and projects that produce, analyze, and steward a wide diversity of data. These include the Spallation Neutron Source (SNS), the Oak Ridge Leadership Computing Facility (OLCF), the Center for Nanophase Materials (CNMS), the Consortium for Advanced Simulation of Light Water Reactors (CASL), the Atmospheric Radiation Measurement (ARM) archive, the Carbon Dioxide Information Analysis Center (CDIAC) and other growing initiatives such as our work with the Center for Medicare and Medicaid Services (CMS). In addition to these facilities and projects, ORNL has a wide range of other research projects that are increasingly reliant on computing and data-intensive capabilities. While each of these centers and projects has a unique mission, they share common requirements and needs in computing, data processing, and management. Perhaps even more broadly, these initiatives share common needs in computing and data intensive capabilities that can be more efficient and cost effective if they leverage common technologies and expertise in computing and data science. These capabilities include data analysis, data fusion, data mining, search and discovery, visualization and many others. CADES provides a new local environment for scientific discovery enabling scientists to free themselves from

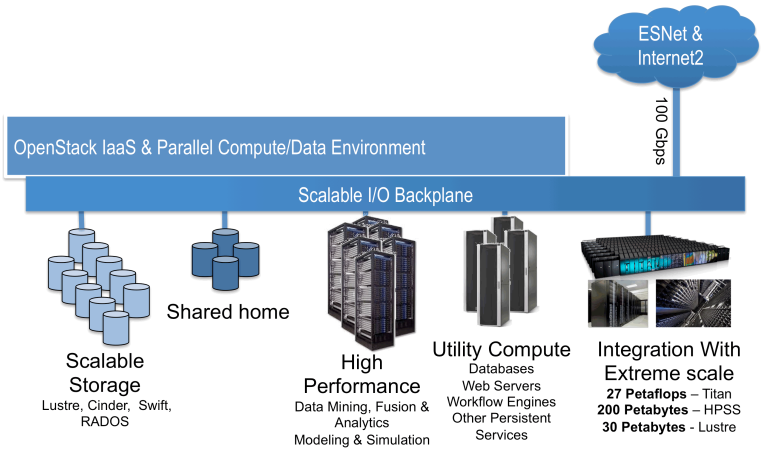



*Figure 3.2: Illustration of the data-intensive storage backplane implemented at the Compute and Data Environment for Science (CADES) at ORNL.*

the difficulties of trying to manage, manipulate and process large data sets in order to concentrate on extracting the scientific meaning of the theoretical, experimental, and observational data.

**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) consists of two tiled display walls: the primary display wall spans 30.5' x 8.5' and consists of eighteen 1920 × 1080 stereoscopic Barco projection displays arranged in a 6 x 3 configuration. The secondary display wall consists sixteen 1920×1080 planar displays arrange in a 4 x 4 configuration providing a standard 16:9 aspect ratio. Three computing systems are currently provided in the laboratory. These consist of a distributed memory Linux cluster, a shared memory Linux node, and a shared memory Windows node.

The EVEREST lab is a venue that serves both as a visualization center and a place for scientists to meet, hold discussions, and present their work.

### 3.3.2 Software Infrastructure

A wide variety of software and algorithms are currently used in the various compute and data-intensive projects. This software ranges from custom-built algorithms and libraries for specific needs; ORNL-packaged software such as ADIOS, EDEN, ESGF, Mercury, Piranha and VisIt; community-developed software such as Git, Hive, Hadoop, and Boost; to commercial software such as Matlab, IDL, and Greenplum. These software packages address requirements such as analysis, data mining, semantic analysis, natural language processing, fusion, indexing, visualization and visual analytics, discovery/dissemination, federation, intelligent user interfaces, and software development lifecycle management.

Via CADES, a data-intensive storage backplane (see Figure 3.2):

- Accessible via a variety of protocols such as: NFS, SMB/CIFS, Web, Globus Online, GridFTP, BBCP, FTP, SCP, among others.

Computational materials and chemical science codes (bold indicates the codes involve NTI development) include:

- **DMRG++**, **DCA+**, **QMCPACK**, **SCFT**, Quantum Espresso, Siesta, Abinit, VASP, NWChem, PSI4, DFTB, LAMMPS, and GROMACS

### 3.3.3 Process of Science

Scientific understanding enabled through the careful design of experiments or computational simulations for the collection and analysis of key data, is one of the primary products of user facilities. Data movement and data processing are critical for scientists and researchers, particularly once they have performed their experiments or simulations. For work at the CNMS, the process can be characterized as listed below.

- Experimentally measuring and computationally simulating materials provide the major means for collecting scientific data and is the primary activity of users; while at the CNMS, the measurement techniques vary widely across the suite of instruments and theoretical/computational approaches, but they have some common steps involving data acquisition and analysis as listed below:
  - Data movement, cataloging, and archiving
  - Data analysis and visualization
  - Fitting to models and analysis of the processed data for structure, function or dynamics information

Whereas computational resources are ephemeral, data is enduring, requiring long-lived allocations of storage resources. Public policy is further motivating this demand, requiring open access to data generated as part of federally funded research. The current data management policy for the CNMS is stated in Table 3.1.

## 3.4 Near-term Remote Science Drivers

### 3.4.1 Instruments and Facilities

CNMS has numerous remote users. These users are primarily cyber-based and connect to the internal compute cluster or LCF using ssh combined with a SecureID and a PIN. Each user has an account on one of the compute systems and accesses the compiled codes, data archives, and necessary math libraries directly. Data from the calculations are typically analyzed and stored on the local compute systems, often for the duration of the project, and then transferred to the users home. The size of the data/results collected varies from projec to project and the compute system (hundreds of gigabytes to hundreds of terabytes). Online data analysis and reduction are important after which GridFTP can be effective for much of the needed data movement between computing resources.

### 3.4.2 Software Infrastructure

Online data analysis and reduction are important and much of this is accomplished within the particular application software. Other components may involve post-processing via VisIt, VMD, VESTA, Matlab, P4Vasp, etc. Typical data formats are ASCII, XML, HDF5, binary, lossless compressed, etc. GridFTP can be effective for much of the needed data movement between computing resources, especially in the post-processed form. Others modes such as NFS, SMB/CIFS, Globus, BBCP, FTP, SCP, can also be useful and are utilized.

### 3.4.3 Process of Science

The size of the data/results produced varies from project to project and the compute system (hundreds of gigabytes to hundreds of terabytes). Online data analysis and reduction are important after which GridFTP can be effective for much of the needed data movement between computing resources. Others modes such as NFS, SMB/CIFS, Globus, BBCP, FTP, SCP, can also be useful. Typical workflow:

*Table 3.1: The guidelines below describe CNMS Data Management procedures, processes and resources that need to be understood by both user projects and in-house research.*

| **CNMS Data Management Policy** |
| --- |
| 1. Limited data management resources. The CNMS has limited data storage resources and storage capacity varies depending upon the instruments used. User data stored at CNMS will only be retained up to three months past the termination date of a user project. There is no lifetime retention. Users are expected to migrate their data out of CNMS prior to the expiration date of their project. Beyond the three-month grace period, the CNMS reserves the right to delete all user _les and directories across all applicable storage media.<br>2. Users must not rely upon data storage and retention at CNMS. Users must not rely upon data storage and retention at the CNMS to meet funding agency requirements for a Data Management Plan.<br>3. The standard User Agreement, approved by DOE, defines ownership of technical data.<br>4. Published data displayed in figures. Researchers are encouraged to make publicly available in tabular form any published data displayed in figures. This requirement could be met by including the data as supplementary information to the published article or through other means. The published article should indicate how these data can be accessed, or researchers could make this information available in their research webpages.<br>5. Release of computer source code. When appropriate, researchers are encouraged to make publicly available the full release of all computer source code used to produce the published results along with self-tests to build confidence in the quality of the delivered code. [see, D. C. Ince, L. Hatton, and J. Graham-Cumming, “The case for open computer programs,” *Nature*, **482**, 485-488 (2012)].<br>6. Instructions explaining how to reproduce relevant data. Researchers are encouraged to include step-by-step instructions explaining how to reproduce all relevant data published. This requirement could be met by including the instructions as supplementary information to the published article or through other means. The published article should indicate how these data can be accessed, or researchers could make this information available in their research webpages.<br>7. These guidelines are subject to change without notice. The latest version will be available at http://www.cnms.ornl.gov/user/Data_Management_Policy.pdf. |

- Collecting scientific data via experiments and computational simulations– the measurement and computational simulation techniques vary widely across the suite of instruments and theoretical/computational approaches, but they have some common steps involving data as listed below:
  - Data acquisition
  - Data movement, cataloging, and archiving
  - Data analysis and visualization
  - Fitting to models and analysis of the processed data for structure, function or dynamics information

Many scientific domains are increasingly dependent on the ability to efficiently capture, integrate, analyze, and steward large volumes of diverse data. In materials science, understanding and ultimately designing advanced new materials with complex properties will require the ability to integrate and analyze data from multiple instruments designed to probe complementary ranges of space, time, and energy. Chemical imaging, another crucial contributor to the design of new materials, brings other challenges as many of the imaging techniques are destructive. These techniques require near real time analysis to determine structural evolution properties of the material while there is still significant material left to image. These and many other scientific pursuits require data science resources that are often distinct from, but complementary to, the computational science resources provided by traditional HPC facilities.

As an example, the Imaging and Nanoscale Characterization group at the CNMS is actively working on the development of real-space, non-destructive scanning probe microscopy techniques for probing bias-induced (ferroelectric polarization switching, electrochemical reactions) and thermal (glass transition, melting) transformations on the nanosocale. In these experiments, the scanning probe microscopy (SPM) tip focuses an electric or thermal field in a small, 5–30 nm region of material, inducing local transformations. In parallel, measured dynamic strain, resonance frequency shift, or quality factor of the cantilever (piezoresponse force microscopy, electrochemical strain microscopy) or tip-surface current (conductive AFM) provides information on processes in the material (polarization, domain size, ionic motion, second phase formation, melting) induced by local stimulus. In the future, the detection strategies can include microwave, Raman, focused X-ray, electron microscopy, and other high-bandwidth local ( 10 nm and below) structural and chemical probes. The uniqueness of this approach is that transformation can be probed in material volumes containing no or single individual extended defects, paving a pathway for studying phase transformations and electrochemical reactions on a single defect level (as opposed to volume averaging for typical materials science methods; compare to the impact of molecular unfolding spectroscopy in biomolecular chemistry), the target of crucial importance for material science to link defect structure to its functionality.

The hardware platforms for these studies can be realized on 30,000+ SPMs worldwide! This necessitates a classical development path of minimizing by noise level, improving drift stability, and introducing proper chemical and thermal environments. However, these studies require drastic improvement in capability to collect and analyze multidimensional data sets, well beyond state of the art (2D imaging or 3D spectroscopic imaging) in the field. This can be demonstrated as follows:

- The spatial scanning necessitates data acquisition over 2D dense grid of points
- The probing local transformation requires sweeping local stimulus (tip bias or temperature) while measuring the response
- All first-order phase transitions are hysteretic and hence are history dependent. This necessitates first-order reversal curve type studies, effectively increasing dimensionality of the data (e.g., probing Preisach densities)
- First-order phase transition often possess slow-time dynamics, necessitating probing kinetic hysteresis (and differentiating it from thermodynamics) by measuring response as a function of time
- The detection of force-based SPMs necessitates probing response in a frequency band around resonance (since resonant frequency can be position dependent and single-frequency methods fail to capture these changes).

*Table 3.2: Development of multidimensional SPM methods at CNMS.*
**Dimensionality is given as (space x space) x frequency x (parameters). Note that the signal can be multimodal (e.g. collect phase and amplitude of response or three vector component of the signal). Highest number of measured variables to date is 8 (phase and amplitude in the on/off state for vertical and lateral signal). The collection of multimodal data multiplies file size by $\frac{N}{2}$.*
***Not realized yet due data acquisition and processing limitations, but is the ultimate goal for data acquisition and analysis developments.*
****Current data acquisition times are limited by the eigen-frequency of the cantilever in the contact mode. However, expected introduction of fast DAQ electronics and small cantilevers is expected to push these by about a factor of 10 in next 2–4 years.*
*****Applications for ferroelectric, electrochemical, and biological/macromolecular systems.*

| Technique | Dimensionality | Current data set* | Target data set*** |
|---|---|---|---|
| Band Excitation (BE) | 3D, space and $\omega$ | (256 x 256) x 64, 32 MB | (512 x 512) x 64, 128 MB |
| Switching spectroscopy PFM | 3D, space and voltage | (64 x 64) x 128, 4 MB | (128 x 128) x 256, 32 MB |
| Time relaxation | PFM 3D, space and time | (64 x 64) x 128, 4 MB | (128 x 128) x 256, 32 MB |
| AC sweeps | 4D, space, voltage | (64 x 64) x 64 x 256, 512 MB | (128 x 128) x 64 x 256, 2 GB |
| BE SSPFM | 4D, space, $\omega$, voltage | (64 x 64) x 64 x 128, 256 MB | (128 x 128) x 64 x 256, 2 GB |
| BE thermal | 4D, space, $\omega$, Temp | (64 x 64) x 64 x 256, 512 MB | (128 x 128) x 64 x 256, 1 GB |
| Time relaxation BE | 4D, space, $\omega$, time | (64 x 64) x 64 x 64, 4 MB | (128 x 128) x 64 x 512, 128 MB |
| First order reversal curves (FORC) | 5D, space, $\omega$, voltage, voltage | (64 x 64) x 64 x 64 x 16, 2 GB | (128 x 128) x 64 x 64 x 64, 32 GB |
| Time relaxation on sweep, BE | 5D, space, $\omega$, voltage, time | (64 x 64) x 64 x 64 x 64, 8 GB | (128 x 128) x 64 x 64 x 256, 64 GB |
| FORC Time BE** | 6D, space, $\omega$, voltage, voltage, time | | (64 x 64) x 64 x 64 x 16 X 64, 128 GB |

These simple physical arguments illustrate that complete probing of local transformations requires a 6D (space x frequency x (stimulus x stimulus) x time) detection scheme, as compared to 1D molecular unfolding spectroscopy. To date, we have realized 5D detection schemes (first order reversal curves, time relaxation within hysteresis loop methods). The development of these techniques is illustrated in Figure 3.3. **It is important to note that the acquisition of these compound data sets brings the obvious challenge of data storage, dimensionality reduction, visualization, and interpretation.**

### 3.4.4 Software Infrastructure

Local resources such as CADES (see Section 3.3.1) for enhanced data storage and near real-time data analysis capabilities, especially for experimental techniques like the imaging case mentioned above (Section 3.4.1). We need to bring data analysis and reduction to the experimental instruments to minimize the time-to-results cycle. We need to have low latency, high-bandwidth dedicated data portals for the user centers or “computational end-stations.”

Hardware and software will continue to advance, requiring the infrastructure to remain flexible such that these technological advances can readily be adopted. The infrastructure must allow adoption of hardware advances such as the transition from homogeneous CPU technologies to heterogeneous platforms

that incorporate CPU cores optimized for single-thread performance along with cores optimized for multi-threaded performance. Advances in high-performance networking technologies such as InfiniBand or Data-Center Ethernet must be easily deployed within the infrastructure when compute and data intensive initiatives will benefit from these advances. Adoption of these hardware advances will require a flexible software management infrastructure allowing various levels of the software stack to be adapted to these new hardware technologies. This flexibility must span operating systems, middleware, and even applications. Similarly, the software management infrastructure must be flexible to new advances in software technologies that may or may not be tied to specific hardware architectures.

### 3.4.5 Process of Science

Computational resources will continue to grow in size/speed, and as such the amount of data generated and the demand to perform analysis and data transfer will also increase. A similar statement holds for experimental approaches, including those that have not traditionally been thought as within the big data area, e.g., imaging (see example in Section 3.4.1). Finally, we expect that it will become more common to find simultaneous users at multiple facilities and thus the need to coordinate improved capabilities for theory/simulation guided experiments, data movement, fusion, and efficient analysis.

## 3.5 Medium-term Remote Science Drivers

### 3.5.1 Instruments and Facilities

While virtually all projects (computational and experimental) require scalable compute and relatively large on-line storage capacities for specific workloads, these workloads are typically transient in nature. This transient property can be leveraged to provide elasticity in both compute and storage for these workloads. Whereas in the past a compute and storage resource would have to be designed to accommodate the most demanding requirements of an initiative, elasticity may allow these data initiatives to only use the "*average*" requirement of storage and compute.

Consider the workflow illustrated in Figure 3.3; data ingress from a large-scale simulation or experiment followed by data analysis to find features of interest as illustrated by the red arrow in the figure. During this workflow the variation in processing resources, both in terms of compute and storage, can be substantial. Ingress of data from a simulation or experiment can require significant high-performance "online" storage capacity and substantial compute resources for subsequent data analysis. These resource requirements are point-in-time and, once data analysis is complete, both the compute and "online" storage resource are no longer needed as the data is moved to archival storage. Other examples of this variation in processing requirements are commonly found in data mining in which data is staged from archive into online storage and compute resources are then used to mine these datasets for correlations and other features of interest. In both of these examples and in many others representing a broad set of workflows in compute and data-intensive initiatives, an elastic infrastructure is a requirement.

### 3.5.2 Software Infrastructure

A number of technologies such as high-performance parallel file systems, scalable archival storage systems, parallel data analysis tools, metadata harvesting tools, and distributed data management technologies have been developed and deployed. While these technologies have addressed many existing requirements, future requirements may require improved capabilities. As increases in computational capabilities lead to increases in data volumes and velocity and the drive to couple simulation with experimental results and observations, this will lead to increasing data diversity as well. There will be a need to develop fundamentally new approaches to data capture, reduction, analysis, and management.

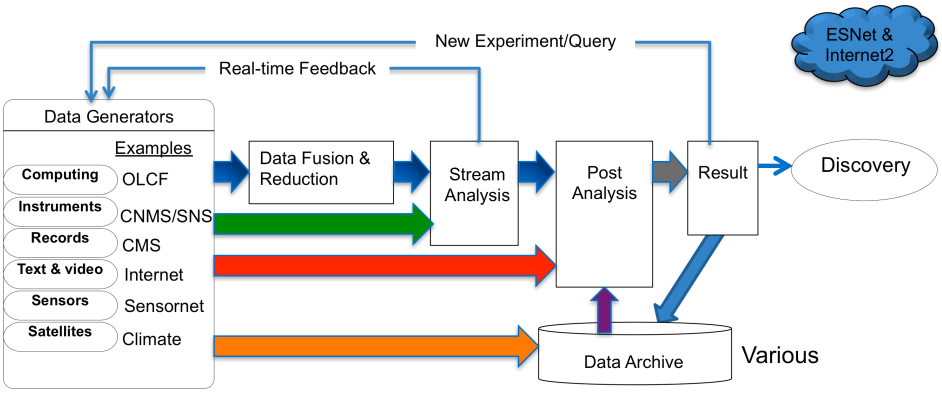



*Figure 3.3: Illustration of the typical scenarios for data workflow from large data generators (shown in the box). The data workflow can range from simple data archiving (bottom orange arrow), to post data analysis (middle red arrow), to data stream analysis (green arrow). More recently, the data workflow has evolved toward requiring real-time data analysis and reduction to provide guided feedback (top blue arrows).*

There will still be continued use of GridFTP, NFS, SMB/CIFS, BBCP for data movement. Traditional simulation services such as simulation frameworks, scientific libraries, and scalable debuggers, will be complemented by data and analytic services such as data mining and data fusion services, data transfer tools, metadata harvesting and management tools, and data discovery and dissemination services. As these services are often composed into complex workflows that incorporate one or more underlying services, higher-level workflow composition and execution services are often employed, such as the Adaptable I/O systems.

In addition to supporting a variety of workflows, the infrastructure must be flexible to support varying security and data policy requirements. This infrastructure must support data that is open-science with few if any restrictions governing access, and at the same time support data with far greater restrictions such as proprietary information as part of an industrial partnership or information that may be classified in nature. In some cases, differing security requirements may simply necessitate differing controls for access and use of the data while sharing the same physical infrastructure. In other cases, elements of the infrastructure may require isolation for compliance. Beyond diverse security and data-access requirements, projects may have differing requirements in data-stewardship and data-retention policies. As an example, most of the simulation platforms currently managed require users to move their datasets off of high-performance file systems almost immediately (two weeks). Although on the compute cluster at CNMS the data management policy is more dilated (see Table 3.1), future demand and capacity may require moving toward the shorter times on storage. Currently on the OLCF, users can move datasets from these file systems to longer-term archival storage, but even in archival; data retention is limited to the life of the project, often just a few years. This is found in constrast with initiatives such as the ARM and CDIAC archives where datasets are held indefinitely. In the future, the underlying infrastructure will also need to have the flexibility to support the diversity of these and many other requirements in security and data-policy. Over time, as common data workflows become more widespread, the software infrastructure will need to allow readily deploying and supporting these common workflows.

### 3.5.3 Process of Science

The trend toward larger numbers of remote and many-facility users will continue. This will demand improved integration and coordination, both in terms of workflows but also for data analysis, data management and data transfers. Simultaneous, multi-modal analysis will become more common, particularly in terms of computational modeling/simulation and experimental study, and also in terms of coordinating multiple modes of experimental interrogation such as X-rays, neutrons, and materials imaging.

While advances in experimental facilities and computational techniques continue, the deluge of data

generated from these techniques represents both a big data challenge and a big data opportunity. For example, instruments at the SNS can now generate millions of neutron events per second and rather than simply collapsing these events statistically, new experiments require the ability to analyze each individual event. Experiments at NSRCs, such as at the CNMS, use multiple real-space probing techniques to observe building blocks at different time and length scales requiring advances in multi-modal data-driven analysis techniques to analyze these high-dimensional datasets (these also include dynamic STEM and TEM capable of generating many tens of terabytes; with 600 instruments worldwide, and increasing by 50 instruments/year). The trend is toward significantly increased resolution and flux that is capable of generating multiple petabytes of data per week and this will require commensurate advances in data management, movement, and analysis techniques. Computational materials design approaches will continue to increase, employing high-throughput parameter space studies to identify candidate materials prior to synthesis. Such studies can generate vast knowledge databases of candidate materials and multiple petabytes of data, requiring advances in data management and parameter space exploration techniques.

## 3.6 Beyond 5 Years

Over the next five or more years, the size of simulation and experimental data sets will continue to increase. This will be a result from both the deployment of enhanced computational resources and from 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 light source facilities to be heavily used and likewise demand a greater need for networking and data storage. The photon and neutron sources now provide considerably higher intensities to individual beamlines and future plans will further enhance this capability. Coupled with corresponding advances in detector technology, this will continue to result in unprecedented rates of data collection in the experiments. The sheer volume and velocity of data from individual experiments/calculations will also increase, as scientists are able to probe increasingly complex questions that probe more subtle properties and phenomena. In the future, bringing all types of disparate data to bear on a particular discovery process or mission outcome could be a new frontier of science and technology.

In the rapid incubation of data-intensive science, the scientific community is attempting to develop technical sophistication and mature code bases on an aggressive schedule. By comparison, this work happened over the course of 10–20 years in parallel computing, while there appears to be the expectation for an accelerated similar process occurring in data-intensive computing over the course of only about 5 years.

## 3.7 Network and Data Architecture

The computational capabilities of the Leadership Computing Facilities (LCFs) will increase 1000-fold over the next decade. The research, development and integration of the resources required to manage, analyze and disseminate data generated at the LCFs will require a comprehensive data and network strategy. This strategy must encompass research and development of new data technologies, industry engagement to productize these technologies, and an organizational structure that is prepared to develop, deploy, and manage these resources in the most cost effective manner. Many of the challenges faced with respect to data are common across a number of other DOE facilities such as the neutron and light sources, the Nanoscale Science Research Centers (CNMS, CNM, CINT, CFN, and Molecular Foundry), as well as many of the Energy Frontier Research Centers (EFRCs) and Innovate Energy Hubs.

Data-intensive scientific studies will continue to become more prominent and will increasingly require high performance computing and data infrastructure, sophisticated data analysis and visualization, and

comprehensive data management.

Clearly high-performance data transfer and big data analysis will be important aspects for the next 5–10 years. Also the work within the national Materials Genome Initative (MGI) will continue to develop large-scale knowledge databases which will enable new capabilities to perform materials design.

## 3.8 Data, Workflow, Middleware Tools and Services

Some applications on the LCFs already routinely create tens of terabytes of data in a single application run and it is not uncommon for a single application run to generate in excess of 100 TB. In addition to large volumes of data, applications can create millions of granules (files) in a single application run—at several hundreds of thousands threads of execution each generating a file per I/O epoch. In this environment, many users are already confronted with extreme data-management challenges. On smaller compute clusters, users also are generating tens of terabytes of data during a project and these data sizes will continue to rapidly increase in the next 5–10 years.

To meet the challenges posed by increased capabilities to produce large amounts of data, a number of technologies such as high-performance parallel file systems, scalable archival storage systems, parallel data analysis tools, metadata harvesting tools, distributed data management and data transfer technologies have been developed and deployed. While these technologies address many existing requirements, future requirements will require completely new capabilities. As continued increases in computational capabilities lead to dramatic increases in data volumes and the drive to couple simulation with multi-modal experimental results and observations leads to increasing data diversity—fundamentally new approaches to data capture, reduction, analysis, and management will be required.

In addition to traditional HPC software packages such as Globus services, HDF5, NetCDF, Glean, ADIOS, VisIt, and ParaView, centers will need to provide new software and services such as Mongo, Hadoop, and AMQP services. Perhaps these services may be provided as a low-level provisioned infrastructure or a higher-level service designed to address fault-tolerance and scalability.

As we move towards Exascale systems in which total concurrency is anticipated increase by a factor of 4000 and total system memory is anticipated to increase by a factor of 200, required storage capacities will approach an exabyte (EB) with an I/O bandwidth requirement of 60 TB/sec. Meeting these challenges using traditional disk-based storage technologies would require 200,000 disk drives even under the scenario that bandwidths and latency improvements continue to trend as they have in the past. It is clear that fundamental research and development of new storage technologies and data transfer will be required. In a similar vein, it is clear that existing approaches to storage system access, primarily POSIX file system interfaces, will face significant challenges in scaling to future systems. In this regard, an example of some of the success enjoyed by industry, in 2010 the largest Bigtable instance (a no-SQL store) at Google already exceeded 70 PB.

As common software stacks are identified and end-to-end workflow patterns emerge, facilities will need to operationalize and support key services for data intensive science communities.

*Table 3.3: The following table summarizes data needs and networking demands for a few key instrument types of CNMS.*

| 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 (quantum and classical; atomistic and mesoscale), mean field approaches, statistical simulations on large computer clusters and capability computing 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. | · Simulations can generate several hundred GB of data (composition: simulation results, input files, and restart files; ~5 files total). Restart files, production data/results (10-100GB) depending on the specific type of calculation: | | |
| · Quantum transport in nanostructures and virtual STM imaging. | · Users will generally desire to transfer data files back to their home institutions for future reference, analysis, etc. | · **Case 1**: CCSDT/aug-cc-pvtz for a heteroatom supramolecular complex of 60-100 atoms (100-500 GB) | · Case 1: ~10 minutes (1 per day per project) | · Case 1: ~10 minutes (a couple per week per user project) |
| · Self-assembly of nanostructured materials. | · Memory ranges from several hundreds of GB on clusters to tens to hundreds of TB on capability machines. | · **Case 2**: molecular dynamics for a million atom molecular system with a trajectory dump for the (**p**,**q**) every 10 fs over each trajectory spanning times of 100 ps - 100 ns (100GB-1TB) | · Case 2: ~10 minutes – 2 hrs (only a few per quarter) | · Case 2: similar data traffic as LAN, longer times for transfer; depending on site of transfer |
| · Statistical physics in non-equilibrium systems. | | · **Case 3**: plane wave periodic electronic structure calculations for heteroatom supercells of 100-500 atoms (10-50 GB) | · Case 3: a few minutes (10 per day) | · Case 3: similar data traffic and transfer times as LAN |
| · Emergent behavior in strongly correlated electronic systems. | | · **Case 4**: DCA++ calculations (5-10 GB) | · Case 4: few minutes (2 per week/project) | · Case 4: Same as LAN |
| · Theory/simulation-driven materials by design (high-throughput screening). | | · **Case 5**: QMC for 100 + atom systems (10-100 GB) | · Case 5: ~10 minutes (1-2 per day/project) | · Case 5: Same as LAN |
| | | · **Case 6**: DMRG calculations on GMR (1-5 GB) | · Case 6: a few minutes (1 or 2 per week/project) | · Case 6: Same as LAN |
| | | · **Case 7**: experimental data from scanning probe microscopies, 5-200 GB per sample (5-10 samples/day); STEM,TEM up to 1 TB/day | · Case 7: a potential growth area; up to 10/day | · Case 7: Same as LAN |
| | | · **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 systems and/or larger basis sets; case 2 longer time scales; case 3 larger supercell sizes; case 4 enhanced details for models of inhomogeneities and multi-orbital Hubbard models, 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. | · All seven: Significant adoption of high-throughput simulation-based screening has the potential to increase data sets sizes and requried LAN transfers by several orders of magnitude. | |

| | | | | |
|---|---|---|---|---|
| **2-5 years** | | | | |
| · Same as above but more use of simulation for guiding materials discovery and design. | · Same as above but 10-100 larger capacities (compute and data).<br><br>· Users will continue to need to move experimental and simulation data/results between SNS, EFRCs, Hubs, CNMS, computational resources (both capacity and capability compute systems), home institutions. | · Same as above but the average data set size will increase toward several TB<br><br>· 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/OLCF/NERSC/ EFRC's/Hubs/…<br><br>· Need improve latency to match data volumes | · Maintaining or improving the time to transfer the key data for the major materials science codes would be important. | |
| **5+ years** | | | | |
| | · By 2020 we expect to have exascale computing capability. | · Computational systems will be considerably large in capacity. This will enhance the amount of simulations completed per day and likewise the number of data set produced.<br><br>· A case study is to dump the entire system (a projected 64 PB task). This would require 32 minutes with a LAN of ~33 Tbs and a high quality service for network reliability and latency. | · Would expect an increase in the number of data sets transferred as well as the size of the data sets. | |

# Case Study 4

# Combustion Research Facility

## 4.1 Background

The combustion research community is diverse and geographically distributed, and involves many research institutions around the world. Research in combustion science aims to provide a fundamental and predictive understanding of the complex multiscale processes that are present in a variety of transportation, propulsion, and power systems. It involves a wide range of computational and experimental techniques that span from the atomistic scales associated with chemical reactions to continuum scales associated with the exchange of mass, momentum, total energy, and species in reacting flow systems. Challenges involve treatment of turbulence, advanced fuels, multiphase flows, and catalytic systems, to name a few. A key objective is the construction of predictive models such as those that describe the multiscale interactions between turbulence and chemistry that can ultimately be assembled into engineering design tools for development and optimization of device-scale combustion processes.

Given the many multiscale and phenomenological challenges, the combustion community relies on interdisciplinary communication and integration of research results through a variety of experimental and simulation approaches. Establishing synergy between different techniques such as advanced laser diagnostics and state-of-the-art computations that make use of the full hierarchy of computational resources (e.g., desktop computers, mid-scale computer clusters, leadership class computing facilities) is an integral part of the research workflow. The scale at which both computational and experimental data acquisition techniques are producing information is now occurring at unprecedented rates. The growth in data output, coupled with the inherent need for collaborative exchange of information, is now challenging many assumptions about how the scientific process can efficiently operate. These changing paradigms often require advanced networking capabilities coupled with new software tools to use them. New experimental techniques involving high-repetition rate laser diagnostics, and numerical techniques such as Direct Numerical Simulation (DNS) and Large Eddy Simulation (LES), currently impose some of the most challenging requirements to ESnet.

## 4.2 Collaborators

Sandia National Laboratories (SNL) Combustion Research Facility (CRF) illustrates an example of collaborative activities that occur in the combustion science community. This facility is a special-purpose DOE Office of Science collaborative research facility. It is distinguished by a multidisciplinary combination of activities that range from 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 150 collaborators per year (those engaged in a research

project with one or more of the technical staff) and about 1500 visitors per year (those working on active research projects onsite, wanting to see various laboratories, attend technical workshops, or discuss current and future collaborations). Many other institutions (e.g., national laboratory, academic, and industrial) follow the same type of approach on an international level. Core research typically involves broad collaborative interactions that combine experiments, theory, and computation. Both local and remote exchange of data is a routine part of the research workflow.

## 4.3 Near-term Local Science Drivers

### 4.3.1 Instruments and Facilities

It has become increasingly recognized that a hierarchy of local and national hardware resources are necessary to support computationally intensive research in all scientific disciplines. Progress depends on ready access to all parts of the hierarchy, and on the complementary integration of its components, both locally, and between centers. This notion was effectively expressed in a seminal report in 1993 to the National Science Foundation entitled "Exploiting the Lead in High Performance Computing" by L. Branscomb (NSB 93-205, August 1993), and reiterated by the BESAC group in an Office of Science report entitled "Opportunities for Discovery: Theory and Computation in Basic Energy Sciences" in January 2005. Since this time, cluster computing and "mid-scale" systems have played an increasingly important role in the scientific enterprise.

The use of such a hierarchy of resources is now a mainstream approach applied in the combustion community. A foundational cornerstone of the CRF numerical combustion program, for example, has been facilitated by the Combustion Research and Computational Visualization Facility (CRCV) established in 2010. The CRCV was designed specifically to support collaborative research in computational combustion and includes an optimal combination of collaboration space, visualization facilities, and infrastructure to support a variety of computationally intensive tasks. It houses a state-of-the-art computer laboratory designed with ample power, cooling, and floor space to handle routine and sustained deployment of a variety of mid-scale high-performance computational equipment. The most recent system is a BES-funded 50-teraflop cluster with 5376 processing cores and a 600 TB parallel (Lustre) file system. The CRCV supports both in-person and remote collaborations. The visualization and conferencing rooms enable both small and large groups to collaboratively view and work with complex data sets. A variety of modern collaborative tools are available to facilitate visitors, remote collaborations, and/or related combustion community activities such as the "International Workshop on Measurement and Computation of Nonpremixed Flames"[1] and the "Engine Combustion Network"[2]. For cases when a given job fits on local systems, production calculations performed on dedicated mid-scale systems provide fast and routine access to pertinent research results without the overhead associated with large-scale resources. Similarly, preliminary calculations performed on mid-range systems help identify critical issues and phenomena of interest prior to accessing large-scale computing resources such as the Oak Ridge Leadership Computing Facility (OLCF) or the Argonne Leadership Computing Facility (ALCF). This enables more efficient use of these facilities.

In addition to the computational hierarchy, experiments in combustion are now also beginning to generate significant needs associated with local and remote data transfer. Groundbreaking high-repetition rate diagnostics, imaging detectors, and data acquisition has significantly transformed experimental investigations of the structure and dynamics of turbulent combustion. For example, the reacting flow program at the CRF involves a variety of experiments that employ multi-scalar point and line measurements and planar laser-induced fluorescence imaging of select flame marker species to obtain the instantaneous scalar mixing field and flame structure. In addition, particle-based velocity measurements are used to characterize various reacting flow fields. More recently, efforts have been focused on time-resolved imaging of transient flame structures undergoing phenomena such as extinction and reignition in turbulent jet flames. The experimental benchmark data serve as a basis for evaluation and development

[1] I.e., the "TNF Workshop," see www.sandia.gov/TNF

[2] I.e., the "ECN Workshop," see www.sandia.gov/ECN.

of turbulent combustion models throughout the combustion community. These efforts are coordinated on an international level through collaborative activities such as the TNF and ECN workshops. Within this construct, researchers meet to exchange data, validate and develop models against the benchmark data, and plan new experiments on a regular basis. Through constant web, email, and in person collaborations and meetings, this consortium of researchers has made progress well beyond that possible by individual researchers. With expanded access to both experiments and simulations, we anticipate that the size and breadth of a given collaborator base in combustion research will increase significantly.

### 4.3.2 Software Infrastructure

Combustion typically involves heterogeneous chemically reacting and/or multiphase mixtures that have a variety of complicating factors including highly nonlinear chemical kinetics, small-scale velocity and scalar-mixing, turbulence-chemistry interactions, compressibility effects (volumetric changes induced by changes in pressure), and variable inertia effects (volumetric changes induced by variable composition or heat addition). Coupling between processes occurs over a wide range of time and length scales, many being smaller than can be resolved in a numerically feasible manner. Further complications arise when liquid phases are present due to the introduction of dynamically evolving interface boundaries and the complex exchange processes that occur as a consequence. At the device level, high performance, dynamic stability, low-pollutant emissions, and low-soot formation must be achieved simultaneously in complex geometries that generate complex flow and acoustic patterns. Flow and combustion processes are highly turbulent (e.g., integral-scale Reynolds numbers of order 100,000 or greater), and turbulence dynamics are inherently dominated by geometry and various operating transients. In modern systems, operating pressures now approach or exceed the thermodynamic critical pressure of liquid fuels (and oxidizers in the case of liquid rocket engines). Operation at elevated pressures significantly increases the system Reynolds numbers, which inherently broadens the range of spatial and temporal scales that interactions occur over. Because of the complex multiscale nature of the problem, no one experimental or simulation technique is capable of providing a complete description of these processes. The highest quality experiments only provide partial information due to limitations associated with various measurement techniques. Likewise, solving the fully coupled equations of fluid motion, transport, and chemical reaction using the Direct Numerical Simulation (DNS) technique can only be applied over a limited range of turbulence scales in highly canonical domains of only a few centimeters in size due to prohibitive computational demands, even at exascale. Thus, a variety of approaches are used for simulation and analysis, and this trend will continue in the long term.

While petascale computing has enabled the application of DNS for treatment of three-dimensional reacting flows with detailed chemistry, the largest DNS runs to date are off by more than an order of magnitude in Reynolds number compared to practical devices. Thus, results from DNS are potentially useful to better select underlying physical assumptions used in modeling and to understand fundamentals related to small-scale phenomena. However, results can only be viewed as a first step and cannot be considered fully conclusive due to the limited dynamic range associated with the canonical flows employed. Conversely, the Reynolds-Averaged Navier-Stokes (RANS) approximation employs filtering in time to derive the governing conservation equations for the mean state. For this approach, all dynamic degrees of freedom smaller than the largest energy containing eddies in a flow are modeled to minimize the cost of the calculations. RANS is currently the primary method used for industry relevant engineering calculations since it is the least costly. However, since the models must represent dynamic interactions over the full range of spatial and temporal scales in a flow, they are the least universal in character and provide the lowest level of fidelity. These models inherently involve many tuning constants that must be calibrated on a case-by-case basis.

The Large Eddy Simulation (LES) technique falls between the two limits of DNS and RANS. The large energetic-scales are resolved directly and the subgrid-scales are modeled. Just like one chooses the resolution at which a photographic image is resolved, one can conceptually choose the resolution at which pertinent broadband structures of a flow are resolved using LES if validated subgrid-scale models are available. As the spatial and temporal resolution is increased, the cost associated with a calculation increases, but the range of scales over which the system of subgrid-scale models must work becomes

proportionately less and they tend to be more universal in character. Given these attributes, LES can represent a range of scales from the DNS limit to RANS and is used as both a tool for scientific discovery (using high-fidelity albeit more expensive first principles models) and for engineering design (using less expensive albeit less universal engineering-based models). DNS and LES impose the most stringent demands on computational and networking infrastructures.

### 4.3.3 Process of Science

The limitations and challenges associated with turbulent combustion research and the related development of chemical kinetics mechanisms requires that a hierarchy of approaches be taken to fully understand key processes and work toward predictive models. From this perspective, the process of science always involves significant interactions between theory, experiments, modeling, and simulations. Each approach complements the others. The process for typical combustion simulations involves three key stages: (1) programming and software readiness, (2) production run preparations and execution, and (3) post-processing. Programming and software readiness involves first determining and testing optimum programming models for multicore processors through performance monitoring and use of (for example) MPI, OpenMP, OpenACC, and/or CUDA. It then involves implementing and testing collective I/O and standardized I/O formats (e.g., HDF5 or NetCDF) as part of the selected programming model. Production run preparations and execution involves first a series of preparatory runs to determine the correct selection of numerical and physical parameters. These are typically performed on local mid-scale clusters. Then, after the correct setup is achieved, production runs are typically (but not necessarily always) run on the larger leadership class platforms such as OLCF or ALCF. Compute time on leadership class platforms is facilitated by either the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program or the ASCR Leadership Computing Challenge (ALCC) program. Production calculations funded by the Office of Science are also run at the National Energy Research Scientific Computing Center (NERSC).

## 4.4 Near-term Remote Science Drivers

### 4.4.1 Instruments and Facilities

The main facilities used remotely for combustion science are the Oak Ridge Leadership Computing Facility,[3] Argonne Leadership Computing Facility,[4], and National Energy Research Scientific Computing Center.[5] Compute time on leadership class platforms is facilitated by either the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program, or the ASCR Leadership Computing Challenge (ALCC) program. Once awarded a grant, the infrastructure associated with these centers becomes available to facilitate the collaborative research activities. Data transfer between these remote facilities and home institutions typically involves movement of large-scale datasets generated using DNS and LES.

### 4.4.2 Software Infrastructure

Production calculations performed on local mid-scale clusters typically require 1–5 million CPU hours and generate 10–50 TB of data. Production calculations performed remotely in leadership class platforms typically require 10–50 million CPU hours and generate 100–500 TB of data. This type of workflow occurs approximately 4–8 times a year. Smaller amounts of data are transferred to various collaborators across the country to universities more frequently (approximately 6 times a year). The size of respective datasets is approximately doubling every year, which highlights a need to shift the workflow more toward

[3] www.olcf.ornl.gov
[4] www.alcf.anl.gov
[5] www.nersc.gov

*in-situ* on-the-fly data processing that makes use of restart capabilities to extract new data from a given case as needed. Shifting the workflow in this direction will make the size of the data that needs to be stored relatively smaller (i.e., restart files of initialized primitive variables versus time correlated frames of processed data fields) at the expense of increased CPU time. Restart files are typically written hourly. As jobs are running, health is monitored (both locally and remotely) using various data analytic and plotting packages.

Post-processing involves a combination of tasks. Raw data from respective calculations are processed locally or remotely depending on the size, and feasibility of moving the output files. In most cases, local data is stored on the same cluster that the runs are done on. Typical local capacities are now on the order of 0.5–1 PB. Calculations being performed on leadership class platforms are typically processed remotely using the large-scale facilities at respective computer centers (e.g., OLCF, ALCF, or NERSC). Data is typically archived for approximately 5 years since it might be revisited multiple times by the modeling community. However, depending on the case, a measurable fraction of the data generated is also replaced by newer more accurate calculations as improvements in either the solution fidelity and/or models employed are made. Typical post processing involves a variety of software packages such as ParaView, EnSight, VisIt, Tecplot, and MATLAB. Remote rendering is employed routinely to minimize the need to move large data sets from remote to local resources.

### 4.4.3 Process of Science

Analysis and visualization of simulation data at remote sites is a common use model for all simulation scientists. Use includes the transfer of images and graphs of a running simulation using *in-situ* visualization. It also includes much more data-intensive operations such as the transfer of processed data for further analysis and visualization. Assuming that *in-situ* processing of data results in a reduction of two orders of magnitude, a simulation using all of the memory on (for example) the OLCF Titan platform would produce 6 TB of processed data per time step. If a simulation takes 10 minutes to compute a time step, this would be equivalent to a data rate of 80 Gbps.

Combustion simulations are currently being transitioned to work more efficiently on hybrid architectures and many of the analytics are being placed inside the simulation workflow at runtime to reduce the net amount of raw data that needs to be stored and/or transferred. As such, interactive remote data processing and visualization is becoming a common necessity for analysis of combustion data. Currently, this method does not place high demands on the WAN. (A typical 1,000 by 1,000 pixel image sent over the network at 5 frames per second only requires 0.12 Gbps.) However, as more users begin to simultaneously perform remote visualization using this method, the data rate required will increase linearly (e.g., 1000 users simultaneously would require 120 Gbps).

## 4.5 Medium-term Local Science Drivers

### 4.5.1 Instruments and Facilities

There are two areas where significant investments will be made in the next 2–5 years. The first will be continued upgrade and expansion of local mid-scale computer systems, which currently support both experimental and computational research especially in the areas of DNS and LES. As described in the example above, the CRF CRCV houses a state-of-the-art computer laboratory designed with ample power, cooling, and floor space to handle routine and sustained deployment of a variety of mid-scale high-performance computational equipment. The most recent is a BES funded 50-teraflop system with 5376 processing cores and a 600 TB parallel file system. The average lifetime of mid-scale systems such as this is approximately 5 years. Each new system is significantly more powerful than the previous. To stay current with evolving computer architectures, while at the same time providing the capacity required to sustain expertise in state-of-the-art numerical combustion calculations, new clusters will continue to be replaced using the most recent hybrid architectures. Mid-scale clusters typically follow

closely the basic design of leadership class systems such as the OLCF Titan CRAY XK7, but at a much smaller scale using mostly commodity hardware. Following this approach, recently proposed mid-scale systems are LINUX-based using computational nodes designed with dual-socket motherboards with (for example) AMD Opteron "Abu Dhabi" 6376 16C, 2.3/3.2GHz CPUs (32 cores/node). In addition, nodes are equipped with (for example) dual NVIDIA Kepler k20 GPUs. The communications interconnect is designed using Quad-Data-Rate (QDR) Infiniband with full bisectional bandwidth. The system is racked in the standard way and powered using standard Power Distribution Units that connect to 208 Volt, three-phase busses widely available in local facilities. Systems are typically configured to provide 2-gigabytes of RAM per core within each node, and include a Network File System over Infiniband with at least 400 TB of usable disk space. They also include provisions to connect to existing local high-speed Lustre file systems. The systems include fully configured operating systems, cluster management software, related licensing, and supporting peripheral equipment required for operation. While the total cost of a given system can be scaled up or down based on available funding levels and needs, an example of a modern system available in 2014 that costs $750,000 provides approximately 45 compute nodes (i.e., 1440 processor cores and 90 GPUs). Such a system will provide a peak theoretical speed of approximately 200-teraflops, which is approximately four times faster than the existing machine deployed in 2010 at 25% less in cost. Networking between local systems and remote computer centers (e.g., ALCF, NERSC, OLCF) will be pertinent, with emphasis on optimal management.

The second investment is related to the amount of data being generated from experiments, which is increasing exponentially as significant capital investments are made in laboratories such as the CRF Advanced Imaging Laboratory. New high-speed imaging systems, for example, are starting to provide information about turbulent flames that will be unique worldwide. Current experimental campaigns involve parametric studies that involve on the order of 10 cases per year that generate on the order of 1 TB of data per case. Over the span of 2–5 years it is anticipated that this volume will increase by at least an order of magnitude. Thus, data acquisition systems that are currently housed and self-contained within the laboratories (e.g., a basic NFS RAID stack with 15 TB of capacity) will be replaced by larger and faster systems housed within mid-scale computer centers such as the CRF CRCV. These data will eventually require the use of larger remote facilities. In addition, multi-core processor capabilities for data analysis will be required that can be used both locally and remotely.

### 4.5.2 Software Infrastructure

Needs related to the required software infrastructure are identical to the current needs described in this case study. However, these will need to be addressed in a manner consistent with the direction in hardware design taken to achieve exascale performance.

### 4.5.3 Process of Science

The process of science will not change over the next 2–5 years compared to that described already. However, it will need to be optimized. An example of a typical DNS simulation over the next 5 years is as follows. A 20-petaflop run will produce 10 PB of data per checkpoint file. I.e.,

$$10 billion node grid * 80 variables * 8 \frac{bytes}{variable} = 6.4 TB. \tag{4.1}$$

If 200 checkpoint files are written out at approximately 1 file per hour in a 7–10 day run, the data generation rate is approximately 20 Gbps. If the data is moved elsewhere at the rate it is generated, then we need a to move data at 20 Gbps. The network will need an even higher bandwidth to account for overheads due to protocol, metadata, contention, etc. The current goal is to stream data to an analysis and rendering machine as it is produced rather than waiting until the run is complete. Then, known analysis tools can be applied to the data as it is generated to get an initial understanding of the underlying physics. Subsequent iterative analysis can then be performed off-line. Automated workflow scripts will facilitate the data streaming, morphing, archival, and analysis. Data will need to be moved to platforms and archival storage within a supercomputing center. Reduced data will be transferred to local facilities

such as the CRF CRCV for refined analysis and rendering. To strike a balance between the size of data to be moved and speed at which it can be generated, typical simulations (e.g., DNS, LES) 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. In addition, web-based portals developed for sharing simulated benchmark data with a modeling community of approximately 50 to 100 international collaborators at national laboratories, universities, and industry will be developed. The goal is to deploy a scalable, extensible framework for analyzing large data from both computations and experiments. The framework will ideally adopt standardized formats, translators, graphics, combustion analysis software, parallel feature detection and tracking libraries, 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.

## 4.6 Medium-term Remote Science Drivers

### 4.6.1 Instruments and Facilities

The needs and goals outlined by major DOE computer centers such as ALCF, NERSC, and OLCF are consistent with the needs for combustion research. Optimizing the interplay between how local and remote facilities are used in a complementary way will then be the key challenge.

## 4.7 Beyond 5 years

A long-term goal in combustion research is the development of predictive multiscale models for turbulent, chemically reacting, and/or multiphase flows. Emphasis is typically placed on priority research directions identified in two recent DOE workshop reports: the SC-BES-sponsored "Workshop on Clean and Efficient Combustion of $21^{st}$ Century Transportation Fuels," and the jointly sponsored BES and EERE-VTO "Workshop to Identify Research Needs and Impacts in Predictive Simulations for Internal Combustion Engines." Needs are centered on two central issues. The first is addressing the basic science issues that limit our ability to simulate turbulent combustion. The second is addressing critical needs related to development of enabling software and the Information Technology (IT) infrastructure required to support collaborative model development. Success hinges on simultaneously advancing capabilities in the area of simulations, experimental diagnostics, and joint analysis of measured and modeled results in a manner that provides detailed validation data and the broad discovery-based research required for rapid development of predictive models. Current capabilities are lacking in all of these areas. Diagnostic techniques are limited. Numerical techniques are prone to many forms of error. Data reduction techniques are subject to many simplifying assumptions. There are many uncertainties. Thus, pertinent questions related to model development cannot be answered definitively. It is impossible to answer these questions through isolated research efforts and loosely formed collaborations. Instead, a concerted effort between experimental and computational researchers is necessary to provide a structured and well-coordinated framework.

Initial successes in facilitating joint collaborations in combustion have been achieved through internationally recognized forums such as the "International Workshop on Measurement and Computation of Turbulent (Non)Premixed Flames"[6] and the "Engine Combustion Network."[7] These workshops have provided a widely recognized level of success within the broad community of experts through annual and/or biennial meetings at various locations around the world, coupled with remote meetings held consistently on a more regular basis (e.g., monthly). To date, support for these forums has been voluntary and

[6] The "TNF Workshop," www.sandia.gov/TNF

[7] The "ECN Workshop," www.sandia.gov/ECN

expansion has been limited. Neither has a truly state-of-the-art web-based infrastructure. Such an arrangement will enable frequent and efficient remote collaborations. Thus, there are significant benefits to be gained through development of a dedicated IT infrastructure that caters to the complementary needs of these types of activities. A central goal is to foster synergistic relationships by providing a platform for collaborative model development and validation while providing a common set of well-documented openly available software and databases with web-based access.

## 4.8 Data, Workflow, Middleware Tools and Services

The dramatic increase in the amount of data generated in future simulations will necessitate a relative decrease in the amount of I/O possible. Currently, the size of respective datasets from DNS and LES is doubling every year, which is not sustainable. This will necessitate a large movement toward *in-situ* data analysis and visualization at runtime. Thus, the role of scientific data management middleware will be key in providing the software infrastructure that exploits current petascale and future exascale architectures ability to overlap data analysis with I/O. The computational power of processors will be used to reduce the need for data output by building I/O systems that enhance massively parallel methods for data movement with extensive processing resources to manipulate and reduce data before it is moved. Since I/O will become a very constrained resource, it is expected that the adoption of I/O middleware libraries will significantly accelerate over the next five years, with the majority of HPC applications using some middleware system after this point. This is especially true for combustion science applications. Scientific workflow tools are also likely to increase in prominence, especially for scientific analyses that involve parameter sweeps or manage ensemble runs. Workflows that extend beyond the individual computer system are also likely to see greater adoption, including management of archival storage, off-site data transfer, exploitation of multiple computing systems, and distribution to web-based portals.

*Table 4.1: The following table summarizes data needs and networking requirements for the CRF.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · 10 petaflop DNS/LES running for 10 days per run on leadership facilities<br>· 50 teraflop LES/DNS running for 10 days on local mid-scale clusters<br>· High-speed laser diagnostics for combustion experiments requiring advanced local storage and post-processing | · 50-100 collaborators around the world<br>· International workshops aimed at joint analysis of experimental and computational results of common target cases<br>· Emphasis on both combustion chemistry and reacting flows | · 100 to 500 terabytes per dataset for LES/DNS<br>· 10 to 50 terabytes per dataset for high-speed imaging experiments<br>· Extensible and scalable I/O using NetCDF or HDF5. Optimal balance between number of files and file size adjustable | · LAN transfer time typically bounded by 1000BASE-T Ethernet | · 150 Gbps WAN bandwidth<br>· Collaborating sites include National Labs, Universities, some Industry<br>· Domestic and International |
| **2-5 years** | | | | |
| · 20 petaflop DNS/LES running for 10 days per run on leadership facilities<br>· 100 teraflop LES/DNS running for 10 days on local mid-scale clusters<br>· Continued development of data acquisition systems for high-speed laser diagnostics with data acquisition handled by mid-scale clusters and storage systems | · Same as near term, but with extensive community-wide collaborations and data sharing through web portals. | · 1 to 50 petabytes per dataset for LES/DNS<br>· 100 to 500 terabytes per dataset for high-speed imaging experiments<br>· Extensible and scalable I/O using NetCDF or HDF5. Optimal balance between number of files and file size adjustable | · LAN transfer time typically bounded by 1000BASE-T Ethernet | · 300 - 2400 Gbps WAN bandwidth unless in-situ processing techniques are implemented at runtime to reduce ouput of raw data for post-processing<br>· Collaborating sites include National Labs, Universities, some Industry<br>· Domestic and International |
| **5+ years** | | | | |
| · DNS/LES code capabilities will continue to be refactored to efficiently scale on the newest architectures with improvements in attaining a greater fraction of theoretical peak architecture speeds<br>· Continued development of data acquisition systems for high-speed laser diagnostics with in-situ data reduction done in parallel on mid-scale clusters and parallel storage systems | · Interactive remote data processing and visualization will continue to become a common necessity for analysis of combustion data<br>· Development of a collaborative information technology infrastructure informed by a diversity of experimental, simulation, and data analysis activities between a large network of academic, industry and national laboratory partners | · Unknown. This will depend on the balance between data analytics performed at runtime versus post processing performed after completion of a run, which requires significantly more data be output, stored, and processed | · Unknown. This will depend on the balance between data analytics performed at runtime versus post processing performed after completion of a run, which requires significantly more data be output, stored, and processed | · Unknown. This will depend on the balance between data analytics performed at runtime versus post processing performed after completion of a run, which requires significantly more data be output, stored, and processed |

# Case Study 5

# Linac Coherent Light Source

## 5.1 Background

Linac Coherent Light Source (LCLS), the world's first hard X-ray free-electron laser, pushes science to new extremes with ultrabright, ultrashort pulses that capture atomic-scale snapshots in quadrillionths of a second. These images can reveal never-before-seen structures and properties in matter, and can be compiled to make movies of molecules in motion. Since its 2009 launch, LCLS has drawn researchers in a wide array of scientific fields from around the globe to explore the innermost workings and properties of common and exotic materials at the nanoscale. LCLS features six specialized instrument stations, each with a dedicated team of scientists and support staff, to conduct pioneering research and assist users with experiments. Each station is equipped with a suite of instruments to assist in gathering a wide range of data using various specialized techniques, from telltale signatures of electrons and ions to the intricate patterns left by crystallized samples struck by the X-ray laser.

## 5.2 Collaborators

Roughly 50 experiments are run at LCLS each year. The scientists work in collaborations ranging from 10–50 collaborators. Beam time is awarded by proposal process with a 25% acceptance rate. The diversity of the science cases is demonstrated by the schedule from early 2014.[1]

LCLS users are responsible for complying with the data management and curation policies of their home institutions and funding agents and authorities.

## 5.3 Near-term Local Science Drivers

### 5.3.1 Instruments and Facilities

The experimental facilities are at LCLS generally referred to as the Far Experimental Hall (FEH) and the Near Experimental Hall (NEH). These facilities are physically separated from the main part of the campus. The instruments, experimental apparatus and experiment Data Acquisition (DAQ) Systems are controlled via kiosk nodes in the experimental control rooms. The typical user will not have any remote access to the kiosk nodes.

Each experimental control room has nodes for local login to LCLS offline compute and disk resources. These nodes also have external access. LCLS supplies on-site disk, tape and compute resources for

[1] See http://www-ssrl.slac.stanford.edu/userresources/documents/lcls-scheduling-run_8.pdf.

prompt analysis of LCLS data, and software to access those resources consistent with the published data retention policy. Data storage is provided for raw detector data and derived data from analysis. The raw data formats are XTC files or HDF5 files on demand. The raw data is written to archival tape storage, with the robotic storage located in the computing center on the main area of campus. Live disk storage is provided, (currently 6 PB total), with raw data maintained for 6 months. Several batch processing farms (Infiniband connected) are located in NEH and FEH. Compute resources are preferentially allocated to recent and running experiments. Each instrument gets a dedicated file system for caching data of on-going experiments. The current experiment gets preferred access to the data cache.

LCLS integrates 2 PB of raw data per year. Data rates vary by instruments from 180 MB/s (3 TB/shift) to 1.5 GB/s for Coherent X-Ray Imaging (CXI) where the event size exceeds 10 MB (30 TB/shift). The system is file-based with typical file sizes of 100 GB.

Connections to major scientific instruments and other facilities include university computational resources associated with the researchers, and computational resources at other major national and international computational facilities such as the National Energy Research Supercomputer Center (NERSC) and the Deutshes Elektronen-Synchrotron (DESY).

### 5.3.2 Software Infrastructure

The software infrastructure for LCLS is largely custom-built.

The DAQ system, hardware and software resources are the same for all six of the LCLS instruments, as distinct from many other lightsources that might have a different system or set of tools at each beamline.

Transferring data to portable media such a storage device or laptop is available in the control rooms via the SLAC Visitor Data Network. A variety of existing tools such as MatLab, IDL, Igor, Python, and custom frameworks are supported.

Raw data files are cataloged via an iRODS-based system. Data security is enforced by group based access control.

XTC is a custom format (with associated read/write libraries) developed at SLAC (circa 1995) for the BABAR experiment, and is used at LCLS by the DAQ as the native raw data format. Simultaneous read/writes facilitate online monitoring and user access for rapid feedback. The need for special libraries reduces the portability. The users can request translation from XTC to HDF5 offline, with roughly 10% of the data being translated. Both formats achieve a factor of 2 compression. A web portal to the tool set provides access to the electronic log book, the file catalog and other data services, such as changing file permissions, requesting data be restored from tape storage and managing the HDF5 translation. A valid SLAC Unix account is required.

LCLS plans to implement additional monitoring to better understand the bottlenecks and usage within the local system, with tools such as Ganglia and Monit.

### 5.3.3 Process of Science

The user community is extremely diverse, with typically two new experimental groups coming to LCLS each week. The typical duty cycle of the experiments is five 12-hour shifts on consequtive days, with two experiments alternating shifts. The typical postdoc has little time to learn the software environment, and the approachs to data analysis vary considerably from experiment to experiment.

There is currently no need or plan to reduce the raw data or filter data during data collection during this phase of LCLS. A fast feedback system is used to tune analysis and algorithms during data collection. Further algorithm development will often take place after the experiment.

## 5.4 Near-term Remote Science Drivers

### 5.4.1 Instruments and Facilities

We support two data export machines each with 10 Gbps connections. The current SLAC to ESnet connection capacity is 2x10 Gbps with plans to upgrade to 100 Gbps.

### 5.4.2 Software Infrastructure

The recommended tools for exporting the data offsite are bbcp and Globus.

### 5.4.3 Process of Science

LCLS Computing and NERSC are actively engaged in joint pilot projects to facilitate data transfers to NERSC, currently to enable the work of individual research groups during there data collection period. This process has identified a number of bottlenecks, and there are plans to upgrade the SLAC to NERSC link.

The nature of these transactions was in a “push” mode. Users can use the standard tools such as Globus to pull the data to the remote sites.

Other user groups do pull reduced data and raw data to their local sites, including transatlantically located sites. Many groups prefer to use the computing resources at SLAC since it can be labor intensive to transfer the data, or there are better computing resources while at SLAC.

## 5.5 Medium-term Local Science Drivers

### 5.5.1 Instruments and Facilities

The number of beam lines will remain constant at LCLS, however, data will be collected simultaneously at multiple instruments.

The data rates are expected to increase by a factor of two to five, due to improvements in detector technologies. Data volumes may also increase due to improved duty factors.

For the LCLS’s Data Management Plan, LCLS plans to continue to offer the current data services, access and curation as it does now. These policies cover raw data only. The users are responsible for curation of the reduced data, data used in publications, and curating the algorithms.

### 5.5.2 Software Infrastructure

We anticipate mostly incremental changes to the infrastructure with the possible exception of a transition to HDF5 if simultaneous read/writes become enabled. Our primary projects in this area are to insure seamless integration with NERSC such that the users will be able to submit analysis workflows transparently regardless of whether the data is at SLAC or at NERSC.

### 5.5.3 Process of Science

Many of the experiments that will run at LCLS will be second or even third generation; we expect that this will increase the familiarity of the users, and the readiness and sophistication of their algorithms during the fast feedback phase. This is likely to encourage more analysis at offsite locations due to competition for resources at SLAC.

The European X-ray Free Electron Laser (XFEL) facility in Germany will begin taking data in 2017. While LCLS collaborates with XFEL and on SACLA, the Japanese XFEL facility, on some data and software issues, the systems have significant differences, and the common users may ask for more commonality.

## 5.6 Medium-term Remote Science Drivers

### 5.6.1 Instruments and Facilities

Working with NERSC, we expect to facilitate the ability of users to transparently use NERSC resources for selected experiments.

We expect that simulations of scientific outcomes, experiments and beam conditions will become common, which will require transfer of data and information from Supercomputer centers to SLAC for local analysis. There may be need for simultaneous allocations of time at supercomputer centers and LCLS. The data volumes may or may not be large, however.

The XFEL will be operation in 2017 and we anticipate that SLAC scientists will be doing experiments there.

### 5.6.2 Software Infrastructure

Software infrastructure for XFEL is unknown.

### 5.6.3 Process of Science

At the XFEL, the process of science in this case will be on a case-by-case basis according to the needs of the collaboration. We may need to plan for significant network traffic coming from DESY.

## 5.7 Beyond 5 years

An upgrade to LCLS is in the design phase. The high-repetition rate (1 MHz) and, above all, the potentially very high data throughput (100 GB/s) generated by LCLSII will require a major upgrade of the data acquisition and storage systems and increased data processing and data management capabilities. The main challenge will be developing high-density, high-throughput, petascale storage systems which allow concurrent access from thousands of jobs. Additional critical capabilities include the deployment of a trigger/veto system, upgrading the SLAC network connection to ESnet and expanding bandwidth and capacity of the tape archive.

## 5.8 Network and Data Architecture

SLAC has two connections to ESnet to provide redundancy. The primary ESnet connection is 100Gbps, and the backup connection is 10Gbps. SLAC also has connectivity to Internet2 via Stanford University (this connection is through the Stanford Research Computing Center which is located on the SLAC campus).

SLAC does not have a specific Science DMZ, but the LCLS provides data transfer nodes for LCLS and SSRL experiments for high-speed data transfers.

## 5.9 Data, Workflow, Middleware Tools and Services

LCLS represented a transition in computing workflows, services and data management. At this point, we are expecting the evolution of tools, particularly in the case of emerging standards or technologies. We will continue to collaborate closely with other institutions and continue to survey our users to target our development efforts.

We are considering using a commercial service for the second curated copy of the LCLS data.

## 5.10 Outstanding Issues

The following represents a personal point of view: the current trends in hardware and other factors are strongly motivating aggregation of hardware resources into extremely large-scale data facilities. If this trend continues, there will need to be profound changes in the way that data is managed by relatively small user groups. This would have considerable impact on networking, security and remote usage point of view in order to provide a good user experience.

*Table 5.1: The following table summarizes data needs and networking requirements for the LCLS.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · LCLS has six instruments.<br>· Custom built framework, data catalog, and Globus or bbcp for transfers. | · User driven, fast feedback in place. Users may use remote resources at their home institution for analysis. Simulations or theoretical computations not included.<br>· Continue DMP support at current level of archiving two copies of the raw data locally at SLAC. | · File based, typical file size is 100GB.<br>· Size of one data set = 500 GB.<br>· Range of data set sizes: 500GB to 3TB depending on experiment.<br>· Data set composition is typically relatively few, large files for raw. Wide variation for reduced sets. | · Steady state: 0.2 GB/s; peak 1.5 GB/s raw data. Local transfers of analysis data, assume steady state of 0.1 GB/s. | |
| **2-5 years** | | | | |
| · None | · Intend to support some simultaneous simulations.<br>· automate integration with NERSC on some workflows.<br>· SLAC scientists may begin working at XFEL.<br>· May use some cloud services for archiving one of the data copies.<br>· Possible re-evaluation of DMP consistent with SC directives. | · Size of one data set 1TB.<br>· Range of data set sizes: 500GB to 5TB depending on experiment.<br>· Typically relatively few large files. | · Steady state: 2 GB/s; peak 15 GB/s raw data. Local tranfers of analysis data, assume steady state of 1 GB/s. | · 25 GB/s steady state, 50 GB/sec peak.<br>· Collaborating institutions: national (NERSC) and international. (XFEL). |

| **5+ years** | | | | |
|---|---|---|---|---|
| · LCLS II will come online, increasing the number of instruments. Detector technology is expected to also lead to a dramatic increase in data volume. | · Expect extensive simulation integration with experiments.<br><br>· Full-scale integration with NERSC and/or other computing centers.<br><br>· Serious investigation of online filtering or zero suppression.<br><br>· Will replace components of software/hardware systems as motivated by changing technologies, emerging standards or bottlenecks.<br>· Need highly performant petascale storage systems. | · Maximum size is being determined and will depend on detector technology and the possibility of on detector data reduction techniques. | · Steady state: 20-50 GB/s; peak 50-100 GB/s raw data. Local transfers of analysis data, assume steady state of 5-10 GB/s. | · Ideally continuously, during data collection, blurring the line between LAN and WAN resources.<br><br>· Collaborating sites: NERSC, International, University resources (derived data), XSEDE resources. |

# Case Study 6

# Materials Project

## 6.1 Background

The Materials Project's (MP) goal is to enable data-driven materials design. It will achieve this objective by (1) using high-performance computing to generate large materials data sets using state-of-the-art theoretical techniques, (2) integrating existing data generated by the materials community, and (3) disseminating the data in powerful ways. Thus far, most focus has been on using HPC noted in the first initiative, with preliminary work on the second and third initiatives. Over 30 million CPU-hours worth of computations have been performed and are freely available at www.materialsproject.org. Over 8000 unique users have registered for access. About 50 users are employing a REST API to access large amounts of data, and data from outside users are already available on the site.

Most of the current work has been focused on in-house contributions using the NERSC supercomputing center; computations are performed, stored, and accessed there. However, going forward we envision more powerful and decentralized ways to build a data resource that requires powerful network technologies; for example, data might be generated and hosted at several computing centers, and user queries would combine data sets from multiple centers. In addition, whereas current data access is to a limited number of post-processed data, in the future we plan to allow downloads of raw data sets and real-time analysis of complex data (e.g., charge densities). Finally, data are likely to flow into the MP automatically from several instruments, e.g. beam lines at the Advanced Light Source, or conversely scientists working at these experimental facilities might require real-time launching and analysis of theoretical calculations. Thus the future vision of MP is to expand beyond a single center/data type in a way that leverages powerful network resources.

## 6.2 Collaborators

The MP works with funded collabortions as well as unfunded ones. The distinction is that funded collaborations require software and scientific developments, whereas unfunded collaborations leverage existing data, infrastructure and software.

*Funded collaborations:* The MP powers the computational design—the Electrolyte Genome and the multivalent cathode design—of the Argonne-led Joint Center for Energy Storage (JCESR) which is five-year, $20 million DOE hub. The MP is also a funded partner in the NSF-funded SI2 Materialshub (www.materialshub.org) led by University Wisconsin at Madison and recently with the NREL-led Center for Next-Generation Materials Design (CNGMD). Furthermore, MP uses its capabilities as a co-PI to predict novel photocatalysts for the Joint Center for Artificial Photosynthesis (JCAP) (http://solarfuelshub.org) and is collaborating with Volkswagen and Stanford University on corrosion-resistant magnesium alloys. Currently, MP partitions funded collaborative, competitive work into secure sandboxes that interact with

the common public data and algorithms. As an example, JCESR has two extra “Apps” with data and customized search algorithms which appear only to JCESR affiliated personnel. As work becomes published, these Apps and data will transition into the public repository. Similar capabilities are planned for the NREL EFRC and the Volkswagen–Stanford collaboration.

*Unfunded collaborations:* The MP lends its software and data infrastructure to a number of individuals and institutions—through a two-way REST interface. Several millions of data records are downloaded each year, and thousands of data records have been published online as user-contributed data. Several currently unfunded pilot projects are being pursued to broaden MP’s scope. A pilot project with the Royal Society of Chemistry (www.RSC.org) is in progress on two-way data sharing. Another project with the Advanced Light Source (ALS) at LBNL is pursuing data flow on spectrocopy and magnetic alloys to MP.

We are also collaborating with and supporting a nascent initiative from UC San Diego on sponsoring computations on demand (MGCloud). The proposal was submitted during the summer of 2014.

## 6.3 Near-term Local Science Drivers

### 6.3.1 Instruments and Facilities

We extensively use the NERSC supercomputing center; this year, we have already used 30 million CPU-hours for the MP and 15 million CPU-hours for a related project, the JCESR. Furthermore, our raw data ( 50TB), databases, and science gateway is hosted at NERSC. We have also purchased a $1 million computing cluster (“Mendel”) that is managed by NERSC.

### 6.3.2 Software Infrastructure

We have developed several in-house, open-source software solutions for materials science applications and for running and managing large high-throughput calculation projects in general. This software includes the pymatgen materials analysis library and FireWorks workflow software.[1] These tools are generally used to manage, run, and error-correct calculation workflows at NERSC.

In our case, “collaborators” are a community of over 8000 users who depend on the data for their research. The raw data is hosted at NERSC and shared via a science gateway developed and maintained jointly by the MP and NERSC and through a RESTful interface. Currently, the data is hosted at only a single source, however we expect that remote replicas will soon be needed to ensure rapid data transfer and good user experiences. In addition, we expect usage of the REST interface and large-scale downloads of data to increase in the future. This is particularly the case if we make more of the raw data (as opposed to processed data) available to the user.

### 6.3.3 Process of Science

Locally, our team uses the networking infrastructure of NERSC to download small data sets for local analysis. However, most data generation and processing is done within the NERSC global file systems.

Occasionally, we encounter the problem that developers and major MP contributors have to be registered NERSC users and go through the NERSC firewalls to access and manipulate the MP data and software. While it is desirable to have a control on core developers, part of the MP team is international and it has occasionally caused delay.

[1] These codes and more are available at http://www.github.com/materialsproject.

## 6.4 Near-term Remote Science Drivers

### 6.4.1 Instruments and Facilities

Currently, almost all our work is performed in-house at NERSC. We have pilot projects in terms of external data contributions from NREL and also the ALS, but these are preliminary. In addition, we have plans to extend our calculation infrastructure beyond NERSC (see below).

### 6.4.2 Software Infrastructure

The software tools developed by MP have been successfully deployed on university clusters and the San Diego Supercomputing Center, and are currently being installed at the Argonne Leadership Computing Facility. Thus, it is expected that calculations will become more distributed in the future, and we will need to manage and compile data generated across several centers. At the current moment, almost all calculations are performed at NERSC.

### 6.4.3 Process of Science

All users of the MP are in a sense "remote" users as they access the data via a science gateway. Currently, they are able to browse processed data sets via a rich, interactive web site and access limited amount of raw data through a RESTful interface. Currently, it is not possible to download the full raw datasets as we expect that thousands of users downloading tens of terabytes of data could not be supported by the network. In addition, we require a good way to expose certain parts of our NERSC project directories to users.

We are also experimenting with allowing users to "suggest" calculations; currently, only a small amount of information is transferred (the crystal structure), and the user is only exposed to the processed results. However, we envision a scenario whereby a user could also interact with portions of the raw data set that are not available today. For example, the full charge density output by a calculation is on the order of 100MB; to visualize and analyze this kind of file over the network, repeated over thousands of users, would produce terabytes of network traffic per month.

At least one external web site, the MaterialsHub (www.materialshub.org), uses data from the MP REST API to power third-party developed applications and toolkits.

## 6.5 Medium-term Local Science Drivers

### 6.5.1 Instruments and Facilities

Locally, we expect to use the new supercomputers in development at NERSC. In addition, we plan to form a partnership with the Advanced Light Source such that computational and experimental data can be integrated under a common science gateway. We also anticipate that the MP will become a partner in possibly several future materials design centers. As specified in the latest DOE RFI on Specific Clean Energy Manufacturing Focus Areas Suitable for a Manufacturing Innovation Institute which mentions the MP as one of three exemplary partners, together with the NSF-sponsored Network for Computational Nanotechnology led by Purdue, and the NIST-sponsored Center for Hierarchical Materials Design (CHiMaD) led by Northwestern. These Advanced Materials Manufacturing Centers (if funded) will require increased data production, management, analysis and dissemination/network capabilities.

### 6.5.2 Software Infrastructure

Development of the current software tools will continue. We plan to further release tools to allow users to easily develop and test their own calculation workflows. We also expect to further develop tools that allow smaller “high-throughput” jobs to run at large supercomputing centers, via techniques such as advanced checkpointing and job packing.

Sandboxes will continue to grow in number and size. Currently, we are testing incremental database building that will facilitate the assimilation of several different sandboxes. However, the current infrastructure will need to change (or be replicated) to sustain numerous sandboxes. Possibilities include mirrored MP databases and smart algorithms for updating.

Direct data conduits to ALS, ALCF may be desirable if those efforts become larger than the current pilot stage.

### 6.5.3 Process of Science

We expect a combined effort on specific science drivers through target collaborations which extend the MP’s capabilities as well as leveraging those capabilities as public resources. We expect to increase our interaction with other materials design centers, highly increased processing and storing of materials data and algorithms that operate on that data. For example, remote facilities, i.e. other supercomputing centers, data hosting centers, and database hosts will be incorporated into MP. Improved algorithms in materials science will be implemented first within the areas in which MP is actively funded.

## 6.6 Medium-term Remote Science Drivers

### 6.6.1 Instruments and Facilities

For data generation, we expect to leverage several supercomputing centers such as the ALCF in performing calculations. In addition, we may need to host our data and mirror our databases at several facilities. Finally, we hope to form partnerships for data sharing with several experimental characterization facilities, such that many types of materials data sets can be united within a common gateway and programmatic interface.

### 6.6.2 Software Infrastructure

We plan to expand community development of our open-source codes. Furthermore, we plan to release tools that allow users to build their own custom third-party applications (Apps) based on the MP data set that can be hosted externally. Thus, development on materials analysis toolkits based on the core data can be performed and hosted remotely. In addition, we may allow external sites to host their own data and expose it via a common REST interface. Thus, a common gateway could query data from several sites and collect them to form a single result, in real-time.

### 6.6.3 Process of Science

We expect that access to the raw data and to larger data sets will become more important, and users will demand the ability to download and locally process datasets that are tens of terabytes in size. As mentioned earlier, the infrastructure may also grow less centralized and thereby more reliant on the network. Rather than a single data resource hosted locally, the MP would instead become a collection of resources that are individually managed but connected by the network.

## 6.7 Beyond 5 years

We estimate that MP—even using a conservative estimate of constant growth—will be hosting 20,000-50,000 users worldwide and supporting increasingly larger data set transfers between MP and outside entities. Until now, we are have relied on NERSC to supply our storage, network capability and data transfer rate with the outside world. Depending on the growth and demands on MP (as well as the role of NERSC) this may need augmentation.

## 6.8 Network and Data Architecture

The network that supports the data transfer to/from MP decides what data can be accessible to the users. If larger data sets can be transferred within a reasonable time frame, that opens up new possibilities. Today, MP stays within its capabilities for storage and dissemination as provided. Improved network and data architecture could instantly help users visualize and download charge densities, *ab initio* molecular dynamics trajectories, molecular assemblies snapshots etc. which are not considered today.

## 6.9 Data, Workflow, Middleware Tools and Services

Better tools for automated backup of large data sets would be extremely useful. In addition, tools to help set up REST interfaces to distributed data resources would be highly appreciated.

One issue regarding the vision for MP involves network security. While distributed databases and workflow systems involving several computing centers are attractive, they can be slowed by network security issues. For example, one issue regarding our workflow software is that it requires worker computers to be able to query a "task queue" via MongoDB. However, many supercomputing centers invoke network security restrictions on their compute nodes that make it impossible to set up such a global task queue with distributed computing. Instead, one must set up different task queues at the different centers, with the database hosted on the internal network.

Finally, our data storage needs are challenging in that the raw output files encompass tens of millions of output files (although the total size is only 50TB). All these files need to be easily accessible because outputs of prior calculations need to be used to feed the inputs of next-generation calculations, in real-time for millions of workflows. As all the files are currently stored on NERSC project directories, issues like latency are not a problem. However, in the distributed computing model a task on Resource A might require output files that are stored on Resource B. For this to succeed, both resources must be available and the transfer of data needs to be fast.

*Table 6.1: The following table summarizes data needs and networking requirements for the MP.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · In-house produced data and software at NERSC and contributed stored at NERSC. Pilot projects with ALS, RSC and ALCF in progress. | · Increased collaborations with materials data repositories and facilities will increase the need for data storage and transfer rates. | · Current size of MP db is 50 TB for 50,000 materials. This data includes removal of files and data for all calculations to fit within reasonable NERSC storage.<br>· Desirable data storage in the hundreds of TBs enabling novel visualization and transfer options. | | |
| **2-5 years** | | | | |
| · Hosting of data and mirroring of databases at several facilities. Partnerships for data sharing with several experimental characterization facilities. | · Increased collaborations with materials data repositories and facilities will increase the need for data storage and transfer rates. | · > 500 TB as collaborative partners and data sets grow. | | |
| **5+ years** | | | | |
| · The data storage and transfer rates available will determine what roles MP can undertake. | · A fully networked material science community could develop comparative algorithms and analysis on data sets that are currently not available unless you are physically at the facility that collected it. | · Possibly petabytes. | | |

# Case Study 7

# Theoretical Modeling of Pump/Probe Experiments in Strongly Correlated Materials

## 7.1 Background

We have assembled a global team that is working on calculations to describe pump/probe experiments in special classes of materials, as performed in large facilities like the Linac Coherent Light Source (LCLS) at SLAC. Most of the production runs of our calculations take place at NERSC, where we currently are using 3-5 million CPU hours per year. Data needs to be transferred from NERSC to the different sites where additional data analysis and visualization tools are applied to the data. This typically includes sites like Stanford University, Lawrence Berkeley Laboratory, Georgetown University, University of Bonn, the Institute for Condensed Matter Physics in Lviv, Ukraine, and the Indian Institute of Science in Bangalore, India. Data files are usually in the hundreds of megabytes to a few gigabytes, and bottlenecks, especially for foreign transfers are typically coming from Internet issues beyond the control of DOE networks.

## 7.2 Collaborators

Collaboratoring institutions include:

- NERSC. This is where most of our computation is performed.
- Stanford/SLAC. Tom Devereaux and Brian Moritz lead the efforts here.
- LBL. Lex Kemper is a collaborator here.
- Georgetown University. Jim Freericks leads the efforts here.
- University of Bonn. Michael Sentef is the collaborator here.
- Institute for Condensed Matter Physics, Lviv, Ukraine. This effort is led by Andrij Shvaika.
- Indian Institute of Science, Bangalore, India. Work here is led by Hulikal Krishnamurthy.

## 7.3 Near-term Local Science Drivers

### 7.3.1 Instruments and Facilities

We create our data at NERSC and store it locally there.

### 7.3.2 Software Infrastructure

Software codes for production runs are created by us. We use facilities at NERSC for transferring files and archiving them.

### 7.3.3 Process of Science

After completion of data runs, we often need to postprocess data and visualize it. This is usually done locally after transferring files from NERSC, although some postprocessing can be done directly at NERSC allowing the postporcessed files only to be transferred.

## 7.4 Near-term Remote Science Drivers

### 7.4.1 Instruments and Facilities

We rely on the Internet for connection to different facilities and between the different collaborators. While the US institutions are part of Internet2, we rarely find high-speed connections across the coast, because the data transfers usually cross over to the regular Internet at some point with the concomitant slow down. The facility we need to maintain connections to is NERSC.

### 7.4.2 Software Infrastructure

We write our own software for performing the scientific calculations. Currently these codes are producing small to moderate data sets, but this might change in the future with newer algorithms or with different types of problems that we are studying. The main need for data transfer is for postprocessing of data and for visualization. Postprocessing is also done with our own codes. Visualization, on the other hand, is performed with different software packages (XmGrace, pgplot, Paraview, etc).

### 7.4.3 Process of Science

Production runs of codes take place at NERSC, as well as initial data storage of results. Data is then transferred to the primary site where the collaborator who is performing the calculations resides. Data is then postprocessed, visualized, and often shared with other members of the team. Usually this summative data results in rather small data sets that can be shared by e-mail or by dropbox or other download sites. We have not yet engaged in more data-intensive collaboration where multiple collaborators are looking at and processing visualization software in real time across the network. This is because the bandwidth usually is not good enough to enable such work.

File sizes are rarely in the few gigabytes of size and usually smaller than that, but Internet connections to the Ukraine and India are painfully slow, so we minimize data transfers there.

## 7.5 Medium-term Local Science Drivers

### 7.5.1 Instruments and Facilities

We plan to continue to use NERSC for production runs and local resources for post production. This is unlikely to change.

### 7.5.2 Software Infrastructure

We are continually developing new algorithms and new software. It is possible that some of the newer algorithms or newer problems we work on will generate significantly more data than what we currently use.

### 7.5.3 Process of Science

We anticipate having a closer interaction with experimentalists and a more materials-specific modeling of data. This may require more live video conferencing and data sharing than before.

## 7.6 Medium-term Remote Science Drivers

### 7.6.1 Instruments and Facilities

Our remote sources will not be changed.

### 7.6.2 Software Infrastructure

Postprocessing and visualization tools are unlikely to be changed.

### 7.6.3 Process of Science

We may have the need for more data sharing, for video conferencing, and for more collaborative science as we work more directly with experimentalists on their results.

## 7.7 Beyond 5 years

Eventually, it would be useful to be able to share efforts in real time for postprocessing, especially for different ways to visualize data. It is not clear how much of a higher load this will place on data transfers by the Internet, but at the moment we do not really think of trying such interactive sessions because they will not work effectively. As we work more directly with experimentalists, we have to factor in their needs for data sharing, postprocessing and data visualization to make more collaborative efforts possible.

## 7.8 Network and Data Architecture

The problems we work on can be classified as extreme data science—problems so big that a database cannot be made of all of the data and the science itself requires one to construct the data by keeping only important data in a causal way in time. As our algorithms grow, it is likely that these "subset" databases will be large and need to be stored and transferred from one location to another, but it is too early to tell how this will pan out.

## 7.9 Data, Workflow, Middleware Tools and Services

Having automated data transfer tools that can restart after they are stopped prematurely and that can transfer data in the background with limited supervision—tools like Globus—would always be helpful.

## 7.10 Outstanding Issues

So far our work has not pushed the data handling issues to the current limits that are available, in part because we find that data transfers can ultimately be slow from one place to another. We try to have workarounds, including generating and transferring the smallest datasets possible. This code of operating might change in the future and allow us to have more collaboration in the ways in which we handle and treat data.

*Table 7.1: The following table summarizes data needs and networking requirements for theoretical research modeling of materials.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · NERSC.<br>· Develop codes to solve the quantum many body problem in nonequilibrium. | · Data generation is for two-time Green's functions which can easily require 10-100 million double complex numbers for each set of parameters that are run. | · 100s of MBs to 10s of GBs.<br>· Usually dumped to a few files (Green's function, self-energy, etc). | · Varies significantly. But being able to transfer 10s of GBs in minutes would be good. Often go for long times with no data transfer then need tens of transfers per day for weeks. | · 10s of GBs in less than an hour.<br>· Bottleneck is when data transfers go over the conventional Internet, which is significantly slower. |
| **2-5 years** | | | | |
| · Constantly upgrading algorithms. | · May change as we collaborate more with experimentalists and try to examine and interpret their data. | · This is unlikely to change much unless new algorithms generate significant data, which is unclear at this time. | · Same as above unless higher data generation requires more performance. | · Same as above unless new algorithms create larger data sets that need to be dealt with. |
| **5+ years** | | | | |
| · None planned. | · If we can be involved in more collaborative data analysis amongst collaborators, especially experimental and theoretical colleagues, that would be good. | · Same as above, unless we change our operating style, in which case data transfers might grow somewhat, but rates might need to be much faster for collaborative engagement. | · Same as above, unless trying to collaboration in real time. | · Same as above unless trying collaborative data analysis in real time. |

# Case Study 8

# National Center for Electron Microscopy

## 8.1 Background

The primary mission of the National Center for Electron Microscopy (NCEM) at the Lawrence Berkeley Nationa Laboratory is to design, operate and model transmission electron microscopes (TEMs). These microscopes use coherent electron waves to image materials science and biological samples in real and diffraction space, using both plane waves and focused probes. The data collected is entirely digital and requires extensive computer analysis and comes in many different flavors (images, diffraction patterns, spectroscopic data, energy loss curves, etc). All recorded data must be moved from microscopes to analysis computers, user data destinations or supercomputing facilities.

NCEM is one of several DOE electron microscopy centers, all attached to DOE Nanoscience Centers. These microscopy centers exchange resources, ideas and even facilities through the remote microscopy project.

## 8.2 Collaborators

NCEM has 7 full time staff scientists and 7 technical staff. It has recently merged with the Molecular Foundry,[1] which has about 20 staff scientists. NCEM serves approximately 250 users from around the world at any given time. We have worked with NERSC on large-data studies a few times during pilot projects.

## 8.3 Near-term Local Science Drivers

### 8.3.1 Instruments and Facilities

The NCEM is a User Facility operated by the Basic Energy Sciences division of the Department of Energy, located at the Lawrence Berkeley National Laboratory in Berkeley, CA, USA. We operate approximately ten TEMs, which collect anywhere from a dozen 2MB images to dozens of multidimensional datasets up to 0.5 terabytes each, per day. We primarily use the local network or external hard drives to transfer data between experimental microscope computers and analysis computers, on which we use

[1] A DOE nanoscience user facility. http://foundry.lbl.gov.

a variety of software programs to perform analysis. We also simulate various experimental geometries, primarily on workstation PCs located at NCEM. Our facility has approximately 100 terabytes of local storage, spread over a dozen or so workstations.

### 8.3.2 Software Infrastructure

Most TEM data is recorded in proprietary formats provided by the microscope vendors. Vendor software is used to process and analyze most of the data. Additionally we have written several thousand Matlab, python and C++ scripts and programs to perform additional data analysis not covered by vendor software. Virtually all of our simulation and modeling software is open source or developed internally. We have recently introduced a new file format based on HDF5 for large datasets but tools for acquisition, viewing and analysis require further development.

## 8.4 Near-term Remote Science Drivers

### 8.4.1 Instruments and Facilities

Currently we transfer some large datasets to NERSC for long-term storage, but we are not yet performing analysis online at NERSC. These transfers are performed on a data transfer node using 10 Gbps fiber connections set up by ESnet.

Through the remote microscopy project, one of our most advanced microscopes (TEAM) can be operated at multiple satellite facilities. These facilities have a local computer that can mirror the various operation screens, as well as inputs from the mouse, keyboard and specialized analog control paddles. This remote operation requires transmission several large screens and control streams with high-frame rates and low latency (about 5 million pixels worth of screen space at 60 frames per second, with about 50 milliseconds of latency). High latency was primarily responsible for similar programs failing in the past. ESnet provides the network transmission guarantees that ensures the required low latency. This year, NCEM published the first micrographs that were obtained by one of our staff members on the opposite coast of the United States.

### 8.4.2 Software Infrastructure

Some users use GUI remote logins for NCEM workstations to run proprietary software not available at their home institutions. Data backups to NERSC are done over command line or the ESnet and NERSC web portal resources, such as Globus.

### 8.4.3 Process of Science

TEM data often require modeling or simulations to fully interpret. It is therefore important for our users to have access to our codes and the proprietary vendor software. Networking allows our users to remotely access expensive proprietary software and our internal analysis codes. Often TEM data is transferred back several times between our staff and our users at their home institutions, with additional analysis, annotation or reconstructions added at each step.

The usual tool for moving our TEM data and analyses is email. Large datasets are transferred using LBL webspace or commercial tools such as Dropbox or Google Drive. The analysis tools are completely decoupled from the networking tools used to move the data around. We have no centralized repository for either data or analysis codes.

## 8.5 Medium-term Local Science Drivers

### 8.5.1 Instruments and Facilities

With the advent and adoption of new cameras, we expect data rates to skyrocket. The previous generation of microscopes generated approximately 1 TB/year, but we now own a direct electron detector that records 16 TB of data in 15 minutes of experiments. If half of our microscopes have new detectors installed in the next 5 years, we expect data recording rates (of experiments that are successful) to climb to approximately 10 TB/day.

### 8.5.2 Software Infrastructure

We hope to develop some kind of central data repository and move to open source data formats over the next 5 years. We would like to exchange software and best practices with other DOE electron microscopy facilities.

### 8.5.3 Process of Science

We plan on sharing far more of our internally developed analysis and simulation codes with users and collaborators around the world. We have set up a few pilot websites and plan to eventually make all internal software available online.

## 8.6 Medium-term Remote Science Drivers

### 8.6.1 Instruments and Facilities

We will deepen our collaboration with NERSC and move some of the data analysis to purely online solutions. Raw data should go up the pipeline to the supercomputer, and results and numbers should be the only data returning.

### 8.6.2 Software Infrastructure

We currently have no funding to develop better networking or analysis software tools. One key element will be convincing vendors to unlock proprietary data formats or convert to open-source standards. We have made some headway in this regard and expect a lot more over the next 5 years.

### 8.6.3 Process of Science

By following in the footsteps of the X-ray and synchrotron communities, we expect that best practices and analysis codes will be far more widely distributed through the electron microscopy community in the next 5 years. We can make use of many of the networking and web tools developed by these communities for our data.

## 8.7 Beyond 5 years

We will purchase at least one new microscope in the next 5 years. We expect that BES will allocate funding to data processing and this will allow us to develop the tools required to work with the vastly increased flow of data generated by modern microscopes.

## 8.8 Network and Data Architecture

All electron microscopy data should be transferred to a supercomputer, where it can be analyzed and visualized using online tools. This removes the need for our users to purchase or use proprietary software remotely. We hope to work with NERSC to migrate our local analysis codes to run purely at the supercomputing facility and make results from the codes available online via a web interface.

We also plan to expand the remote microscopy program. This would allow the DOE to invest in equipment usable by scientists from all over the world, significantly lowering the barrier to entry for user facility access for those researchers who are not local.

## 8.9 Data, Workflow, Middleware Tools and Services

We currently have no standardized analysis workflow for data. We are not aware of tools beyond networking applications (such as Globus) available via ESnet. Our small facility is not able to develop fully mature software solutions internally. We (and our users) are woefully unprepared to handle the massive increases in data collection enabled by the new generation of detectors.

Ideally we would have some kind of drop-in networking tools than can manage very large volumes of data, with some kind of access control for users and staff. We are motivated to work with ESnet or other relevant DOE organizations to develop and/or test such tools.

## 8.10 Outstanding Issues

Internal data transfer speeds and storage is insufficient for our current experimental data load. We use external hard drives to move data room to room, but without any overarching organization or constant backup practices. Dataset fragmentation is a real problem. We want to build an NAS system to centralize all NCEM data but do not have the expertise to know what hardware to buy, let alone how to set up the internal network properly to connect microscope PCs and user machines.

*Table 8.1: The following table summarizes data needs and networking requirements for NCEM.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · Newest detector can collects 2000x2000 pixel images/s.<br><br>· Typical TEM cameras record ~100 images daily, a few megabytes a piece.<br><br>· Lots of time series, spectroscopy datasets, etc. | · Many different analysis routines and software codes are used, for many different experiments (extremely heterogeneous data).<br><br>· Some open-source software, some proprietary vendor programs and some developed internally. | · 2D images, 2 MB.<br><br>· 3D datasets such as time series, focal series or 3D tomographic data, ~100 MB-1GB.<br><br>· 4D datasets such as 2D grids of 2D probe images, maxing out at 500 GB. | · Images require only seconds, while it can take up to 12 hours to move large 4D datasets.<br><br>· With portable drives we can move anything up to 4TB. | · We have a data transfer node, which combined with ESnet fiber setup gives 1-2 TB transfers in an hour to NERSC tape archive. |
| **2-5 years** | | | | |
| · Additional cameras that will again dramatically increase data collection rates. For example we're developing a detector with 100,000 frames/s (1 MB each). | · Better sharing of software codes<br><br>· More use of open source file formats. | · Same dataset sizes as above, but more of them recorded daily. | · Similar transfer times, unless we build an internal NAS with high-speed network ports. | · Same as above. |
| **5+ years** | | | | |
| · Next generation microscopes that record data in many channels simultaneously. | · Move more codes online to allow remote processing of data at NERSC. This will speed up analysis and allow users to perform analysis without proprietary software. | · Next generation 4D datasets may contain multiple TB each, recorded in 1-5 minutes. Perhaps a dozen of these in one day. | · Need to move many large TB-scale datasets at once from multiple microscope computers to lab workstations or supercomputers.<br><br>· Need method to share large datasets with users. | · We will need to move many large TB-scale datasets at once from multiple microscope computers to lab workstations or NSERC supercomputers. |

# Case Study 9

# National Synchrotron Light Source II

## 9.1 Background

National Synchrotron Light Source II (NSLS-II) will be a new state-of-the-art, medium-energy electron storage ring (3 billion electron-volts) at the Brookhaven National Laboratory in New York. Designed to deliver world-leading intensity and brightness, NSLS-II will produce X-rays more than 10,000 times brighter than the current NSLS. The superlative character and combination of capabilities will have broad impact on a wide range of disciplines and scientific initiatives, including the National Institutes of Health's structural genomics initiative, DOE's Genomics: GTL initiative, and the federal nanoscience initiative. The first six beamlines will use a variety of techniques: submicron resolution X-ray spectroscopy (SRX) analysis pipeline (XANES), inelastic X-ray scattering (IXS), X-ray powder diffraction (XPD) , and coherent soft and hard X-ray (CSX/CHX). HXN uses a number of techniques including: X-Ray fluorescence microscopy, differential phase contrast, tomography with self-absorption correction, image correlation and registration, ptychography and Bragg–ptychography, nanodiffraction and XANES. To effectively use these new capabilities, new methods are required to collect, analyze, store and visualize the data.

## 9.2 Collaborators

The NSLS-II is a user facility that will initially provide the community with 15 beamlines. There are two basic types of beamlines, those that run samples that are submitted and those that support interactive experiments. The beamlines that run samples queue up experiments or experiments are selected through a process and a team of scientists are given access to one of the beamlines for 1–14 days. These groups come from international laboratories, universities, and private industry. Their data is proprietary in most cases. They can be working at a major lab with access to all of the network bandwidth and computing that they are endowed with or they could be at a coffee shop in their university with a laptop.

*Table 9.1*

| Beamline | 1-D Data types | 2-D Data types | 3-D Data types | Analysis Codes | Visualization needs |
|---|---|---|---|---|---|
| **SRX** | Line: ($E_f$, counts)<br>Line: ($E_i$, Absorption) | XY: (x, y, Absorption)<br>XY: (x, y, Fluorescence)<br>Stack: (x, y, Fluorescence, ROI)<br>Stack: (x, y, $E_i$, Absorption)<br>Stack: (x, y, $E_i$, $E_f$ ROI) | | pyMCA | 1-D line plot<br>1-D stack plot<br>2-D image rendering<br>2-D image stack<br>1-D interactive ROI |
| **IXS** | Line: (E, counts)<br>Stack: (N-detectors, counts)<br>Stack: (E, counts, Q) | | | pyMCA<br>1-D Non-linear fitting<br>1-D curve registration | 1-D line plot<br>1-D stack plot |
| **XPD** | Line: (Q, counts/Intensity)<br>Stack: (Q, counts, EV)<br>Line: (2D->0D, EV)<br>Line: (1D->0D, EV) | XY: (x, y, I)<br>Stack: (x, y, I, EV) | -Reciprocal space reconstruction, set of 2-D images to 3-D (h, k, l, counts) | Billinge Group<br>Caking & Integration | 1-D interactive ROI<br>1-D stack plot<br>2-D image stack<br>2-D interactive ROI<br>3-D plotting for space$^{-1}$ |
| **CSX/CHX** | Line: (time, $R^2$)<br>Line: (line from 3-space, I)<br>Stack: (time, $R^2$, EV) | XY: (x, y, I)<br>XY: ($t_1$, $t_2$, $R^2$)<br>XY-Stack: (x, y, I, EV)<br>XY-Stack: (XY-XPCS, correlation)<br>Slice from 3-space: (x, y, I) | -Reciprocal space reconstruction (h, k, l, counts) | pySpec<br>-Andrei's code in Yoric<br>-Existing XPCS code in python/C | 2-D image stack<br>2-D interactive ROI<br>3-D plotting for space$^{-1}$ |

## 9.3 Near-term Local Science Drivers

### 9.3.1 Instruments and Facilities

The NSLS-II storage ring is 792 m in length[1] and there will be 60 to 80 beamlines.[2] NSLS-II is designed to deliver photons with high average spectral brightness in the 2 keV to 10 keV energy range exceeding $10^{21}$ ph/s/0.1%BW/mm$^2$/mrad$^2$. The spectral flux density should exceed $10^{15}$ph/s/0.1%BW in all spectral ranges. This is considered cutting-edge performance and requires the storage ring to support a very high-current electron beam of 500 mA. The ring will operate at 3 GeV, which is considered medium range, yet the brightness and flux are what will make NSLS-II unique. The beamlines will use state of the art area detectors capable of taking 1000 frames per second as well as correlation counters that can resolve time to 125 ns. Motion controllers provide coordinated motion between optics from the insertion devices to the detector with a resolution of 1 ms. The facility timing system is extended to all instrumentation from the storage ring to the detector that time stamps data with 8 ns of resolution. These new capabilities from the X-ray source, to the detectors and coordinated motion control are capabilities that drive data rates.

### 9.3.2 Infrastructure

A modular hardware and software architecture is planned to provide an infrastructure that can grow and evolve over the life of the facility. It must support the new detectors and techniques as they continue to evolve. They must also support analysis of the quality of the data and the performance of the beamline.

### 9.3.3 Software Architecture

The software architecture is built as a set of tools that use standard Application Program Interfaces (APIs). The software architecture is built taking advantage of the tools and communication protocols that exist in the Experimental Physics and Industrial Control System (EPICS). The software architecture supports all data handling from the detector and other instrumentation, through analysis, visualization, and storage at each phase. There are two aspects to this software architecture, the data stores and the processing of the data.

[1] https://en.wikipedia.org/wiki/National_Synchrotron_Light_Source_II#cite_note-NSLS-II_Specs-5.
[2] https://en.wikipedia.org/wiki/National_Synchrotron_Light_Source_II#cite_note-NY_Times_NSLS-II-6.

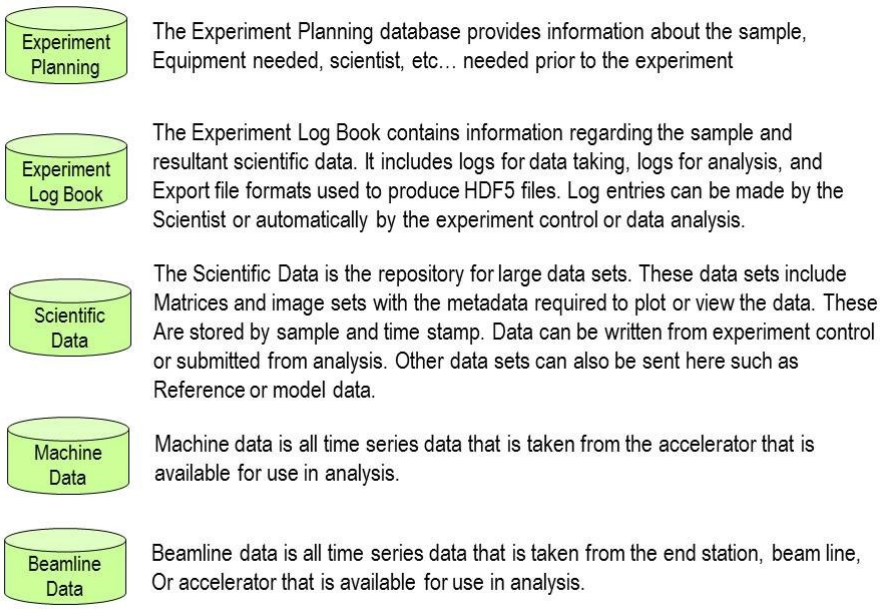



*Figure 9.1: Data stores for scientific use cases. Each is physically separate, and may be optimized for its intended purpose.*

**Data Stores**

A typical experiment involves not only the raw data from a detector, but also requires additional data from the beamline, information from the accelerator complex, logbooks, and administrative information (such as safety reviews or user information). To date, this information is largely kept separated and manipulated individually. A much more effective approach would integrate these different data sources, and make these easily accessible by data analysis codes. This new architecture is a modular design that allows solutions to evolve independently for each of the individual aspects of the problem. This overall approach is easily scalable and exploits existing data analysis codes in the data processing pipeline.

The classes of data are identified along with the use cases for each, starting by defining each of the data stores pictured in FIgure 9.1 These five classes of data storage will likely be comprised of many physical Relational Database and Archive Data instances. To create a modular architecture, these data stores need to have clearly defined interfaces for use by the different applications that require the ability to add or reference the information that is contained in the stores.

Each of these data stores has different requirements for storage rates, retrieval rates, and query capability. In this model, the experiment planning and experiment log book(s) are envisioned to be relational databases that permit data deposits either directly or through applications that allow tracking all of the peripheral information about the experiment, notes from the project team from proposal submission through analysis, history of extraction formats, and all data provenance. The scientific data store is the source of the large data sets. This requires fast access, the ability to make references to the data sets from the relational data stores, and the ability to name and access the data directly. These large data sets should contain all of the metadata needed to understand the data set, but exclude any other peripheral data that is available from the relational databases or the machine data. The machine data is the storage of all dynamic conditions around the detector, the beamline and the accelerator. These data must be available on a name, function or time basis for export to analysis codes.

The use cases for each of these data stores are illustrated in Figure 9.2. These include experiment

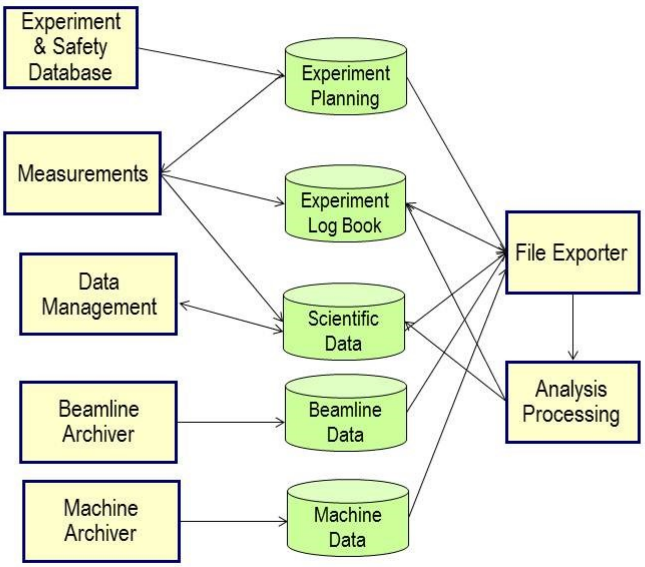



*Figure 9.2: Data store use cases with applications.*

proposal submission, a description of data collection algorithms that are used to create the data, management of the scientific data that is created, a collection of any of the process variables at the facility, beamline, or end station that may be used to help characterize or analyze the scientific data. On the analysis side, users must be able to interactively search their project, sample, and environmental data. They need to be able to create file formats (such as a specific HDF5 format), in which to extract or reference data that can be fed into an analysis processing chain. This analysis processing chain must be able to log the provenance and store the results of the analysis.

Many elements of this architecture exist and are readily available. Computing hardware includes, disks and servers, the size and speed of which would be determined by the scale of the system. Many of the software elements are also available. EPICS provides the machine and beamline archive data, and an electronic logbook on which to build. Data acquisition codes trigger detectors to write to the scientific data store in a format best suited to the detector performance, and existing analysis codes may be used to analyze the experiment data. Networking and data management tools may be required for optimization to assure effective system performance.

The architecture outlined here promotes the development of whatever storage methods best meet requirements, provides the ability to search through all of the experimental data collected or created through analysis, supports the ability to create any file format to meet the input requirements for an analysis code, and supports documentation of provenance in an electronic logbook.

### Software Components

The software components support experiment control, data acquisition, interactive data analysis, and data export. At each level of the software, Application Program Interfaces (APIs), are used to support the incremental addition of functionality. The Experimental Physics and Industrial Control System (EPICS) is used to provide all interfaces to the hardware and make it available over the Channel Access network protocol. The EPICS Version 4 protocol (PVAccess) is used to provide access to complex and large data from either the Input/Output Controller (IOC) or middle layer services that are built on top of network available services such as the Archive Appliance that is used to archive scalar data from the beamline.

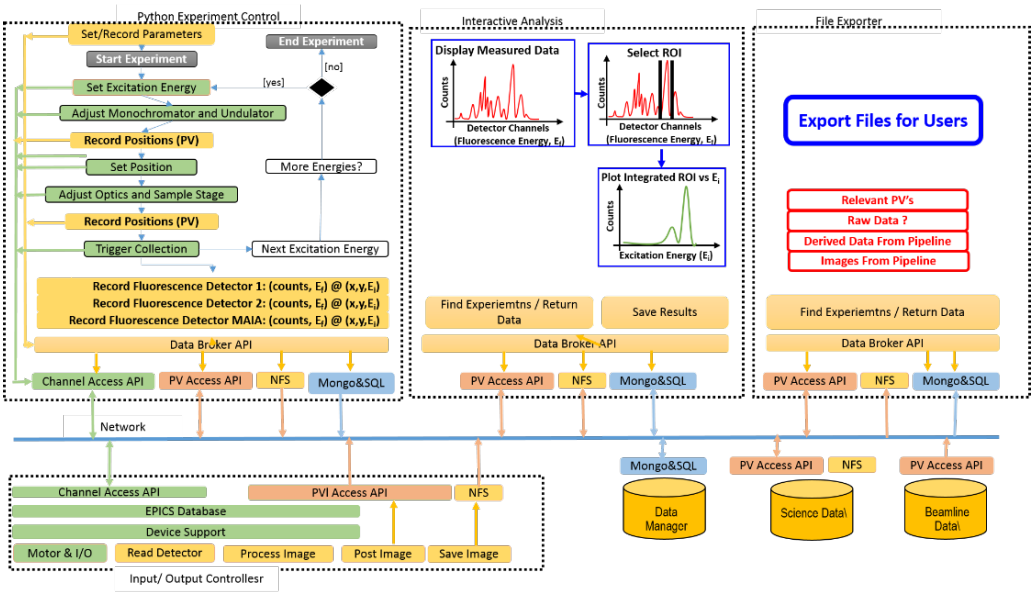



*Figure 9.3*

Run headers, configuration data, and a dictionary of properties are stored in the data manager and searchable by subsequent applications through a network service that is encapsulated by the Data Broker API. All user applications interact with the data stores through the Data Broker API. Experiment control has additional connection to the instrumentation through the Channel Access protocol.

### Hardware

The hardware architecture is present at the beamline in the first-phase protocols. There are file servers for the Data Manager, Science Data, and Beamline Archive Data. The science data will be migrated from the beamline to a central storage facility after the scientist leaves the facility. The Data Management System contains the current location of the Science Data and the Data Broker API gives the user applications access to that data so the relocation of the data is transparent to the user. The hardware layout for the data storage facilities is shown in Figure 9.4.

There are four IOCs: three for controlling all network-based instrumentation that includes motor controllers and industrial I/O. There is an IOC that provides integration into the facility-wide timing system as well as any low-latency I/O required. An IOC is provided for integration of area detectors with the ability to process images in this multicore environment prior to the storage of the large data sets onto disk. Each of these IOCs runs EPICS to integrate the I/O and provide it through the Channel Access and PVAccess. In addition to these services and the I/O, there are the lab-wide PASS system for experiment planning, the Machine Data Archive Appliance that provides all machine historical data to the beamlines users, and a Channel Finder Service to aid the users in discovering the appropriate name of the EPICS process variable among the over 20,000 that will exist on each beamline. The combined software/hardware architecture is illustrated in Figure 9.5.

### Experiment Control

Experiment control consists of several layers: intelligent motor controllers, integration of all control hardware into the EPICS process database, sequences of commands that automate an experiment, and the user interface to observe and control the experiment.

Many motion controllers used at NSLS-II have embedded computers. These controllers are programmed to control the acceleration and velocity of each axis. For most of our optics, virtual axes (such as pitch,

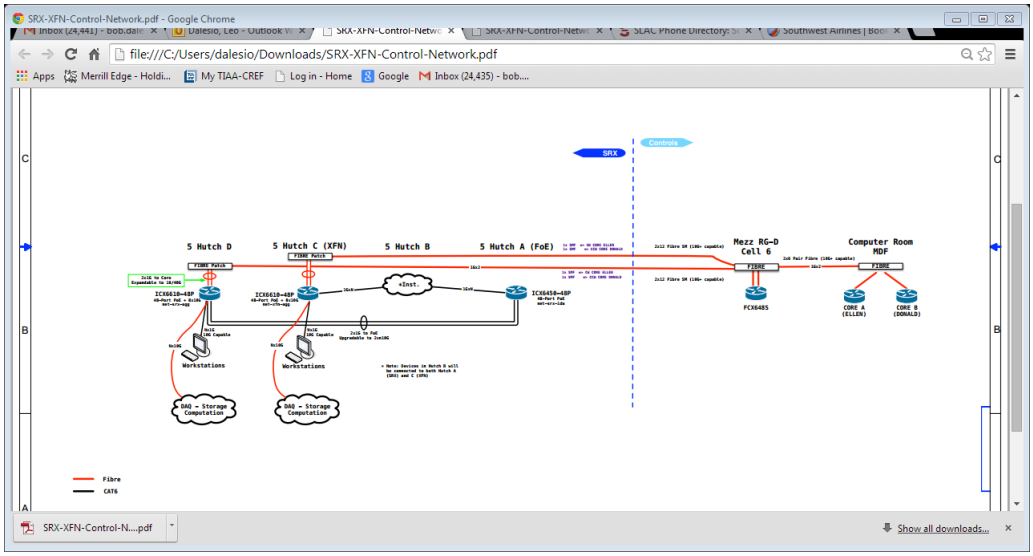



*Figure 9.4*

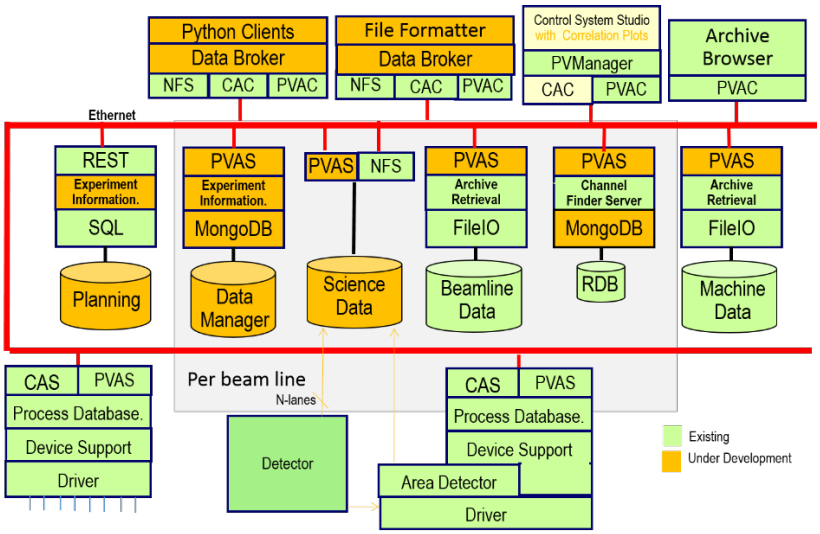
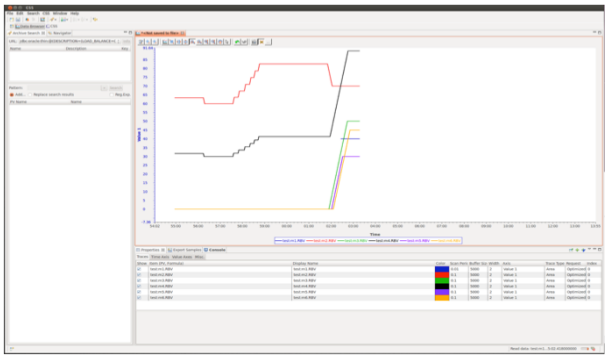
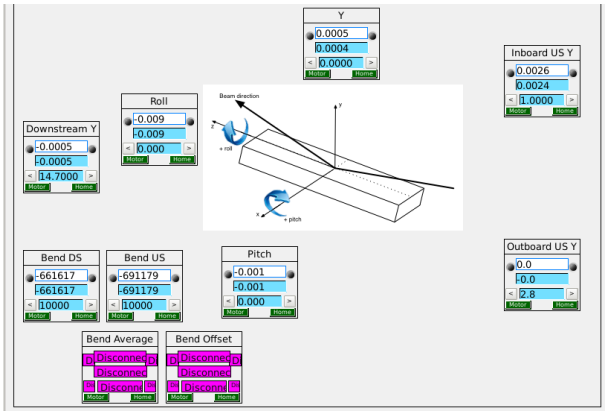



*Figure 9.5*



*(a)*




*(b)*

*Figure 9.6: (a) Data Browser Client to Strip chart Live and Archived Data. (b) 1CSStudio Screen of Optics Motion.*

roll, phi, gamma, etc.) require the coordinated motion of several motors. This coordination can also be programmed into the motion controllers.

All controls and monitors are integrated into the EPICS process database. The process database records process at approximately 1 us per record. Processing consists of obtaining a time stamp from the time server, converting engineering units to raw units, writing outputs, reading inputs, converting to engineering units, checking for alarms, and notifying every client that is monitoring any field in the record that has changed. Each signal or virtual signal is exposed through the Channel Access protocol for read/write access to any client on the network. Writes in this environment are executed across the network in the order of 35 ms. All control parameters and monitor signals are available to be monitored over the Channel Access protocol and are delivered with the facility wide time stamp and the alarm severity. Tens of thousands of monitors can be sent per second. Many clients are available to present alarms, collect history, or present the process variables to the user.

Python scripts will be used to allow scientists full control over the beamline instrumentation. Standard scripts will be maintained in a library that provide a more integrated experience. In the following lines taken during a session, the motor “delta” is moved absolutely to a position and a status is returned. The next command does an absolute scan controlling the motor, delta, over a linear space.

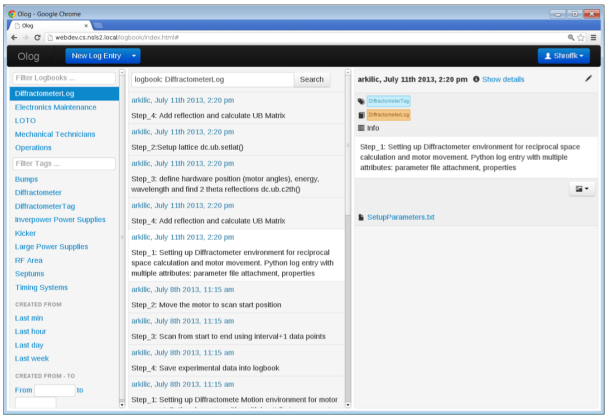
**In [4** ]: move(delta, 12)

**Out [4** ]: 1

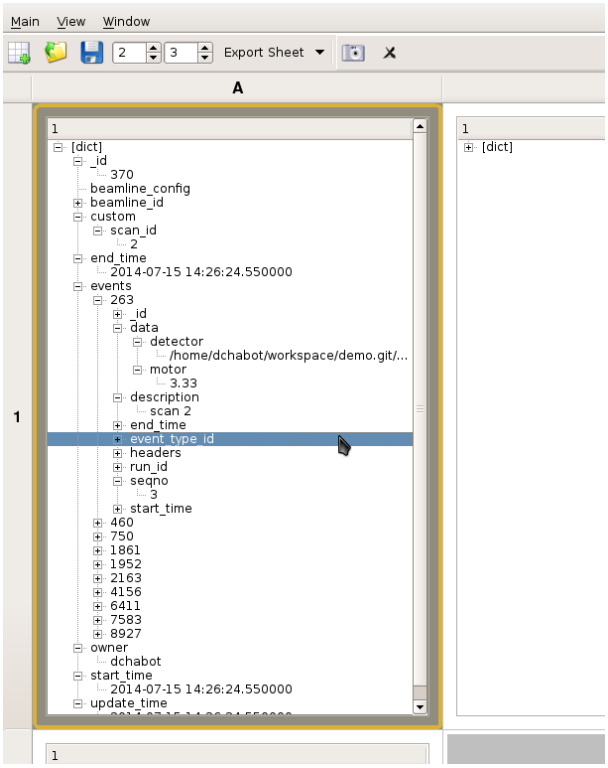
**In [5** ]: ascan(motor=‘delta’, trajectory=linspace(0,10,10))

**Out[5** ]: 1

The macro ascan, also makes records into the LogBook and the MetaDataStore that allow these runs to be found and viewed. There are three Log Viewer applications that can be used to view these entries:


Olog
New Log Entry
Shroffs
DiffractometerLog
Electronics Maintenance
LOTO
Mechanical Technicians
Operations
Bumps
Diffractometer
DiffractometerTag
Inverpower Power Supplies
Kicker
Large Power Supplies
RF Area
Septums
Timing Systems
Last min
Last hour
Last day
Last week
Search
arkilic, July 11th 2013, 2:20 pm
Show details
Info
SetupParameters.txt

*(a)*


Main
View
Window
Export Sheet
A
1
[dict]
_id
370
beamline_config
beamline_id
custom
scan_id
2
end_time
2014-07-15 14:26:24.550000
events
263
_id
data
detector
/home/dchabot/workspace/demo.git/...
motor
3.33
description
scan 2
end_time
event_type_id
headers
run_id
seqno
3
start_time
460
750
1861
1952
2163
4156
6411
7583
8927
owner
dchabot
start_time
2014-07-15 14:26:24.550000
update_time

*(b)*

*Figure 9.7*

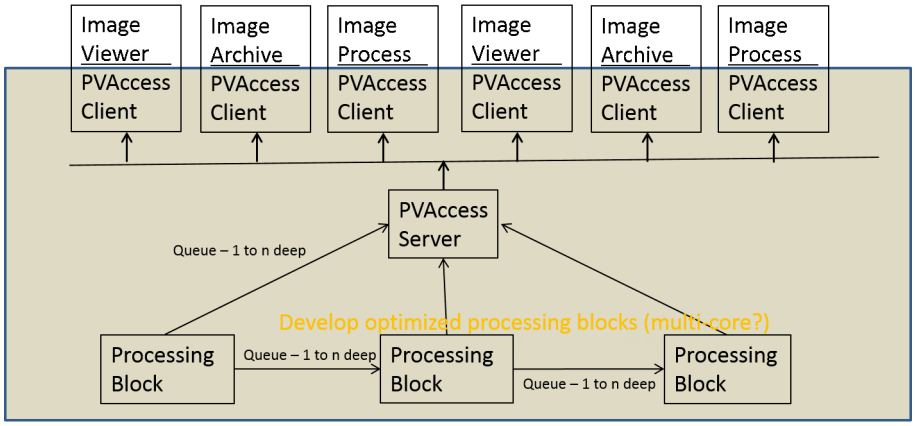



*Figure 9.8*

one in command-line, one in CS-Studio and a web browser application. MetaDataStore and LogBook are accessible through the DataBroker API. The MetaDataStore describes each run that is registered with it. Run header information includes: Run_id, Principal Investigator (PI), start_time and end_time time for each run (which includes multiple scan events), beamline ID, run status (in progress/complete), event header information (ascan has standard header information as do other predefined events), data associated with a series of events within a run (motor coordinates, image URLs, some time series data, and any additional user requested data which can be saved as name/value pairs to search for data such as detector or the value of a given Process Variable (PV)), as well as data analysis results. Time series data can be accessed from Channel Archiver via a python library (which will be seamless through the dataBroker in the future). The configuration data for the run header is currently handled using "conf" text file that includes database authentication and user/beamline information. In the near future, this text-based configuration schema will be replaced with a NoSQL database that holds multiple configuration parameters (i.e., session information) for various users. This configuration database will be flexible and searchable. DataBroker API will serve as the glue between these multiple systems. This experiment control environment provides the scientist with a library of routines that control experiments and store information into standard services that make the data available to the DataBroker.

### Data Acquisition

The data acquisition collects frame rate data from detectors and stores the data onto disk. There is an ever increasing number of high speed detectors. Many of these read the frames directly on to solid state disks or memory that is part of the detector. In our data acquisition, we will integrate all of the detectors through the areaDetector application that runs under the EPICS process database.[3] It provides a number of drivers for existing detectors and cameras. Once the frame data is read into areaDetector, it can be processed through a number of areaDetector plugins that provide some level of simple processing (dark frame subtraction, region of interest, etc). For all of the areaDetector plugins, there is EPICS device support that makes all of the parameters for the filters available through hundreds of EPICS records. This in turn makes all of the configuration available to any EPICS client to allow change, monitor, and archiving. There is a plugin that writes the frame data over the network to disk. The frame data is also available through device support that reads the frame data into EPICS records and makes it available to Channel Access as a byte array. There is also a plugin that makes the frame data available to PVAccess (EPICS V4) protocol which serves the data as a standard data multi-dimensional data type with axes information as part of the data structure. PVAccess is able to multicast the frame data to multiple clients at the same time. Python or Java scripts can be written that monitor these data streams and process them in parallel while being written to disk. As rates continue to increase and the need for immediate

[3] It is an open-source project that is available on Git Hub.

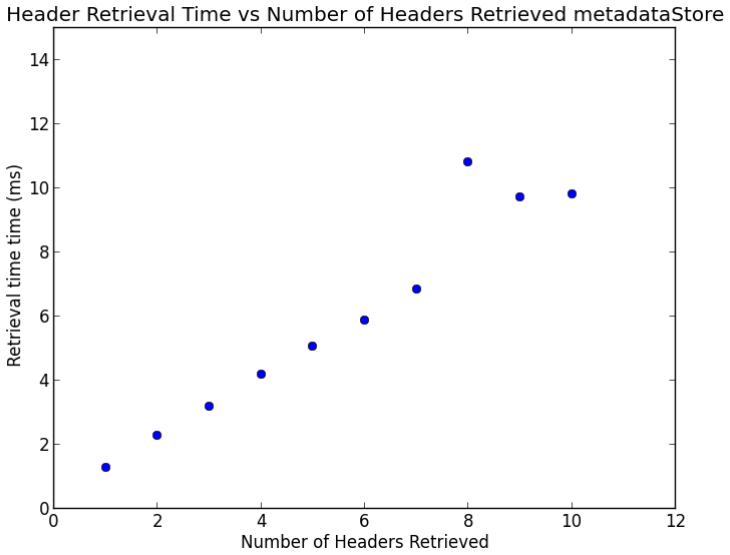



*Figure 9.9*

analysis grows, multi-core processors and large memories can be used to minimize the latency between capturing frame data and processing it.

### Data Management

Data management consists of several areas: configuration data, beamline time series data, machine time series data, science/detector data, and run data. The frame rate data is the large data payloads from the detectors and the storage is under consideration. The configuration and run data are metadata that identifies the science data, including configuration, sample information, and experiment conditions, and are stored in the metaDataStore. The beamline time series data is the subset of the 20,000 signals from the beamline that are archived over time. The machine time series data is the subset of the one million machine signals that are also available. Both machine and beamline time series data are archived in standard EPICS data archiving applications. These data are kept online at the beamlines for the lifetime of the beamline.

The science data store is not yet selected. Under consideration are HDF5 files accessible through some file system, HADOOP, and an EPICS V4 archiver that manages NTNDArrays. The science data will be available for analysis as early as possible as a network Process Variable. It may be first stored on local beamline storage thus allowing dedicated resources at the beamline to be brought to bear for immediate analysis. Once an experiment is completed, the science data will move to a central disk storage for relatively fast access over some period of time. Finally, the science data may be moved to tape for inexpensive long-term storage.

The metaDataStore is the primary source to locate data and a potential source for storing the metadata from analysis. The metaDataStore keeps all configuration data, sample data, or analysis metadata to enable queries to locate specific science data. The metaDataStore does not store the science data, just a URI for the location of the science data. When the science data is moved from local storage, to central storage, and then perhaps tape the metaDataStore simply changes the reference for the frame data to the current location.

The major challenge for the management of data, is the timely retrieval of data sets. In our preliminary tests on data sets, which store all metadata with the frame, searches by run ID, configuration, or conditions, were taking unacceptable portions of time. In a test with a 1000-frame scan that stored the metadata in the frames, the search on energy took most of the 20 minutes required to retrieve, sum

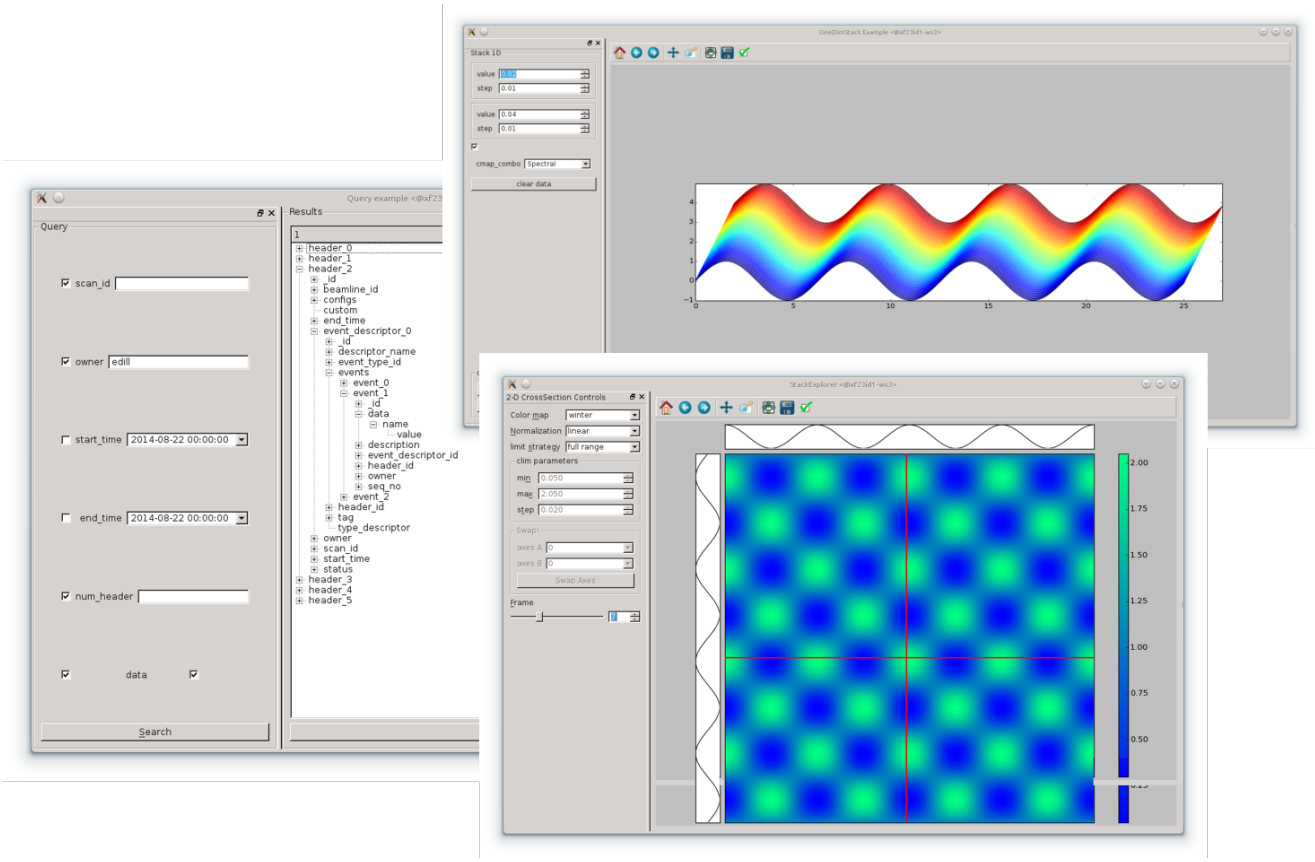



*Figure 9.10*

and plot. Our first prototype for the metaDataStore moved the metadata into a relational database. As we tried to determine what standard metadata would be stored, it became clear that finding a standard set of metadata was a task that would not be achievable in our time frame. We stored all metadata as name/value pairs. We then scaled the prototype to have one-million entries with 120 properties each. Random searches were taking 15 seconds to locate the frame data. With our goal to find, analyze and plot data in a few seconds, we did not find this acceptable. Our latest metaDataStore now uses a NoSQL database, and achieves this test in 10 milliseconds. As frame data is the largest consumer of storage, it is the portion of the data that will be moved. The time series data and the metaDataStore will be maintained on the beamline for at least five years.

### Data Visualization

After a survey of the first set of beamlines built at NSLS-II was done to understand the priorities for visualization, it was determined that support was needed for 1D line, 1D stack, 1D stack waterfall/2D surface plot, 2D, 2D grid, and 2D flipbook. Integrating the metaDataStore to find data and MatPlotLib, a python visualization tool has been put together to search, select, and plot data.

### Data Analysis

There are many techniques and a large collection of analysis codes available in the community. As a first step, the community was canvassed and the analysis needed for early science on the first six project beamlines was identified. There are many codes that are targeted at the same problem. At NSLS-II, we are developing a python environment to incrementally expand a set of analysis routines.

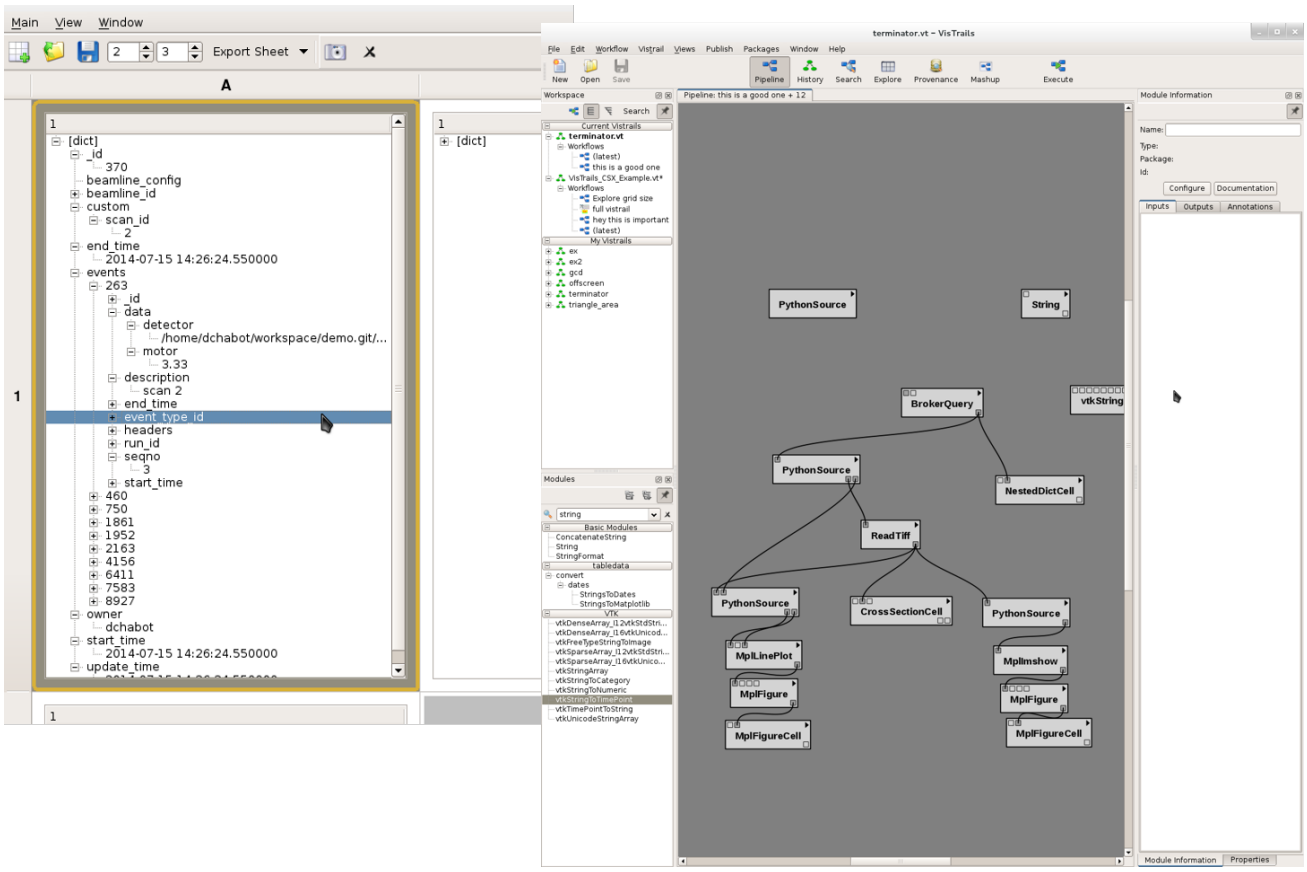

Main View Window
Export Sheet
A
[dict]
_id
370
beamline_config
beamline_id
custom
scan_id
2
end_time
2014-07-15 14:26:24.550000
events
263
_id
data
detector
/home/dchabot/workspace/demo.git/...
motor
3.33
description
scan 2
end_time
event_type_id
headers
run_id
seqno
3
start_time
460
750
1861
1952
2163
4156
6411
7583
8927
owner
dchabot
start_time
2014-07-15 14:26:24.550000
update_time
terminator.vt - VisTrails
File Edit Workflow Vistrail Views Publish Packages Window Help
New Open Save Pipeline History Search Explore Provenance Mashup Execute
Workspace
Pipeline: this is a good one + 12
Module Information
Name:
Type:
Package:
Id:
Configure Documentation
Inputs Outputs Annotations
Current Vistrails
terminator.vt
Workflows
(latest)
this is a good one
VisTrails_CSX_Example.vt*
Workflows
Explore grid size
full vistrail
hey this is important
(latest)
My Vistrails
ex
ex2
gcd
offscreen
terminator
triangle_area
Modules
Basic Modules
ConcatenateString
String
StringFormat
tabledata
convert
dates
StringsToDates
StringsToMatplotlib
VTK
PythonSource
String
BrokerQuery
vtkString
PythonSource
NestedDictCell
ReadTiff
PythonSource
CrossSectionCell
PythonSource
MplLinePlot
MplImshow
MplFigure
MplFigure
MplFigureCell
MplFigureCell
Module Information
Properties

*Figure 9.11*

The routines are being collected, characterized, wrapped for access from python, refactored to conform to our APIs, or rewritten to conform to our performance requirements. This project is being planned and executed with expansion, access for experts, flexibility for the semi-expert, and the ability to wrap analysis in a GUI for the non-expert users. VisTrails is being considered to provide the environment for integrating and deploying these analysis modules at the analysis console. The analysis codes are taking advantage of the services and environment for data management and data visualization. The analysis codes are being written in a modular fashion to maximize reusability and portability to other computing environments.

### 9.3.4 Process of Science

The data comes off of the detectors at each beamline through network ports. Many of the modern detectors have 4 parallel 10 Gbps ports that transfer the local detector data to SSD storage as part of the detector. The frame data is moved from this local disk to storage at the beamline and may be shipped in parallel to a compute processor for analysis. Once the data is move from the SSD detector storage, the next data set can be taken. Each beamline requires dedicated storage and immediate access to processor farms as some partial analysis is needed with the minimum latency. Providing some partial analysis with minimum latency is required to take full advantage of the new flux being provided and the new detectors being used.

## 9.4 Near-term Remote Science Drivers

### 9.4.1 Instruments and Facilities

Remote institutions have a variety of equipment at their disposal. In most cases, a small cluster is a reasonable expectation. They will likely be limited to commercial network connections. Most users are not expected to download their data sets over the network. If they require their data, they will have some sort of disk drives to take their data sets with them when their experiment is over. If their data sets are too large to remove from the facility, then their network is also too small to transport the data set in a timely fashion. It is foreseen that we will provide remote access to data stored at our facility and access to compute servers at our facilities to perform some additional analysis, data reduction, and data formatting. Reduced data sets in custom formats are envisioned.

### 9.4.2 Software Infrastructure

Users will log in to computers at the NSLS-II and will have access to the same software that was available when they were on site.

### 9.4.3 Process of Science

The same tools for data analysis, data management, and data visualization will be used to provide remote access to the data and the analysis environment to the scientists.

The workflow is as described in the three related chapters above. The retrieval and speed of analysis is no longer as critical as it was when the user was on site, as they are not wasting beam time while they wait for results. In this case, the analysis could be queued so that the computing resource at the facility can be shared among users.

## 9.5 Medium-term Local Science Drivers

### 9.5.1 Instruments and Facilities

The current and projected NSLS-II beamlines will support a wide range of disciplines from biology to physics with a wide range of techniques. Higher beam quality and new generations of area detectors are increasing data rates quickly. Increasingly all disciplines are using multi-element detectors to measure and record the results of the experiment. Current state-of-the-art detectors are capable of producing megapixel images with frame rates in the kilohertz range which results in data rates of up to 800 Mbps. Such data rates are fully compatible with current off-the-shelf RAID storage technology, effectively allowing the data to be written to a storage medium as it is produced. The challenge for NSLS-II, however is to be able to *process and analyze* the data at this rate, with these volumes to allow the experimenter to not only make the "first-cut" for decision-making during the experiment, but to retrieve data and analyze it for publication. The solution to this should be compatible with:

- The current, and future user base of NSLS-II,
- The distributed nature of the control system and information stores,
- The large volumes of detector data, and
- A modular flexible approach to analysis.

It is recognized that the user-base of a synchrotron facility such as NSLS-II has shifted somewhat towards users who are not x-ray experts. Increasingly, synchrotrons are becoming tools that are used by scientists in multi-discipline studies with the users being the experts on the science, but not the technique. Furthermore, the research groups accessing the facility do not focus their resources on the necessary data analysis development nor do they have the hardware capable of performing the complex analysis. NSLS-II therefore recognizes that the beamlines must be treated as instruments with the facility providing the necessary resources and infrastructure to provide the user with all the tools to reduce, analyze and present their data in a format ready for publication.

The beamlines at NSLS-II will be among the highest instrumented, and controlled of any worldwide. This has been enabled by two main factors. 1) The maturity of the EPICS tools used to control and monitor the beamline hardware and 2) the reduced cost and therefore availability of high-performance servers and network technology. While it is now possible to record almost the entire state of a beamline during operation, this comes with the challenge that the data stores become much more distributed in nature. The challenge is therefore to synchronize these data stores to allow cross-correlation of the data during the analysis process. This also opens up the opportunity to pull data from other databases such as the safety approval database or proposal database, and make this available to the analysis workflow.

With modern X-ray detectors capable of streaming data at over 800 Mbps management of the detector data will be a crucial aspect of the data collection and analysis architecture. Here the challenge is the lifecycle of the data, from the initial storage of the data at the beamline during the experiment to the eventual long term archiving of the data. To fully cover the data lifecycle many different storage mediums will be used and the data management solution will keep track of where the detector payload is stored. By tracking this detector payload, the user does not need to be aware of where, or on what medium the data resides but can request the data though a standard interface. This allows for a flexible management plan which can both grow with the facilities needs and advantageous changes to storage technology.

Given the breadth of analysis performed at NSLS-II, a flexible solution is needed to cater for all possible data analysis workflows. While often the end analysis is unique to the experiment being performed, there is a great deal of common tasks that are undertaken on the data spanning the breadth of all disciplines. For example, most beamlines with 2D detectors wish to subtract the dark image from the detector, perform a flat-field correction and remove cosmic rays. We therefore propose a modular approach to data analysis with small modules performing operations on the data available to all beamline users. In order to both manage the flow of data through the modules a workflow manager allows the user to

construct their specific workflow. Finally a suite of visualization modules will be available to view the data in 1D, 2D and eventually 3D.

### 9.5.2 Future projections

In order to maximizing the productivity of NSLS-II for beamlines which require complex data analysis and computation it will be necessary to provide resources to users to guide their analysis to publication. We propose a model for the future that draws a parallel to the current model of user access and support at NSLS-II. Currently, each user experiment is assigned a duration of beamline and a "local contact," responsible for aiding the users in their experiment. Due to the complexity of the experiments the local contact is increasingly involved as a collaborator on experiments and often becomes part of the experimental team. We propose that a similar model be applied to data analysis and computation for both access and collaboration.

For experiments requiring more than average computation, users can request analysis time on NSLS-II resources during the proposal process. Here any special needs can be captured early to enable planning for the experiment, not only from a technical aspect but also from a computational aspect. Such a mechanism would also allow access to shared laboratory-wide resources, such as computational clusters. Load managing of these resources also becomes possible in the same way as load managing of resources, such as detectors from a detector pool, are currently performed today.

In order to fully aid users in their analysis and computation, we propose that a data analysis "local contact" be assigned to experiments in the same way a technical "local contact" is today. Such a person would reside most likely in a data analysis group within the Photon Division and would work collaboratively with users to enable their data analysis and computation both at the time of the experiment and once the user has returned to their home institution. For some experiments this may result in a great deal of interaction between the data scientist and has a number of advantages. Firstly, it fully engages the data scientist in the analysis and science being performed which provides good feedback from users to such a group. Secondly, it allows the data scientists to become collaborators on these experiments which is both good for the career path for these scientists and allows for active development of state-of-the-art analysis and computation.

### 9.5.3 Requirements

There are several key requirements that are derived from studying the current planned beamlines.

Data Rates

- Each beamline can produce 8 MB frames up to 1 kfps, with data sets up to 1 TB per minute.
- Detectors buffer all of this data on SDD storage and does not pose a challenge.

Data Analysis

- There are many techniques used to study various samples at NSLS-II.
- Closer to the raw data, many of the algorithms used are common across beamlines.
- There are some very specialized analysis routines required as well.
- Provenance of data must be preserved to validate the scientific process.
- "Scientists would rather use someone else's toothbrush than their analysis routines."
- Users will want to use different, competing, evolved, higher performance, or just better codes.
- To allow access to raw and processed data, a Data Broker API is provided.
- A standard library of python processing routines is produced and supported.
- Processing routines use standard data types.

- Results from processing can be stored back into the data management system.
- A visualization library is supported for 1D and 2D data and data sets.
- A data flow system is being developed to support the development of data analysis. VisTrails is being evaluated as a tool to provide this.

Data Processing

- Scientists may be at the beamline and interact with the experiment control. In this case, we want to minimize the latency required to process the data and provide information.
- Scientists may have submitted samples and interact with a scheduler and analysis report. In this case, we want to batch process the data and provide remote analysis results and reanalysis requests.
- There are many architectural elements being modified or developed to address this.
  - Zero copy transportation of large arrays in front-end computers.
  - Modified EPICS locking to use multicore front ends for array processing.
  - Requests to vendors to provide parallel SFP ports to pipe data into our FPGAs.
  - Multicast network protocol to send data to multiple clients.

Data Management

- 1 data set per minute, with 60 beamlines operating 20 hours per day.
- 72,000 data sets per day, over 2 M data sets in one year.
- 20,000 signals are controlled and monitored on each beamline and recorded as time series.
- A MetaDataStore service is being provided to support the tracking of data.
- A ChannelArchiver is provided to store time series data from the beamlines.
- A Science Data Store is being provided to store large frame rate data sets.

Data Storage

- Data must be kept online for some time and may be stored to tape for some time.
- In 2016 we will need 2 PB to store one year of data, by 2020 35 PB will be needed.
- There is a requirement to track data provenance and store intermediate analysis results.
- Commercial solutions exist for this scale of data.

Data Export

- Scientists need to put together frame rate, time series, and configuration data in specific formats for existing analysis codes.
- Scientists may want to export the data life cycle—from raw data through all analysis.
- The DataBrokerAPI can be used in Python to access all data. Code must be developed to repackage data from different data stores into a desired export format.

Remote data access problem

- There is 1–10 Gbps link into BNL.
- Our users sit at universities, coffee shops, small offices (not on a high performance network port).
- Analysis computers are being planned at BNL.
- A user interface to the analysis codes developed and maintained at BNL is being developed.

### 9.5.4 Software Infrastructure

Over the next 2–5 years, the set of analysis libraries, data management, and data visualization tools available will expand to support new techniques as they arise.

### 9.5.5 Process of Science

Over the next 2–5 years, it is expected that the rate at which detectors can collect data will follow Mohr's Law. The rate and size of the images will double every year. The challenge throughout will be to provide robust and near immediate analysis results as the experiments are being run.

## 9.6 Medium-term Remote Science Drivers

### 9.6.1 Instruments and Facilities

Over the next 2–5 years, the NSLS-II will grow from 12 experimental beamlines to 30–40 beamlines. The number of scientists that will access their data will increase by a factor of 3.

### 9.6.2 Software Infrastructure

The tools are expected to evolve over the next 2–5 years (e.g., beyond the current fiscal year's budget cycle and out to 5 years) and hopefully converge to one analysis code for each function.

### 9.6.3 Process of Science

The process of science of the coming 2–5 years is seen to remain the same. As long as 1 Gbps (or even 10 Gbps) networks are the commercial standard, users will need to use storage and analysis resources at the facility to handle their large data sets.

## 9.7 Beyond 5 years

The storage of all frame rate data for one year is estimates to be on the order of 35 PB to store all of the raw data with zero loss compression and the results of any analysis that has been performed.

NSLS-II will need 70 PB of storage to keep 1 year of scientific data This number is based on current techniques and detectors. The required data growth may be limited by the high cost of replacing detectors (on the scale of millions of dollars). Additional beamlines are assured until the facility is fully populated. Software continues to require authentication and authorization to find and access data. Online analysis capabilities need to grow to support new techniques, increase performance, and expand export capabilities.

Network capabilities within the facility must continue to expand to support the transfer of data from the beamlines to the central storage and analysis facility. Upgrading the network within the NSLS-II facility from the beamlines to central storage will be required. Transferring these files to either USB drives or to the thousands of scientists that we have at their hundreds of institutions is not feasible with current technology.

## 9.8 Network and Data Architecture

Each detector comes with sufficient network and disk architectures to store their data from the detector to solid state disks. After the data is collected, it will be moved to local storage with capacity for 30 days of data for the given beamline. Once the scientists leave, the data is moved from the beamline storage to the central storage. There are 10 Gbps links to each beamline to support this data transfer. The transfer of data does not have a latency requirement. The transfer capability must support the volume of data taken. Once the scientists leave the facility, their data must be available from the central storage for whatever period of time is required by the facility's data policy. Typical data policies indicate 30 days, 90 days, or 1 year.

## 9.9 Data, Workflow, Middleware Tools and Services

The flux of the X-rays are allowing scientist to take clearer images at a much higher rate. New detectors are becoming commonplace that produce an ever increasing frame size and rate, which is currently 8 MB and 1kfps. The limit of the data sets appears to be the solid state disk size that can be used to buffer the data from the detector. These large data sets have made it impractical for scientists to move their data from the facility either with removable media or to transfer the data via the network. As a result, data storage and management is becoming an essential task of the facility. In addition to this, analysis tools and workflow management must also be supported at the facility. The new volumes of data require the development of more time efficient analysis codes. This requires a higher degree of technical skills to implement the analysis codes.

There are two key factors that distinguish light sources from large physics data. First, they are proprietary data sets that represent corporate or intellectual property that must not be made available to anyone outside of the science team. Second, is that there are a very large number of scientists that use these facilities and they live and work virtually anywhere. This means that only a very small part of the data should ever be available and that these small number of large data sets should be accessible anywhere, any time.

## 9.10 Outstanding Issues

The current belief is that all data resides at the facility. However, there is no stated data policy yet. Other facilities have 30-, 90- or 365-day online data policies. In no case, does anyone state that they will guarantee the data against some catastrophic failure. This results in data being stored on RAID arrays, but not redundantly nor backed up. Is this understood and acceptable to the science community or is it the only thing that our budgets allow? If the data must be kept under all circumstances, is there some national facility that would provide a better solution? Does that facility provide the security that is required by this community?

The analysis and visualization codes required for these larger data sets requires more sophisticated software engineering. The number of scientists that are also capable of this level of software engineering is very limited. This software has to be accessible for efficient use by a large community of scientists that range from highly skilled programmers to those that can only interact with a user interface. This tool set does not exist for the materials science community. Building a tool set that serves the most well understood sector in this community has taken over 9 years and was developed for the Mx community. All of the other techniques are less well understood and the efforts to develop some sort of standard are just beginning. There is a huge challenge to provide tools that can take advantage of the new facilities and detectors that have been coming online in the recent years.

*Table 9.2: The following table summarizes data needs and networking requirements for the NSLS-II.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · NSLS II light source ready to accommodate 60 beam lines at 8 hours per day. | · Scientists either send a sample to NSLS II to be analyzed or a team comes to the beam line and is given some period of time from 8 hours to 14 days to perform studies on their samples. | · 100 KB to 1TB. | · Data is stored at each beam line during the experiment. | · The distribution of users and the network infrastructure does not support transfer of data sets to the user. |
| · 15 beam lines in production that will go online in the next 2 years. | · A workflow engine is used to integrate analysis codes. | · Up to 1 TB per minute, up to 8 hours per day. | · Data sets are transferred to central storage to be available to the scientists after they leave the facility. | |
| · A software infrastructure is being developed to manage, find and access data. | · Beam line scientist work with visiting experimenters to take data and perform initial analysis to successfully collect the data required from the sample. Ongoing analysis of the sample is done by the scientists with the data and analysis machines at the NSLS II facility. | · The data sets are 800 MB frames that may be taken up to 1kfps. | · Considering the variety of beam lines, it is expected that 20TB/day will be moved from the beam lines to the central storage. Data will be moved from some subset of the 15 operational beam lines and will require hours. | |
| · A software infrastructure is being developed to manage and integrate existing analysis and visualization codes. | · Many experiments will evolve to perform fly scanning to take advantage of the higher flux and faster detectors. | · In addition, there is related time series data and configuration data. | | |
| · Detectors are available that take 8MB 2D frames at 1kfps with 1TB of storage. | | · Total of 2 PB of storage is anticipated in these early years. | | |
| **2-5 years** | | | | |
| · NSLS II light source ready to accommodate 60 at 20 hours per day. | · More beam lines will move to fly scanning. | · 100 KB to 1TB. | · Data is stored at each beam line during the experiment. | · The distribution of users and the network infrastructure does not support transfer of data sets to the user. |
| · 30 beam lines in production that will go online in the following 3 years. | · With improved techniques for time resolved experiments, more beam lines will use faster detectors to take data frames at the highest possible rates. | · Up to 4 TB per minute, up to 20 hours per day. | · Data sets are transferred to central storage to be available to the scientists after they leave the facility. | |
| · A software infrastructure will be evolving to improve the abilities to manage, find and access data. | | · The data sets are 1600 MB frames that may be taken up to 1kfps. | · Considering the variety of beam lines, it is expected that 90TB/day will be moved from the beam lines to the central storage. The data will be moved from some larger subset of the 30 operational beam lines and will require hours. | |
| · A software infrastructure will continue to evolve to better manage and integrate existing analysis and visualization codes. | | · A total of 35 PBs is anticipated to store 1 year of data. | | |
| · Detectors may be tiled and they will certainly grow as SSD allows larger intermediate storage of data sets. | | | | |

| **5+ years** | | | | |
|---|---|---|---|---|
| · All 60 beam lines will be built out. | · Data analysis will continue to be supported at BNL with scientists able to view, analyze, and manage their results remotely, using the storage and analysis at BNL. | · Data sizes are expected to continue to grow as detectors, SDD, and storage technology allows. | · Time to transfer a data set on the local network must continue to keep up with the rate at which good data is collected. | · Perhaps faster disks and computers will develop that enable us to transfer the scientists' data onto some media that they can take with them.<br><br>· A much faster ubiquitous network would be required to allow scientists to download their data to whatever institution (or coffee house) may suite them. |

# Case Study 10

# Reactive Molecular Dynamics Simulations of Materials

## 10.1 Background

In broad areas such as physics, chemistry, and materials science, there is urgent need for performing large **quantum molecular dynamics (QMD)** simulations, which follow the trajectories of all atoms while computing interatomic forces quantum mechanically from first principles. Recent advances in first-principles-based **reactive molecular dynamics (RMD)** simulations using environment-dependent reactive force fields have further extended the spatiotemporal scales of atomistic simulations to study the intricate coupling of complex materials structures to their thermo-mechanical properties and chemical processes far from equilibrium.

With state-of-the-art O(N) algorithms such as divide-conquer-recombine [F. Shimojo et al, *J. Chem. Phys.* **140**, 18A529 ('14)], where the computational cost scales linearly with the number of atoms, N, there has been dramatic increase of the scale of RMD and QMD simulations. In 2010, scientists at the University of Southern California (USC) performed 48-million-atom RMD simulation on 65,536 IBM Blue Gene/P cores at the Argonne Leadership Computing Facility (ALCF), to study sulfur segregation-induced embrittlement of nickel (Figure 10.1). This is an important problem for the design of the next-generation nuclear reactors to address the global energy problem.

Another important energy application of combined RMD and QMD approaches is ***self-healing nano-***

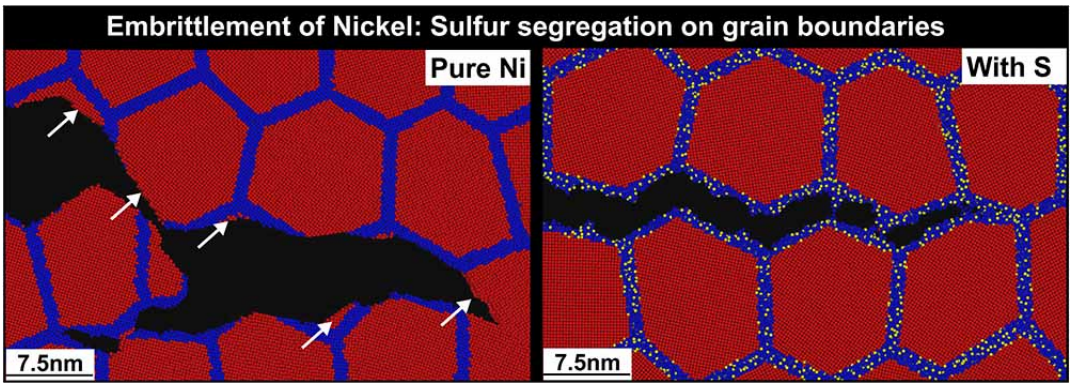



*Figure 10.1: 48 million-atom RMD simulation of fracture in nanocrystalline nickel without and with amorphous sulfide grain-boundary phases on 65,536 IBM Blue Gene/P cores at ALCF [H. Chen et al., Phys. Rev. Lett.* ***104****, 155502 ('10)].*

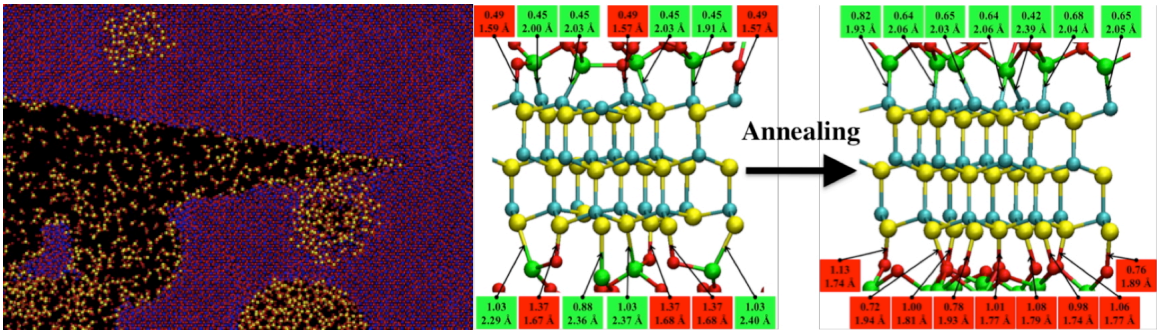



*Figure 10.2: (Left) RMD simulation of crack self-healing by molten silica in alumina. (Right) QMD simulation revealed bond-purification by thermal annealing at a SiC/$Al_2O_3$ interface [K. Shimamura, et al., Phys. Rev. Lett.* ***111****, 066103 ('13)].*

***material systems*** capable of sensing and repairing damage in harsh chemical environments and in high-temperature/high-pressure operating conditions. Self-healing can significantly enhance the reliability and lifetime of materials while reducing the cost of manufacturing, monitoring and maintenance for high-temperature turbines, wind and solar energy, and lighting applications. An example is autonomous self-healing of cracks in alumina ($Al_2O_3$) containing silicon carbide nanoparticles (n-SiC). Near a cracked region, n-SiC gets oxidized in the high-temperature oxygen environment, resulting in the formation of amorphous silica ($SiO_2$). Silica flows into the damage zone, heals the crack, and restores the thermomechanical behavior of the composites. RMD simulation demonstrating self-healing (Figure 10.2) was validated against QMD simulation. The latter also revealed a surprising role of thermal annealing to cause bond-strengthening and bond-purification at a SiC/$Al_2O_3$ interface, accompanied by the formation of an $Al_2O_3$ interphase with a thickness of 2–8 Å. (Figure 10.2).

In 2013, the USC group performed the largest RMD simulation involving a billion atoms on 163,840 Blue Gene/P cores at ALCF to investigate damage caused by a shock-induced collapse of nanobubbles in water near a silica surface (Figure 10.3). They found that the damage on the surface can be mitigated by filling bubbles with inert gas. They also performed a 16,611-atom QMD simulation on 786,432 Blue Gene/Q cores at ALCF (Figure 10.3), and thereby proposed a novel nano-architectual design for a rapid, high-yield production of hydrogen gas from water using Aluminum particles to provide a renewable energy cycle. Their simulation showed that orders-of-magnitude faster reactions with higher yields can be achieved by alloying Aluminum particles with Lithium.

Simulations are performed on local and remote data centers. We have 200 million core-hours/year of computing on the 786,432-core Blue Gene/Q (Mira) at ALCF under an INCITE project. After collecting the simulation data, we perform analysis locally at USC. In the case of simulations performed on remote data centers, they are transferred to USC through WANs. For visualization, we use dedicated computers at our lab. We use the USC campus network (at speeds between 0.1–1 Gbps) to transfer data.

## 10.2 Collaborators

Collaborators:

1. Priya Vashishta, Rajiv K. Kalia, Aiichiro Nakano, Ken-ichi Nomura, University of Southern California
   - Large-scale RMD and QMD simulations
2. Fuyuki Shimojo, Kumamoto University, Japan
   - QMD simulation methods
3. Paul Messina, Nichols Romero, Venkat Vishwanath, Joseph Insley, Argonne National Laboratory
   - Performance optimization, and real-time data visualization and analysis

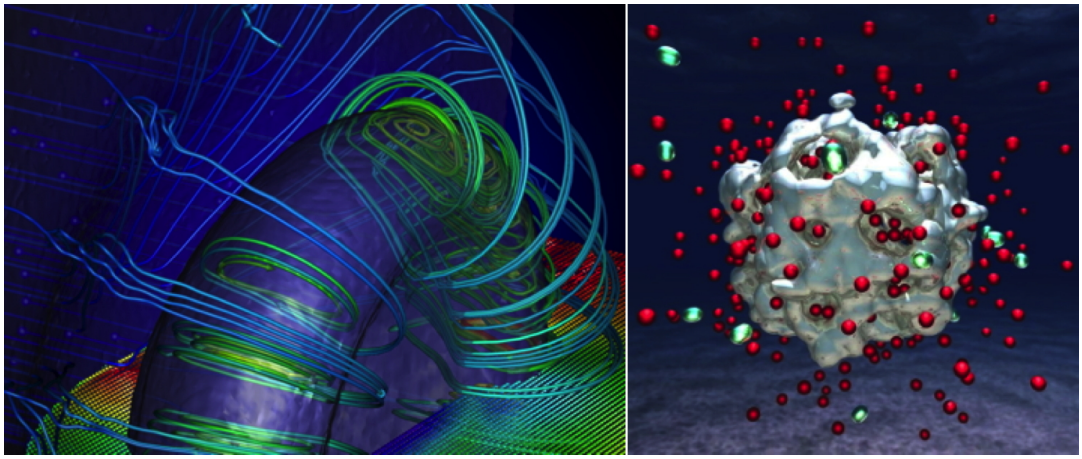


*Figure 10.3: (Left) Billion-atom RMD simulation of cavitation bubble collapse in water on 163,840 IBM Blue Gene/P cores at ALCF [A. Shekhar et al., Phys. Rev. Lett. **111**, 184503 ('13)]. (Right) 16,611-atom QMD simulation of hydrogen-gas production from water using a LiAl particle on 786,432 IBM Blue Gene/Q cores at ALCF [K. Shimamura et al., Nano Lett. **14**, 4090 ('14)].*

4. Manaschai Kunaseth, NANOTEC, Thailand
    - Performance optimization
5. Aaron Knoll, University of Utah
    - Visualization of large datasets

Facilities:

1. Center for High Performance Computing (HPC), University of Southern California
2. Argonne Leadership Computing Facility
    - Mira (Blue Gene/Q) for production run, Tukey (AMD x86_64) for visualization and analysis, and Cetus (Blue Gene/Q) for code development

## 10.3 Near-term Local Science Drivers

### 10.3.1 Instruments and Facilities

We use two major facilities to generate simulation data—USC-HPC and ALCF. At USC-HPC, we have 256 of dual hexcore nodes (3,072 cores) on 10 Gbps Myrinet and 88 of dual octacore (1408 cores) nodes on 56 Gbps FDR Infiniband.

### 10.3.2 Software Infrastructure

We use common Linux tools, such as scp and rsync, to transfer data, and tar and gzip to archive them. To visualize simulation data using the VisIt software,[1] we use the silo format[2]

[1] https://wci.llnl.gov/simulation/computer-codes/visit.
[2] https://wci.llnl.gov/codes/silo/index.html.

### 10.3.3 Process of Science

For simulation data produced at USC-HPC and ALCF, visualization and analysis are performed on two visualization computers equipped with one GPU per node in our lab. Simulation data are transferred to the visualization computers through the USC local network, post-processed, then after creating high-quality simulation movies, they are projected to a tiled display wall by twelve projectors in our visualization room.

Custom software converts the acquired data to HDF5 data files using the standardized NeXus data format[3] and saves to central data storage. Infrastructure based on ActiveMQ has been developed to automatically reduce data on a number of instruments. Data reduction (automated and interactive) on most instruments uses the Mantid toolkit.[4] Data are cataloged using ICAT.[5]

## 10.4 Near-term Remote Science Drivers

### 10.4.1 Instruments and Facilities

We perform small-to-medium size simulations at USC-HPC first for testing, then moved on to ALCF's Blue Gene/Q for large-scale production runs. Initial configurations for the production runs may also be generated using the USC cluster (e.g., 80 GB for the one-billion-atom nanobubble collapse simulation in Figure 10.3) and transferred to ALCF via WANs.

### 10.4.2 Software Infrastructure

We mostly use the same set of tools as described in Section 10.3.2. We plan to use Globus[6] for large dataset transfers.

### 10.4.3 Process of Science

To minimize unnecessarily data transfer, we plan to use VisIt or Paraview as a remote rendering tool from the USC-HPC cluster.

## 10.5 Medium-term Local Science Drivers

### 10.5.1 Instruments and Facilities

USC-HPC staff are currently working toward an intercampus high-speed network called the Trojan Express Network II (TEN-II), providing 100 Gbps bandwidth to transfer data between the University Park Campus, Health Sciences Campus, and Information Sciences Institute. This network will enable rapid data transfer between the USC data center and the visualization room, or real-time streaming of the rendered images from the computing cluster without moving simulation data.

[3] www.nexusformat.org.
[4] www.mantidproject.org.
[5] www.icatproject.org.
[6] https://www.globus.org.

### 10.5.2 Software Infrastructure

We expect more data-intensive, high-throughput simulations based on scientific workflow systems such as Pegasus and Swift. Also, we plan to perform parallel replica dynamics (PRD) simulations to study long-time dynamics, where thousands of simulation instances are coordinated within the scientific workflow (see Figure 10.4).

### 10.5.3 Process of Science

The PRD simulations would generate intermediate data for each simulation instance. Efficient data management is necessary to achieve enough data exchanges with replicas to detect rare events.

## 10.6 Medium-term Remote Science Drivers

### 10.6.1 Instruments and Facilities

Once the simulation data have been stored on disk, continuous growth in the simulation data size will make analysis and visualization almost impossible at the leadership-computing-facility-scale. Real-time or *in-situ* analysis will enable us to check errors, investigate results, and visualize simulation with sufficient temporal resolutions.

### 10.6.2 Software Infrastructure

GLEAN will be used to accelerate I/O and simulation-time data analysis based on its flexible and extensible framework design.

### 10.6.3 Process of Science

Using Mira, we will first investigate the concept of the staging cluster, where simulation data being continuously transferred from a more powerful computing cluster.

## 10.7 Beyond 5 years

***In-situ* data analytics:** The goals of RMD and QMD simulations include finding rare but interesting events in simulation output that may represent key material processes underlying material properties of interest, discovering common sequences of events, and discovering causality among events. Currently, we employ data-sampling techniques to decrease the size of simulation output to make it possible to store, transfer, and post-process the output data. In future, we envision *in-situ* data analysis to include all spatiotemporal trajectories for higher quality knowledge discovery.

**Parallel replica dynamics simulations:** To study long-time dynamics of materials, we will use massively PRD simulations, where thousands of simulations run in parallel (Figure 10.4). In this high-throughput computing approach, the results of these replicated simulations are analyzed and correlated, and the results of that analysis drive the next set of replicated simulations. This would require machine-learning approaches to automatically detect key material behavior events and provide feedback to simulations run on remote supercomputers, thereby significantly increasing the network requirement.

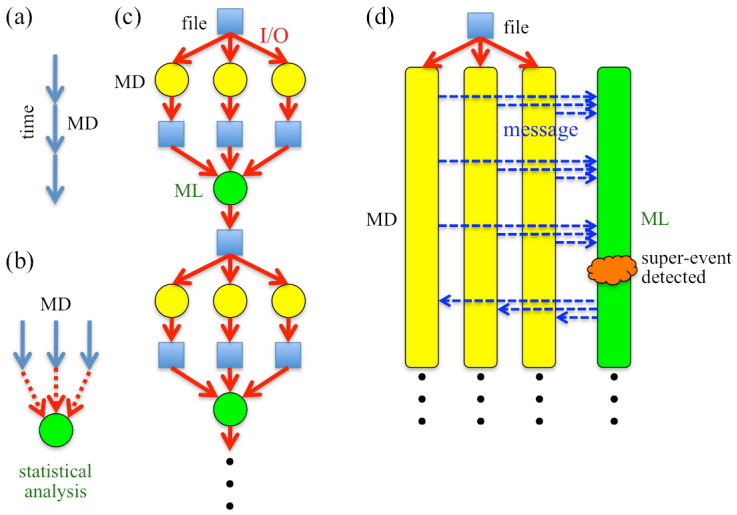



*Figure 10.4: (a) Conventional molecular dynamics (MD) simulation has sequential time dependence. (b) parallel replica dynamics (PRD) predicts the long-time behavior through statistical analysis of multiple parallel MD trajectories. Conventional file-based (c) and new in situ (d) PRD simulations where ML represents machine-learning tasks.*

## 10.8 Network and Data Architecture

Streaming data through TEN-II network and ESnet, we will be able to monitor simulations and perform analysis in real-time at USC while the simulation is running on the LCF supercomputer. Also, steering simulations via real-time feedback to the simulation remotely would be possible.

## 10.9 Data, Workflow, Middleware Tools and Services

Tools examples include Globus or other data transfer tools, automated data transfer toolkits, distributed data management tools, etc.

## 10.10 Acknowledgements

This research was supported by the DOE BES Theoretical Condensed Matter Physics Grant Number DE-FG02-04ER46130. The computing resources for this research were provided by a DOE—Innovative and Novel Computational Impact on Theory and Experiment (INCITE) award.

*Table 10.1: The following table summarizes data needs and networking requirements for molecular dynamics simulations at the University of Southern California.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · USC HPC and Blue Gene/Q at ALCF.<br>· VisIt and Paraview for visualization, Globus for data transfer. | · Small-to-intermediate size simulation on the USC HPC cluster and large-scale production run on Blue Gene/Q at ALCF. Visualization, rendering and analysis on Tukey at ALCF and HPC USC cluster. | · 80 GB per frame.<br>· Current maximum data size is 300 TB depending on the number of frames to be saved.<br>· 1,000–10,000 files and 1–80 GB each file. | · The data transfer rate on USC campus is 1-10MB/s, which gives 100 GB-1 TB/day transfer rate between USC HPC and lab computers or the visualization room. | · The data transfer rate between ALCF and USC HPC is a few MB/s. 100 GB data takes a day to transfer. |
| **2-5 years** | | | | |
| · LCF and USC HPC.<br>· GLEAN for simulation-time in-situ data analytics, Pegasus and Swift for high-throughput simulation and time accelerated dynamics simulation. | · Parallel Replica Exchange on Scientific Workflow. Simulation-time *in-situ* data analysis and visualization. | · 80 TB per replica exchange iteration.<br>· 80TB–800TB.<br>· 1000 files per iteration and 80 GB per simulation instance. 1-10 frames to be saved per simulation. | · TEN-II network provides 100 Gbps as the theoretical peak bandwidth. 2-20 hours to complete transferring 80-800 TB data. | · Connecting ESnet and TEN-II networks, similar ETA to LAN transfer time can be achieved. |
| **5+ years** | | | | |
| · LCF and USC HPC. | · Steering molecular dynamics simulation. | | | |

# Case Study 11

# Spallation Neutron Source and High Flux Isotope Reactor

## 11.1 Background

The Spallation Neutron Source (SNS) and the High Flux Isotope Reactor (HFIR) are DOE BES Scientific User Facilities at the Oak Ridge National Laboratory. These facilities are operated to support research using neutrons in a variety of science areas including structural biology, crystalline materials, polymers, magnetism and superconductivity, chemistry, and disordered materials.

Experiments are typically 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 is scattered and 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 collects event data showing the position of the neutron on the detector array and its time of flight (energy), along with metadata for the sample's environment (temperature, pressure, electric or magnetic field, etc.) and instrument configuration (neutron choppers, motion controls, etc.). The data are processed to produce an HDF5 data file using the NeXus data format. A single measurement is called a "run," and an experiment is typically composed of a number of runs. Collections of runs may be grouped into a "reduced" data set. As reduced data sets tend to be histogram sets and thus include all the zeros. They can be an order of magnitude larger than the raw data set.

SNS data sets tend to be larger than those produced at HFIR. For the lower range of SNS, a run data set may be on the order of tens to hundreds of MB, and on the upper end, a run data set may be many GB. The highest data rate SNS instruments are capable of producing data sets approaching 1 TB. An imaging and tomography beam line is being planned which will have the potential to produce larger data sets.

The facility's data network is largely intra-ORNL at present. Data flow from the instruments to a centralized data management system and are processed either locally on SNS computers or on other ORNL computers SNS provides computing resources and software for users to process and analyze their data. These computing resources are available while the user is on-site conducting the experiment and remain available for remote login afterward.

New horizons in science give us insights into more network-centric science research environments. Researchers more commonly use multiple instruments at the same or different user facilities to shed light on the materials problem under investigation. For example, crystallographers often perform experiments

at the APS or other light sources prior to coming to the SNS where they perform neutron scattering measurements to co-refine the structures using their data obtained from both techniques.

SNS is developing the ability to stream live data during an experiment, which could be streamed to high performance computing resources to perform near real-time data analyses in combination with online simulation and modeling. Near real-time computing during an experiment could be used for determining structure, orienting the sample within the beam, or for controlling or sequencing more complex experiments. As a recent test case, molecular dynamics simulations were carried out on EOS at the same time phonon dispersion measurements of $SrTiO_3$ were carried out on the HYSPEC instrument (a hybrid spectrometer). These types of studies would mean the user is not only interested in passing around a single 100GB data set, but also a similar sized simulation data set. Initially these studies were all carried out at ORNL. However the computing resource could be NERSC or ALS as well.

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.

## 11.2 Collaborators

Between the two user facilities we have roughly 1000 people who wish to access their data remotely. They come from national and international research facilities. They have a varied computational skill set and resources at their home institution. Some are happy with an *NxClient* remote login interface into our systems, while others prefer to analyze their data on the home institution.

## 11.3 Near-term Local Science Drivers

### 11.3.1 Instruments and Facilities

SNS provides shared analysis computing resources for its users. A typical analysis computer has 32 cores and 256 GB of RAM. A computing cluster is available for processing larger data sets. The centralized data storage is a Lustre Parallel File system with 1.1 PB of data storage. A remote login infrastructure provides remote users access to both data and computing resources.

As part of an ORNL-wide initiative, the SNS is an active part of the Compute and Data Environment for Science (CADES) described in the following paragraph.

ORNL hosts a number of signature facilities and projects that produce, analyze, and steward a wide diversity of data. These include the Spallation Neutron Source (SNS), the Oak Ridge Leadership Computing Facility (OLCF), and the Center for Nanophase Materials (CNMS), the Consortium for Advanced Simulation of Light Water Reactors (CASL), the Atmospheric Radiation Measurement (ARM) archive, the Carbon Dioxide Information Analysis Center (CDIAC) and other growing initiatives such as our work with the Center for Medicare & Medicaid Services (CMS). In addition to these facilities and projects, ORNL has a wide range of other research projects that are increasingly reliant on computing and data-intensive capabilities. While each of these centers and projects has a unique mission, they share common requirements and needs in computing, data processing, and management. Perhaps even more broadly, these initiatives share common needs in computing and data-intensive capabilities that can be more efficient and cost effective if they leverage common technologies and expertise in computing and data science. These capabilities include data analysis, data fusion, data mining, search and discovery, visualization and many others. CADES provides a new local-environment for scientific discovery enabling scientists to free themselves from the difficulties of trying to manage, manipulate and process large data sets in order to concentrate on extracting the scientific meaning of the theoretical, experimental, and observational data.

### 11.3.2 Software Infrastructure

Custom software converts the acquired data to HDF5 data files using the standardized NeXus data format (www.nexusformat.org) and saves to a central data storage. Infrastructure based on ActiveMQ has been developed to automatically reduce data on a number of instruments. Data reduction (automated and interactive) on most instruments uses the Mantid toolkit.[1] Data are cataloged using ICAT.[2]

### 11.3.3 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 describes 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

## 11.4 Near-term Remote Science Drivers

### 11.4.1 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.

### 11.4.2 Software Infrastructure

Users can remotely access data analysis clusters and the installed software through *NxClient* and transfer data off-site through a variety of standard tools such as sftp, scp or download through a web portal. For larger volumes we recently established a Globus online endpoint allowing access to SNS data.

[1] www.mantidproject.org
[2] www.icatproject.org

### 11.4.3 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.

## 11.5 Medium-term Local Science Drivers

### 11.5.1 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.

## 11.6 Medium-term Remote Science Drivers

### 11.6.1 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 on particular aspects of their research and 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.

### 11.6.2 Software Infrastructure

The typical facility user will utilize multiple facilities across the globe. A 2012 user survey, including users from 9 European photon facilities, 6 European neutron facilities, and 1 US neutron facility (SNS) is available at http://pan-data.eu/node/99. It shows for example, that 15.4% of the SNS/HFIR users used at least one other neutron source in Europe, and 5.8% of the SNS/HFIR users used at least one European photon source. A study across U.S. neutron and light sources is in progress.

Additionally more extensive simulations are used to compare neutron (and other) experimental results. As part of the Center for Accelerating Materials Modeling (CAMM) using SNS data, we have developed a prototype framework allowing the refinement of force fields in classical molecular dynamics (MD) simulations based on quasi-elastic neutron scattering data measured on the BASIS instrument at the SNS.

Materials science of the future will utilize data from multiple user facilities coupled with simulation results using multiple compute resources. Our infrastructure needs to allow for efficient access (single sign-on if possible) to all those data and resources.

### 11.6.3 Process of Science

In the next 2–5 years, users will expect there to be something which could be characterized as "Google Science" available to help facilitate their research. This will probably be a cloud-computing-based envi-

ronment 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.

## 11.7 Beyond 5 years

Beyond five-years, 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. If this is the case, users will rely more upon their own computing resources to perform tasks rather than remote computing resources to accomplish this. This 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.

## 11.8 Data, Workflow, Middleware Tools and Services

Tools examples include Globus or other data transfer tools, automated data transfer toolkits, distributed data management tools, etc.

For the SNS and HFIR, emerging services such as commercial cloud computing, storage, etc. include 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. More specifically many of our users' skill sets are outside of the high performance computing realm. Therefore we need middleware tuned more towards a typical computer user and take care of the high-speed networking behind the scene.

Cloud computing is not yet in mainstream planning for HFIR and SNS, however these could be promising for facilitating research if sufficient bandwidth between these resources and the data repository can be achieved. Cloud computing in the sense of data on the cloud and remote services like NX and the web monitor are scratching the surface.

There is also an increasing interest in leveraging collaboration tools such as Skype, Remote Desktop, and remote application sharing, web-based notebook tools like Evernote or Google Docs.

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 located at the instruments.

## 11.9 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 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.

*Table 11.1: The following table summarizes data needs and networking requirements for the SNS.*

| 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 has 17 instruments in the user program. | · Data reduction<br>· Visualization<br>· Experiments performed on-site by researchers at the facility.<br>· Concurrent simulations performed at HPC facilities. | · 1 to 2 TB/day<br>· 1000 to 10,000 files per day | · 1Gbps networking utilized.<br>· Initiate 10Gbps networking. | · 1Gbps via traditional network.<br>· >10Gbps via ESnet DTN.<br>· A few data transfers between ORNL and other ESnet or Internet2 sites. |
| **2-5 years** | | | | |
| · ~18 to 20 instruments in the user program. | | · 2 to 4 TB/day<br>· 1000 to 10,000 files per day | · 10Gbps networking standard.<br>· Multiple 10Gbps lines supporting high-volume instruments individually. | · >10Gbps available to all users.<br>· Data transfers occurring more regularly between SNS, and x-ray sources.<br>· Data transfers with Leadership Computing Facilities.<br>· Initiate international collaborations at 1Gbps. |
| **5+ years** | | | | |
| · 24 at first target station instruments in the user program (beyond 24 is with second target stations).<br>· Begin construction on the second target station in this timeframe, which would eventually double the number of instruments at SNS. | | · 5-10 TB/day<br>· 10,000+ files per day | · Upgrade internal networking to 40 Gbps or higher. | · Utilize 40 Gbps or higher between DOE user facilities.<br>· Utilize 10Gbps between international facilities. |

# Case Study 12

# Stanford Synchrotron Radiation Lightsource

## 12.1 Background

The Stanford Synchrotron Radiation Lightsource (SSRL) is a pioneering synchrotron radiation facility known for outstanding science, technological innovation and user support. Now in its fifth decade, SSRL is still evolving and is well-positioned to make significant contributions to scientific discovery for decades to come.

SSRL provides extremely bright X-rays that scientists use for a wide range of research that probes matter on the scales of atoms and molecules. Studies target advances in energy science, human health, environmental cleanup, nanotechnology, novel materials and information technology, among others. As one of five light sources funded by the U.S. Department of Energy Office of Science, SSRL enables research that benefits every sector of the American economy. SSRL also provides unique educational experiences and serves as a vital training ground for the nation's future scientific workforce.

SSRL operates approximately 9 months each year with an extremely high reliability – delivering more than 95% of scheduled x-ray beam time. Access to the SSRL is competitive based on peer-review. Users disseminate their findings through talks and publications, and new highlights are featured monthly. Experiments conducted at SSRL resulted in 11,600 scientific publications since 1974.

## 12.2 Collaborators

Of the approximately 1,700 scientists who annually participate in experiments at SSRL, over 30% are first-time users, and over 40% are postdoctoral associates or graduate and undergraduate students. Each researcher completes training to operate SSRL equipment. Students and new researchers learn from and assist more senior researchers with taking measurements and analyzing data during their experiments. The personal experience received during beam time benefits each participant by providing a framework and understanding of the mechanisms and difficulties that can be encountered during research experiments.

Scientific users also have the opportunity to participate in schools and workshops which provide in-depth, hands-on experience with specific data acquisition and analysis techniques as well as practice with sharing research findings through talks and poster presentations. SSRL research results in approximately 500 scientific papers annually, which includes about 20% that are theses prepared by students who relied upon access to SSRL to complete their dissertations.

With this diversity of users and the general level of experience in the science, the experimental techniques and the computing, training is an important issue and we would plan to increase our ability to create web-based video training.

## 12.3 Near-term Local Science Drivers

### 12.3.1 Instruments and Facilities

SSRL generally operates November through August, using the shutdown period for upgrades and maintenance projects.

The supported technique(s) (and associated beamlines) are outlined as such:

- X-ray Absorption Spectroscopy
  - Biological x-ray absorption spectroscopy
  - Materials / catalysis / chemistry x-ray absorption spectroscopy
  - MEIS x-ray absorption spectroscopy
  - X-ray absorption spectroscopy imaging
  - Single crystal x-ray absorption spectroscopy
  - Grazing incidence x-ray absorption spectroscopy
  - Tender x-ray absorption spectroscopy
  - Tender x-ray absorption spectroscopy imaging
  - Photoemission spectroscopy
  - X-ray absorption spectroscopy, near edge, soft energy
  - Elliptic polarization, soft energy photoemission spectroscopy
  - Ultra-high energy resolution photoemission spectroscopy
  - Angle-resolved photoemission spectroscopy
- Imaging / Microscopy / Microprobe
  - X-ray absorption spectroscopy imaging
  - Tender x-ray absorption spectroscopy imaging
  - X-ray microscopy (hard x-ray)
  - Transmission x-ray microscopy (soft x-ray)
- Macromolecular Crystallography
  - Multi wavelength anomalous diffraction (MAD) and monochromatic crystallography
  - Single wavelength anomalous diffraction (SAD), limited MAD, and monochromatic crystallography
  - Microbeam macromolecular crystallography
- Scattering / Diffraction
  - Biological small angle x-ray scattering (SAXS)
  - Macromolecular solution x-ray scattering
  - Materials small angle x-ray scattering

  - Time-resolved biological SAXS
  - Fiber diffraction
  - Lipid membrane diffraction
  - Small-angle single crystal diffraction
  - Powder diffraction
  - Thin film diffraction
  - X-ray scattering
  - X-ray diffraction
  - Coherent soft x-ray scattering
- Other Techniques
  - White light station

Some of the groups at SLAC who perform experiments at SSRL beamlines have local computing resources.

### 12.3.2 Software Infrastructure

The software infrastructure varies by beamline. The data acquisition systems are typically windows-based. Many of the beamlines have custom data analysis software developed by the instrument scientists. This software is used to format the data, provide fast feedback and to enable preliminary analysis. Training in these tools is offered.

Transferring data to portable media such a storage device or laptop is available in the control rooms. Depending on the size of the files, they may be distributed as email attachments. Computing and software infrastructure are not considered a barrier to scientific productivity.

### 12.3.3 Process of Science

The process of science is proposal driven, with beam time awarded to user groups. Those users might not need to be physically at SLAC to take date due to the remote source handling capabilities at many beam times. Once the user has the data, processing takes place, the data is analyzed and papers are published. The tools and software used are based on beamline tools or may be passed down within a research group. Within some disciplines, community databases are used as repositories. Examples of these include the Protein Data Bank for solved structures.

While not a network issue, providing simple-to-use authentication and authorization procedures that would federate across laboratories could be a gain.

## 12.4 Near-term Remote Science Drivers

### 12.4.1 Instruments and Facilities

The SLAC-based scientific groups have access to in-house computing resources.

### 12.4.2 Software Infrastructure

Data is often transported to remote locations by portable media loaded at the beamlines. The recommended tools at SLAC for WAN transfers for exporting the data offsite are bbcp and Globus. Other tools are in use as well.

## 12.5 Medium-term Local Science Drivers

### 12.5.1 Instruments and Facilities

Several new beamlines will come online in the next few years. These beam lines are not data-intensive, though.

The Stanford Research Computing Center (SRCF) will become the computing center for the Stanford/SLAC joint institutes: The Stanford Institute for Materials and Energy Sciences[1] (SIMES), PULSE Institute for Ultra Fast Energy Science[2] and SUNCAT Center for Interface Science and Catalysis.[3]

### 12.5.2 Software Infrastructure

As needed, the software infrastructure may incorporate additional functionality that leverages LCLS software infrastructure or in some cases, the infrastructure developed at other laboratories. It is expected that this will be driven by scientific needs and user requests, rather than by large scale changes in the data rates.

### 12.5.3 Process of Science

Within the process of science, we expect to see a tighter coupling of the experimental work and theoretical predictions. The Stanford and SLAC joint institutes provide a local example of such integration.

High throughput characterization (HTC) is one methodology for facilitating the process of material discovery and will identify new structure-property relationships to inform computational materials design. In HiTp, “libraries” of combinatorial materials are produced across a single substrate, enabling the researchers to probe the properties of a continuum of varying compositions under controlled production and experimental conditions. The materials libraries are “characterized” at synchrotron light sources (such as SSRL) using X-ray diffraction (XRD) techniques, systematically varying parameters such as temperature that can affect the structural and performance parameters of the materials. HiTp requires interplay between materials library production, the XRD characterization, and the use of sophisticated data mining algorithms to extract unforeseen features and relationships. Developing the software tools to manage the data, to analyze characterized materials libraries and developing simulations to predict the properties will require significant intellectual effort across multiple domains.

[1] http://simes.stanford.edu

[2] https://www6.slac.stanford.edu/blog-tags/pulse-institute-ultrafast-energy-science

[3] http://suncat.slac.stanford.edu

## 12.6 Medium-term Remote Science Drivers

### 12.6.1 Instruments and Facilities

In addition to increased sophistication of the detectors and data, it seems likely that some user groups will increase the use of numerically intensive calculations such as density functional theory (DFT) codes and *ab-initio* quantum chemistry codes such as (VASP or QuantumEspresso) either before, during or after the data collection. Several groups are developing infrastructure that would simplify the code interfaces to enable novice users to get started, enabling them to use the codes in progressively more sophisticated ways. This infrastructure ultimately will enable the direct comparison of data and theory. The computing requirements (and sometimes installation requirements) suggest that many of the codes will not be run on local resources at scale, but will be run at large HPC centers or on cloud/grid resources.

### 12.6.2 Software Infrastructure

A number of interesting software infrastructure projects are taking place at ALS, APS and NSLS-II as well as the local SLAC infrastructure developed for LCLS. For selected techniques, SLAC staff are exploring data collaborations with other labs. Additionally, there are a number of infrastructure tools that can be developed in common, such as extensions to HDF5. It should be stressed these collaborations are likely to be small scale and of mutual benefit. We expect that the user community conducting experiments across the US light sources may begin to request some commonality of tools.

### 12.6.3 Process of Science

See Section 12.6.1 for a discussion of the changes in the scientific process motivated by a tighter coupling of theory and experiment. There are certainly some networking tool implications. The users will have to be able to authenticate across different systems in order to access the experimental and calculated data. Individual facilities may deploy cataloging systems that will need federating.

## 12.7 Network and Data Architecture

SLAC has two connections to ESnet to provide redundancy. The primary ESnet connection is 100Gbps, and the backup connection is 10Gbps. SLAC also has connectivity to Internet2 via Stanford University (this connection is through the Stanford Research Computing Center which is located on the SLAC campus).

SLAC does not have a specific Science DMZ, but the LCLS provides data transfer nodes for LCLS and SSRL experiments for high-speed data transfers.

## 12.8 Data, Workflow, Middleware Tools and Services

To make a distinction, we are currently in a revolutionary era of personal IT as exemplified by the iPhone, and some of the processes and societal changes that have resulted from that revolution have not manifested in lab or research computing environments. Lab IT services are typically costly, based as they are on enterprise level needs for robustness—this limits the types of applications that can be supported or developed. There could also be an assessment of cyber security as it applies to various states of research data. However, taking a modern look could lead to remote controlling a beam line with an app that then returns plots immediately to the mobile device. Of course, the basic infrastructure has to built, however we should be considering adding value and usability to that infrastructure.

## 12.9 Outstanding Issues

To make a distinction, we are currently in a revolutionary era of personal IT as exemplified by the iPhone, and some of the processes and societal changes that have resulted from that revolution have not manifested in lab or research computing environments. Lab IT services are typically costly, as they are based on enterprise-level needs for robustness—this limits the types of applications that can be supported or developed. There could also be an assessment of cybersecurity as it applies to various states of research data. However, taking a modern look could lead to remote controlling a beam line with an app that then returns plots immediately to the mobile device. Of course, the basic infrastructure has to built, but we should consider adding value and usability to that infrastructure.

*Table 12.1: The following table summarizes data needs and networking requirements for the SSRL.*

| Key Science Drivers | | | Anticipated Network Needs | |
|---|---|---|---|---|
| **Science Instruments, Software, and Facilities** | **Process of Science** | **Data Set Size** | **LAN Transfer Time needed** | **WAN Transfer Time needed** |
| **Near Term (0-2 years)** | | | | |
| · SSRL has 33? beamlines and delivers beam with an uptime of 95% during the 9 month running period.<br>· Custom tools at the beamline for data acquisition, fast feedback and user analysis. | · Proposal driven.<br>· User driven with dependencies on techniques.<br>· Data is stored at the local beamlines for two weeks then deleted. SLAC based experimenters will move the data to local resources, non-SLAC users will copy the data, often to portable media and sometimes email. | · File sizes vary from KB to a few GB.<br>· Data sets have similar variation. | · Most data remains at the beamline. Local transfer needs are small and covered by existing solutions. | · Files may need to be transferred within two weeks of data collection. |
| **2-5 years** | | | | |
| · Upgraded detectors | · Expect to see increased use of simulation codes and integration of theory and experiment<br>· Supporting automated high throughput characterization and other<br>· Possible re-evaluation of DMP consistent with SC directives. | · Expect continued wide variation in the data sizes and data rates.<br>· If needed, can deploy the LCLS infrastructure to handle large files or data sets. | · Expect more LAN traffic between SSRL and the Stanford Research Computing Center (located on the SLAC campus).<br>· Expected to be easily accommodated by planned 10gbps connection. | · Volumes remain small compared to LCLS.<br>· Collaborating institutions: national (NERSC), other conventional light sources in US.<br>· Theory groups<br>· Hubs |
| **5+ years** | | | | |
| · Detector technology is expected cause a dramatic increase in data volume. | · Expect extensive simulations with possible integration with experiments.<br>· Expect a need for sophisticated on-the-fly data mining.<br>· Will replace components of software/hardware systems as motivated by changing technologies, emerging standards or bottlenecks.<br>· Expectation of some robust shared infrastructure for data management and analysis (technique) dependent. | · Maximum size is being determined and will depend on detector technology. | · Unknown but expected to be small relative to LCLS. | · Collaborating sites: US and international lightsource community. NERSC, International, University Resources (derived data), XSEDE resources. |