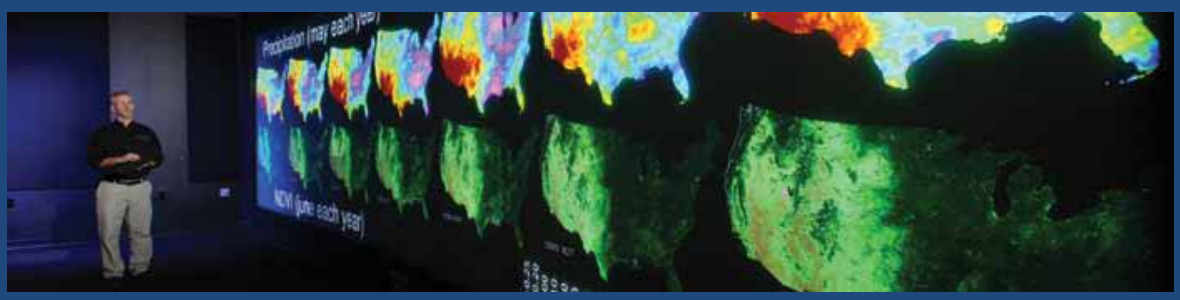
Data and Communications in Basic Energy Sciences: Creating a Pathway for Scientific Discovery

Report of a Workshop Linking Experimental User Facility Needs
with Advances in Data Analysis and Communications
June 2012



Report of the Department of Energy Workshop on:

Data and Communications in Basic Energy Sciences: Creating a Pathway for Scientific Discovery

October 24-25, 2011

Chairs:

Peter E. Nugent, Lawrence Berkeley National Laboratory
J. Michael Simonson, Oak Ridge National Laboratory

Panel Leads:

Workflow Management
Mark Hagen, Oak Ridge National Laboratory
Ross Miller, Oak Ridge National Laboratory

Theory and Algorithms
Bobby Sumpter, Oak Ridge National Laboratory
Chao Yang, Lawrence Berkeley National Laboratory

Visualization and Analysis
Dean Miller, Argonne National Laboratory
James Belak, Lawrence Livermore National Laboratory

Data Processing and Management
Amber Boehnlein, SLAC National Accelerator Laboratory
Eli Dart, Lawrence Berkeley National Laboratory

Front cover: Experimental and computational user facilities supported by the Office of Science, U.S. Department of Energy

Table of Contents

Executive Summary

The Department of Energy (DOE) Workshop on "Data and Communications in Basic Energy Sciences: Creating a Pathway for Scientific Discovery" was held at the Bethesda Marriott in Maryland on October 24-25, 2011. The workshop brought together leading researchers from the Basic Energy Sciences (BES) facilities and Advanced Scientific Computing Research (ASCR). The workshop was co-sponsored by these two Offices to identify opportunities and needs for data analysis, ownership, storage, mining, provenance and data transfer at light sources, neutron sources, microscopy centers and other facilities.

Their charge was to identify current and anticipated issues in the acquisition, analysis, communication and storage of experimental data that could impact the progress of scientific discovery, ascertain what knowledge, methods and tools are needed to mitigate present and projected shortcomings and to create the foundation for information exchanges and collaboration between ASCR and BES supported researchers and facilities.

The workshop was organized in the context of the impending data tsunami that will be produced by DOE's BES facilities. Current facilities, like SLAC National Accelerator Laboratory's Linac Coherent Light Source, can produce up to 18 terabytes (TB) per day, while upgraded detectors at Lawrence Berkeley National Laboratory's Advanced Light Source will generate ~10TB per hour. The expectation is that these rates will increase by over an order of magnitude in the coming decade. The urgency to develop new strategies and methods in order to stay ahead of this deluge and extract the most science from these facilities was recognized by all. The four focus areas addressed in this workshop were:

- *Workflow Management - Experiment to Science:* Identifying and managing the data path from experiment to publication.

- *Theory and Algorithms:* Recognizing the need for new tools for computation at scale, supporting large data sets and realistic theoretical models.

- *Visualization and Analysis:* Supporting near-real-time feedback for experiment optimization and new ways to extract and communicate critical information from large data sets.

- *Data Processing and Management:* Outlining needs in computational and communication approaches and infrastructure needed to handle unprecedented data volume and information content.

It should be noted that almost all participants recognized that there were unlikely to be any turn-key solutions available due to the unique, diverse nature of the BES

community, where research at adjacent beamlines at a given light source facility often span everything from biology to materials science to chemistry using scattering, imaging and/or spectroscopy. However, it was also noted that advances supported by other programs in data research, methodologies, and tool development could be implemented on reasonable time scales with modest effort. Adapting available standard file formats, robust workflows, and in-situ analysis tools for user facility needs could pay long-term dividends.

Workshop participants assessed current requirements as well as future challenges and made the following recommendations in order to achieve the ultimate goal of enabling transformative science in current and future BES facilities:

Theory and analysis components should be integrated seamlessly within experimental workflow.

Develop new algorithms for data analysis based on common data formats and toolsets.

Move analysis closer to experiment.

Move the analysis closer to the experiment to enable real-time (in-situ) streaming capabilities, live visualization of the experiment and an increase of the overall experimental efficiency.

Match data management access and capabilities with advancements in detectors and sources.

Remove bottlenecks, provide interoperability across different facilities/beamlines and apply forefront mathematical techniques to more efficiently extract science from the experiments.

This workshop report examines and reviews the status of several BES facilities and highlights the successes and shortcomings of the current data and communication pathways for scientific discovery. It then ascertains what methods and tools are needed to mitigate present and projected data bottlenecks to science over the next 10 years. The goal of this report is to create the foundation for information exchanges and collaborations among ASCR and BES supported researchers, the BES scientific user facilities, and ASCR computing and networking facilities.

To jumpstart these activities, there was a strong desire to see a joint effort between ASCR and BES along the lines of the highly successful Scientific Discovery through Advanced Computing (SciDAC) program in which integrated teams of engineers, scientists and computer scientists were engaged to tackle a complete end-to-end workflow solution at one or more beamlines, to ascertain what challenges will need to be addressed in order to handle future increases in data volume as well as replicating such an effort across the DOE BES facilities.

Data and Communication Needs in BES User Facilities

Basic Energy Sciences (BES) supports the design, construction and operation of major national facilities for user-driven research. These include major capabilities in advanced synchrotron and free electron laser light sources, electron beam microcharacterization centers, steady-state and pulsed sources and instruments for neutron scattering, and five unique nanoscale science research centers engaged in the synthesis and characterization of novel nanoscale materials and systems. Together these facilities represent a significant national resource for fundamental physical sciences, engineering and biosciences research, and contribute directly to applied research and development supporting new technologies. These facilities, including the intellectual contributions of staff scientists, are made available to the broader scientific and technical community through user programs, allocating resources to users based on peer review of research proposals. The user community has grown over time to over 14,000 individual users each year, many of whom carry out multiple experiments each year. Historically the use of these facilities has engaged individuals or small teams of researchers carrying out individual experiments on specific beamlines at each facility. Rates of data acquisition from individual experiments were relatively modest as compared with fields such as high-energy physics or astronomy. The data management and analysis processes used to extract scientific insight from experiments were often addressed on individual beamlines, or in some cases developed for individual experiments. More recently, the landscape of data needs in BES user facilities has changed markedly. New photon and neutron sources are providing much higher intensities to individual beamlines; coupled with corresponding advances in detector technology, this is resulting in unprecedented rates of data collection in the experiments. The sheer volume of data from individual experiments has also increased as scientists ask increasingly complex questions on deeper and more subtle properties and phenomena. It is increasingly apparent that the combination of results from a series of experiments using different techniques and facilities, and often incorporating theoretical guidance or the results of numerical simulations of the system, pays high dividends in scientific understanding of complex phenomena. All of these factors have combined to create a scientific environment in which the pace of discovery may not be limited by experimental constraints, but by the ability of research groups to manage, analyze, and ultimately understand the data resulting from the experiments.

As these experimental advances have gone forward, the pace of computational science has been expanding essentially with Moore's Law. Challenges of handling and extracting information from data generated by new computational models have been addressed and solutions developed, in part through research sponsored by the DOE-SC Advanced Scientific Computing Research (ASCR) program. Information grid

resources have expanded to enable fast transmission of data, in keeping with the computing power available within individual facilities. Here again, user-driven research in ASCR-supported facilities has been a strong driver for upgraded communication and data-handling capabilities.

The advances in experimental and data-intensive research have not always been closely linked. Past workshops linking these areas have pointed out specific opportunities (e.g., "Computational Scattering Science 2010," National Science Foundation; "Mathematics for Analysis of Petascale Data," DOE 2008; "Scientific Collaborations for Extreme-Scale Science," DOE 2012), but have not directly addressed the application of new data techniques in a range of experimental projects, or effectively outlined areas for focus that will enable the more effective, efficient use of experimental facilities for user-driven science. The drivers for this workshop, and recommendations from this report, are centered on enhanced communications between researchers in experimental and computational sciences. Each of the focus areas within this workshop presents opportunities for future discussion, applications of lessons learned and currently-available approaches, and possibilities for additional research targeted specifically to the data needs of experimental user facilities.

Current BES Facilities

For light sources, recent improvements in detector speed and light source brightness are yielding unprecedented data rates that exceed the capabilities of most data management and data analysis approaches utilized in past experiments. Unlocking the full potential of these facilities for scientific discovery and technological advances requires state-of-the art networking and computing facilities and a new generation of analytical methodologies and tools. Of the four fundamental parameters are used to describe the physical world (energy, momentum, position, and time), three correspond to the three broad categories of synchrotron experimental measurement techniques: spectroscopy (energy), scattering (momentum), and imaging (position). The fourth parameter—time—can in principle be applied to all the techniques. Experiments that directly measure specific sets of parameters (e.g., momentum and energy) can be correlated with complementary measurements in position and time.

Synchrotron light sources are making advances that will put them squarely in the data-intensive category, including new sources (National Synchrotron Light Source II (NSLS-II) under construction at Brookhaven National Laboratory (BNL)) and upgrades to sources and instruments at the Advanced Photon Source (APS) (Argonne National Laboratory (ANL)) and Advanced Light Source (ALS) (Lawrence Berkeley National Laboratory (LBNL)) facilities. Even where individual beamlines do not push the limits of relatively sophisticated data-handling systems, the aggregate data volume and rate from the facility can pose challenges in data

communication, reduction and subsequent interpretation to impactful science. An example of the total data produced at the ALS over time is illustrated in Figure 1. From the current base load, anticipated improvements in detector technology (speed and efficiency) and the implementation of new experimental techniques (e.g., pulse slicing for enhanced time resolution) will significantly increase data rates and volumes from existing light-source facilities.

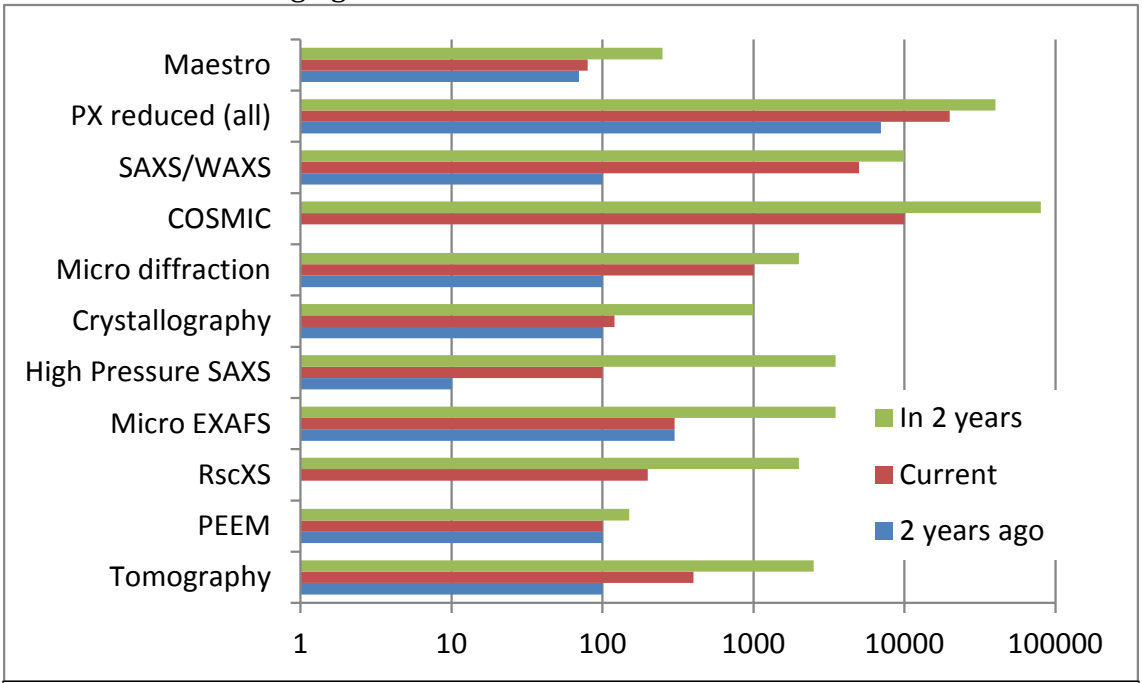


Figure 1: Results from a recent user survey at ALS that summarize the average data volume in GB/month for a variety of beamlines collected during a single month. The light blue bar highlights the average data volume two years ago, red the current volume and yellow the projected volume due to beamline and detector updates in two years. Currently, the 18 beamlines in the survey generate about 300 TB of data per year. This is expected to rise to almost 2 petabytes.

The Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory currently represents the outside of the envelope in data intensity among current BES user facilities. The information intensity and relatively low number of individual experiments at this facility have tended to lead to the formation of interdisciplinary proposal teams that are somewhat larger than at ring-based synchrotron light sources or other facilities, further driving the need for advance data transfer and analysis across the team. For the LCLS, high-performance data systems have been a recognized need throughout the planning and construction process, and new data-streaming and parallel-file capabilities are being developed to address these data needs.

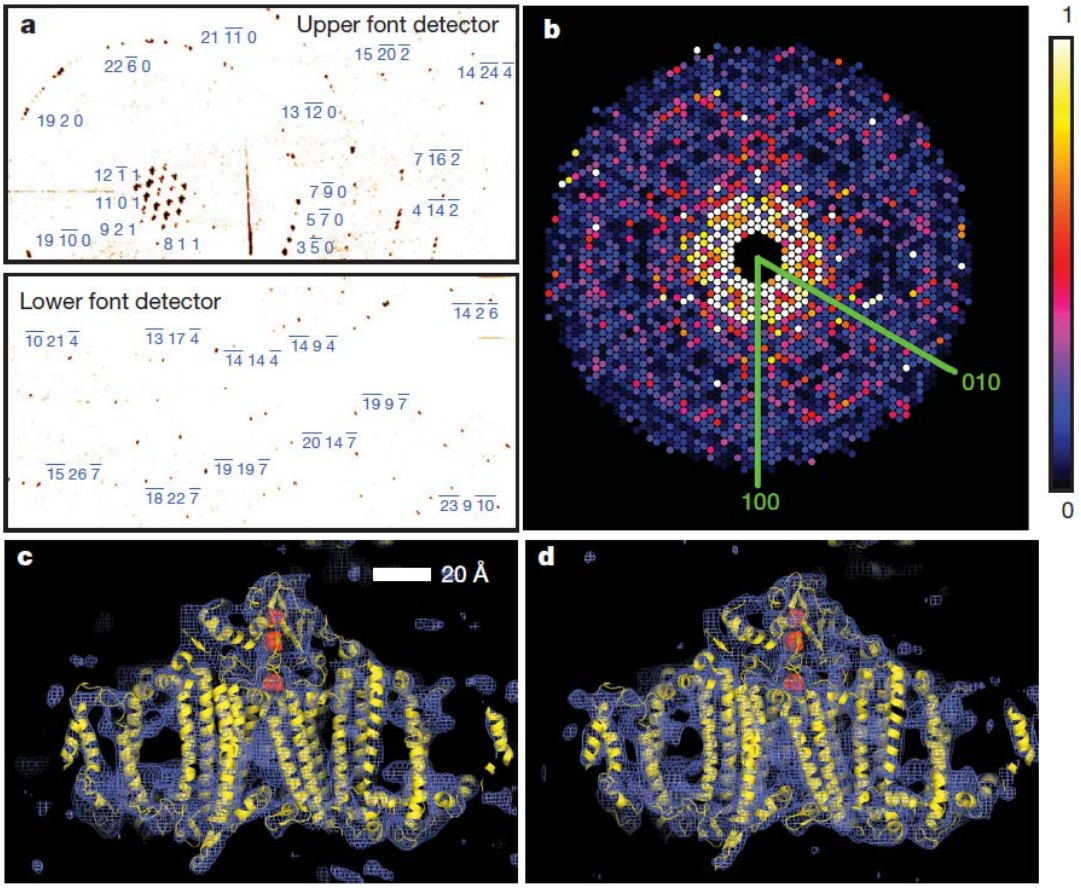
DOE BES user facilities for neutron scattering present similar data challenges. User facilities for pulsed-neutron experiments include the Lujan Center at Los Alamos National Laboratory (LANL) and the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory (ORNL). The High Flux Isotope Reactor (HFIR) at ORNL is a

steady-state reactor source. Data-handling and analysis strategies are continuing to be refined to take into account multiple types of experimental results on a given sample or system. For HFIR and the Lujan Center, a local instrument-specific model for data acquisition and analysis has proven over time to meet the needs of the user community. As a pulsed source with time-averaged flux on sample comparable to steady-state sources, SNS is presenting challenges in data acquisition and analysis similar to those in light source experiments. Using the pulse structure to capture dynamic system response across a range of time scales requires recording of time-resolved individual scattering events ("event mode") coupled with metadata on experimental conditions at the time of the event ("slow controls"), leading to large data sets. A recent estimate predicts that at full operation at > 1 MW power levels, the instrument suite at the SNS will produce data at a peak rate of ~4 GB/s, with sustained data rates near ~90 TB/day.

Electron-Beam Microcharacterization Centers (EBMCs) and Nanoscale Science Research Centers (NSRCs) present different data-intensive issues. The electron-beam centers (and some x-ray and neutron scattering beamlines) were early pioneers of "remote access" technologies, enabling remote users to observe and in some cases control experiments in real-time. Visualization, image translation and data-transfer bandwidth were identified early as issues in the remote-access process. While significant advances have been made, opportunities remain to enhance this timely and cost-effective access mode for state-of-the-art electron microscopy and microcharacterization. The five NSRCs (Molecular Foundry, LBNL; Center for Nanoscale Materials, ANL; Center for Functional Nanomaterials, BNL; Center for Nanophase Materials Sciences, ORNL; and the Center for Integrated Nanotechnologies, Sandia National Laboratories and LANL) have unique missions of nanomaterial synthesis in their programs, along with advanced characterization techniques. Each of the NSRCs incorporates active theory, modeling and simulation user research with the experimental programs, with several of the centers interacting extensively with local and/or remote DOE-supported computational facilities for access to additional modeling expertise and computational resources. Data and computational needs for these facilities include high-speed communication for effective use of remote computers, visualization of results for comparisons between models and experiments, and transfer of both experimental and simulation results between users and facilities.

Example: Unique, data-intensive science at current facilities

A limiting factor in understanding the structures of membrane proteins is the difficulty of growing large well-diffracting crystals—fewer than 300 unique structures have been solved for this reason. A demonstrated scientific advance using a coherent beam at LCLS is the ability to take 'diffraction' snapshots of nanocrystals. Photosystem I is a biological factory in plant cells that converts sunlight to energy during photosynthesis. In an experiment at LCLS, millions of nanocrystals containing copies of Photosystem I were exposed to the X-ray beam with laser pulses striking the nanocrystals at various angles and scattered into the detector, forming the patterns needed to reconstruct the images. Of the three million diffraction patterns, 10,000 were used to generate the information needed to solve the known molecular structure of Photosystem I. The structure of Photosystem I has been extensively studied at conventional synchrotron radiation facilities in crystalline form; the comparison between conventional techniques (right) and LCLS results (left) are seen below—the structure of Photosystem I is solved in both cases.



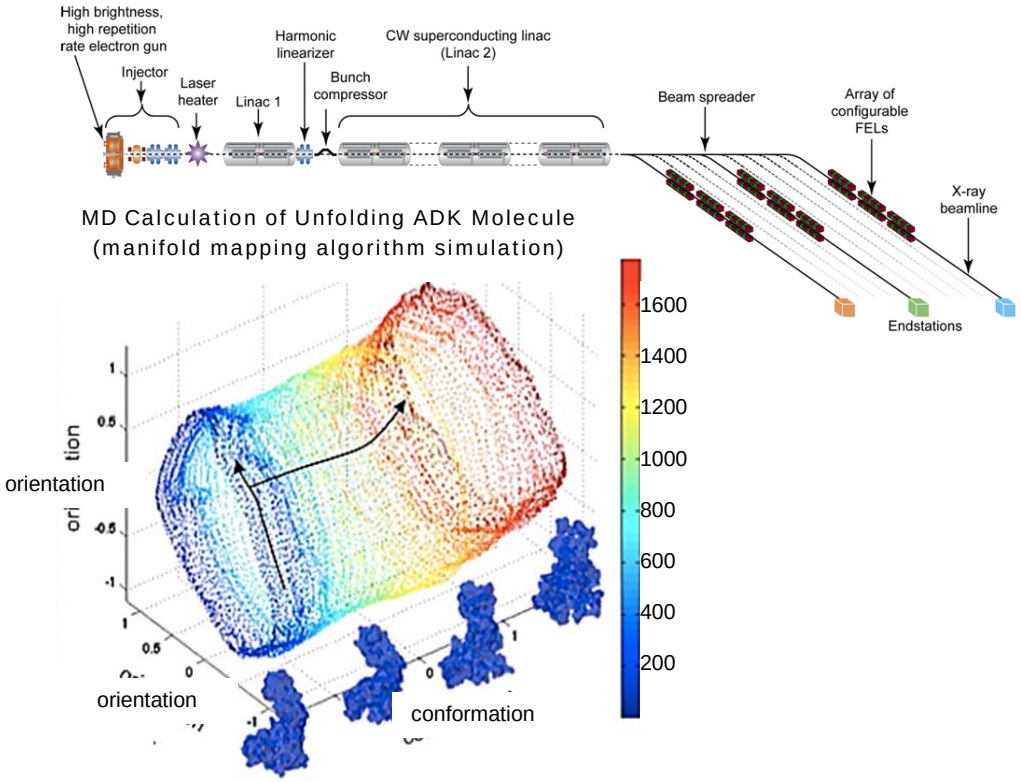
Electron density map of Photosystem I obtained from the LCLS (left) and a reference experiment using synchrotron radiation (right). From the LCLS data, out of 1,850,000 recorded patterns 112,000 were identified as hits and 15,000 could be indexed. Image taken from H.N. Chapman et al., Nature 470, 73 (2011)

Future BES User Facilities

The success of the user community in producing high-impact science using existing facilities is spurring expansion, modernization, and construction of next-generation facilities. In order to fulfill their scientific promise, these next-generation facilities will produce dramatic increases in data rates over the next decade. As LCLS is pushing the envelope for data processing from experimental results, LCLS-II (being designed and constructed) will have even greater challenges in data systems since multiple experiments will be conducted simultaneously. Currently, the NSLS-II at BNL is in the final stages of construction. Operation of the facility is expected to start in 2014 with a limited number of beamlines, and then progressively adding new instrument towards full productivity within approximately 5 years. With about 22 out of a total of 58 beamlines operating at modest levels of productivity, the NSLS-II may produce an aggregate of ~100 TB/day or ~600 TB/week. A full complement of beamlines operating at optimal performance would increase the data rates by a factor of five, producing in excess of 2 petabyte per week of experimental data. Rapid data acquisition enabling high-throughput science will increase the user demand for remote-access experimental interaction and control, further increasing the need for effective data systems. The current community operations model, based on user travel to central facilities, informal collaboration of independent research groups and performing data analysis in quasi-isolation using dispersed resources will break down, if only due to the sheer number of experiments and size of the datasets. Science capabilities at existing light and neutron sources are being enhanced through continuing upgrades of experimental systems. State-of-the-art high-speed detectors, such as single photon counting detectors now in use at the ALS, permit millisecond small angle x-ray scattering and micro-x-ray diffraction measurements but generate so much data that the current analysis methodology simply cannot scale. Acquisition of large volumes of increasingly complex data will inevitably result in a mismatch of data analysis with scientific goals and degrade the efficiency in design of experiments. One free-electron laser facility currently in planning will utilize ultra-fast detectors capable of generating more than 100 megabytes of data per second per beamline and thereby magnify this problem by an order of magnitude and more. With such high data volumes, matching science productivity to experimental capability will depend on the development and application of new approaches to enable faster data analysis turnaround times. Efforts in computing paradigms for multi-core architectures, such as those of GPUs and hybrid CPUs/GPUs, are expected to lead to critical advances for overcoming these computational barriers. The effective use of these radically improved user facilities in advancing scientific discovery will be related directly to the corresponding advance in techniques and capabilities for data acquisition, communication, analysis, and management. The initial discussions in this workshop have shown that there are important advantages for current and projected user facilities in adopting and building on data capabilities developed with the support of the DOE-SC ASCR program.

Example: Data Needs for Future Facilities

Future facilities such as free-electron lasers will enable unique science, but require support of very large data sets and acquisition rates to carry out their science missions. The figure below shows an example beamline schematic for an experiment on the structure of a large molecule - the orientation reconstruction of an ADK (Adenylate kinase) molecule. The input data rate is 10^5 images/second at 10^6 pixels imaging rate (4TB/sec). One has 10^5 images of diffraction patterns representing 2D projections of the sample in random orientations. The best available orientation algorithms require $\sim N^6$ flop (note N=1000 for an NGLS detector). The total performance required is 10^{18} flop/second for a pulse rate of 10^5 images/second - a sustained exaflop computing capability *just* for orientation reconstruction.



Summary Results from Workshop Discussions

The scope of issues involved in data needs for user facilities warranted breaking the problem into more tractable areas for detailed discussion within the workshop. A number of breakdowns could have been identified, and those chosen were neither unique nor self-contained. However, those chosen do represent a cross section of needs within the process of taking data from experiments, simulations, or combinations and processing those data to achieve the end result – impactful science from the BES user facilities. The selected topics were Theory and Algorithms, Visualization and Analysis, Data Processing and Management, and Workflow Management. The following sections are brief synopses of the discussions. Each should help to form a springboard for further, in-depth consideration of these areas for specific research, and should be considered together for the guidance they offer on paths forward to address data needs.

Theory and Algorithms

The problems addressed by current and future BES user facilities are extremely diverse. The science questions range from biology, material science, physics to earth science and archeology. While the science challenges are diverse, the data processing and analysis techniques are connected by common themes and underlying theory and algorithms. To place the challenges in context, the data analysis from X-ray light source instruments can be broken down into three general areas: Real Space, Reciprocal Space, and Spectroscopy.

Real Space. X-ray imaging beamlines have been generating terabytes (or more) of raw image data per month. With the development of new and faster cameras, data rates will double or triple in the near future. As the data rate and volume have increased, the need for on-site analysis has increased. The imaging data analysis involves reconstruction of large 3D images, segmentation of the images into sub-regions, discrimination of multiphase solid materials, identification of microstructures, calculation of statistical correlation functions (e.g., surface, pore-size) and extraction of channel networks. The amount of data demands both automation and new ways of thinking about the underlying calculations. Analysis algorithms should be re-factorized into software that can use multiple cores across several nodes. In general, this approach is characterized as high performance analytics (data centric). Once the need for revised algorithms has been met, issues of disk I/O, efficient threading and distributed communications will play a major role in providing the necessary tools.

Reciprocal Space. Just as in imaging beamlines, scattering instruments have seen a revolution with respect to flux and detectors. High brilliance beams, efficient x-ray focusing optics and fast 2D area detector technology allow the collection of thousands of diffraction patterns occupying terabytes of disk space. The need for real time analysis also stems from the need of being able to modify an experiment on the fly in cases the next action depends on the outcome of the previous scan. Additional computational needs for the light sources is the integration of modeling and simulation as tools accessible to the users in real time - to get the most bang for the buck while at the facility. Techniques such as coherent diffractive imaging, nanocrystallography and ptychography under development at light sources are the fruit of advances in reconstruction techniques. In these techniques, the need for algorithms capable of solving large-scale ill-conditioned, underdetermined, noisy inverse problems has never been so clear. Our ability to take advantage of reconstruction techniques in X-ray microscopy and imaging will depend on our ability to extract maximum information from diffraction data.

Spectroscopy. The data rates for spectroscopy are traditionally not as high as for the previous described techniques; however, like the others, its analysis relies very heavily on quite complicated analysis algorithms and simulations.

Additional progress can be made across the field by working on algorithms that allow the combination of multiple data sources and imaging techniques to provide more reliable solutions. Being able to couple both real and virtual (simulated) experiments and/or having *ab initio* theory guide experiments for data triage/reduction should also be highly beneficial.

The urgent need to address the data challenge at BES facilities cannot be easily met without a deeper understanding of the nature of the measurements. Once the theoretical relationship between the measured data and the object to be uncovered or the problem to be solved can be clearly described in a quantitative manner, a suitable mathematical model can be developed that allows the data analysis problem to be solved in a systematic and efficient fashion. Only when such a model or problem formulation is available, can reliable computational algorithms and their implementations begin to be developed. This approach will require a close collaboration between theory and experimental scientists at BES facilities, applied mathematicians and computer scientists supported by ASCR. The members of the Theory and Algorithms panel identified the following set of issues and problems at the existing and future BES facilities. A number of suggestions were made on ways to promote a close collaboration between BES and ASCR scientists in theory and algorithm development.

Hypothesis-driven science, our traditional approach, is now becoming strongly dependent on a new paradigm of data-driven discoveries. While computational algorithms have largely maintained pace with Moore's law, precise data analysis has not. This represents an impending future problem. Additionally, the current state-

of-the-art in computing means that computational science also provides a viable instrument of its own, capable of very large data production. An example of how the time from data generation/collection to analysis can be reduced for neutron reflectivity is shown in Figure 2. In this particular case, by using standard off-the shelf algorithms (numerical optimization such as the Levenberg-Marquardt algorithm) along with the underlying theory of reflectivity, a more robust fit (model-free or also known as free-form fitting) could be achieved on a shorter time than the case of script based parameter modification based on intuition and information from other characterization experiments (left side of figure). A further and more substantial step is to incorporate explicit physics based coarse-grained modeling tools such as self-consistent field theory (SCFT) for polymers to enable prediction and a feedback loop to the fitting tasks of the reflectivity (model-based fitting - right side of figure). By merging the SCFT with the theory of neutron reflectivity, a more reliable solution for the inhomogeneous structure of the sample can be obtained. Experimentally observed neutron reflectivity profiles can be reproduced by using numerical optimization and experimentally relevant parameters in the coarse-grained modeling.

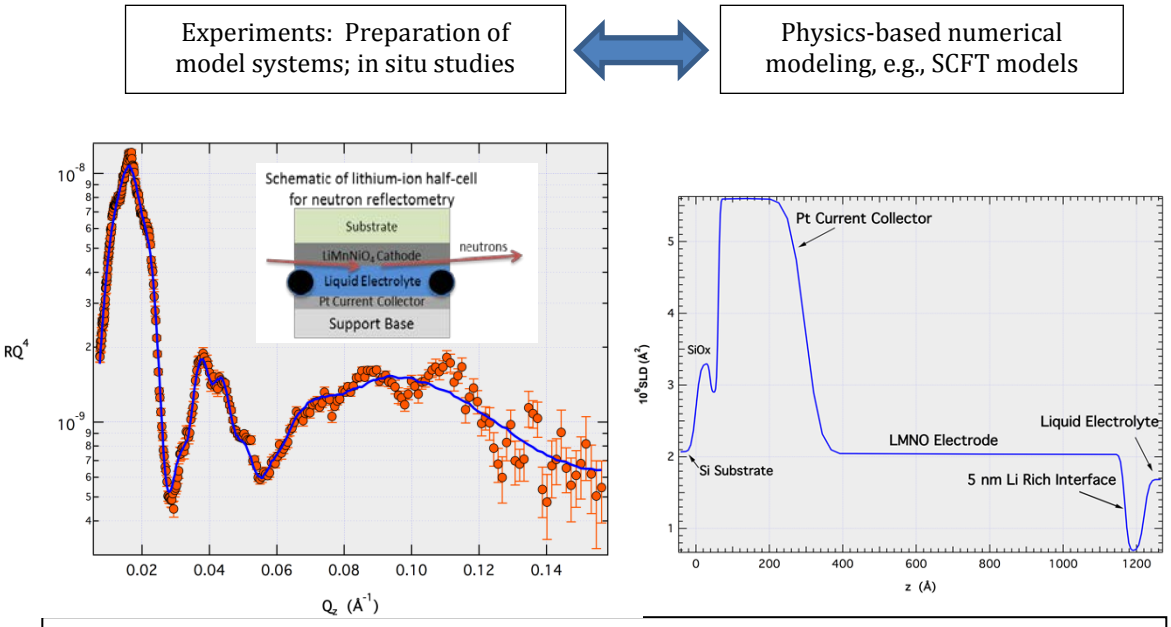


Figure 2. Complementary relationships between reflectivity measurements and computational experiments.

Currently, data analysis at BES facilities is typically done by a small group of facility scientists and users. There is no consensus on how to best model the analysis problem or what algorithms to use. Typical approaches use a combination of commercial software packages that are not designed to solve the particular data problem at facilities and some type of "home brewed" codes/scripts that are neither robust nor flexible. On the other hand, ASCR has provided some support in large-scale data analysis. However, most projects typically develop general

methodologies that are not tailored to any specific data problem found at BES facilities.

It was recognized by the members of the panel that the most productive way to meet the theory and algorithms challenges at the facilities is to organize coherent (perhaps similar to the "glue" Computational Materials and Chemical Sciences Networks (CMCSN) concept in BES) interdisciplinary teams consisting of researchers in theory and experiment, applied mathematicians (including statisticians) and computer scientists to examine both the problem formulation and solution strategies simultaneously. Ideally, it would be optimal to have one team for each end station at each facility. However, before a clear path is identified, it is prudent for BES and ASCR to support some pilot studies that are targeted at specific scientific problems with strategic alignment for producing major breakthroughs. All panel members agreed that it is imperative to provide long term and sustainable support for such type of activities.

Several areas were discussed in the context of the core role of theory in understanding experimental results. In the general area of inverse problems and solution algorithms, phase retrieval for non-crystalline diffraction, tomography, and single-molecule diffractive imaging were identified as particular areas of interest for collaborations between experimental teams and applied mathematicians. Similarly, feature extraction and image analysis including model-based constraints can provide an important first step toward scientific discovery. This capability is currently limited in the number of appropriate tools available at BES user facilities.

Combining multiple data sources and imaging techniques will provide more reliable solutions to complex data analysis problems. The nonlinear and non-convex nature of many data analysis problems arising from the facilities suggests that it may be difficult to rely on a single data source or imaging technique to provide a complete solution to inverse problems. It may be beneficial to combine X-ray diffractive imaging with, for example, electron microscopy to provide a more reliable solution. Multiple imaging techniques and data sources may also be combined to deliver a multi-resolution imaging capability at many facilities.

Following the concept of model-based constraints in data analysis, *ab initio* theory could be used more broadly to guide experiments needed to optimize data reduction. Panel members pointed out that at least for some experiments (e.g., spectroscopy), data volume might be reduced if *ab initio* theory can be used to guide how the experiments are carried out.

Many of these concepts could come together in the context of computational end-stations that couple virtual and real experiments. For a facility user, performing a real experiment is often a daunting task that involves careful planning and a significant amount of effort in tuning parameters and adjusting instruments. The process itself is often tedious and time-consuming. Productivity and scientific

throughput can be increased immensely if computational end-stations can be developed to allow virtual experiments to be carried out prior to real experiments.

Visualization and Analysis

One common theme across many scientific programs, not just those at user facilities, is that datasets from both experiments and simulations are becoming larger, more complex, and are being generated at a faster rate than ever before. This rapid growth in data rate results in a new set of challenges and opportunities.

A central theme of discussion within this workshop session was the role of visualization as a step (perhaps a first step) of an overall data analysis workflow. For some time ASCR has supported an active program in visualization research and development as a key to understanding the results of complex simulations, yielding massive data sets with complex interrelationships. Some of the techniques developed in this simulation-driven visualization research have direct application to displaying and understanding experimental results (e.g., the 4-D spectroscopic data referenced above). Particularly for relatively sparse or incomplete data sets, such as those from the beginning of an experimental run cycle, close coupling of visualization with available modeling results could give early insight into the progress of an experiment, or offer guidance on particular areas of interest within the experimental phase space.

For the experimentalist, the traditional way of doing research has involved collecting and storing of data, then performing analysis once all data has been collected as a post-processing step. In this post-processing mode, the duty cycle of collect-store-analyze data becomes increasingly long as data sizes and rates grow. Many BES science teams have expressed the desire to be able to more rapidly process experimental data as it is collected, and for multiple reasons. One is to alter the experiment while it is running and/or to perform instrument calibration, which may not be possible in a strictly post-processing mode with a long duty cycle. Another is to perform data reduction operations so that the final data stored for later use is much smaller in size than the raw data coming from experiment. In this setting, the term *in situ* processing means being able to accommodate data processing/analysis/visualization at the rate it is being generated by the experiment, and likely "close to" the experiment source, rather than moving full-resolution data over the network to a remote facility for processing.

Another theme heard throughout BES was the incredible benefit that could be derived from merging visualization with modeling/simulation to enhance the analysis of experimental data and to better design future experiments. Here one would have the ability to use the simulation to map out where within the experimental space you would want to do the experiment as well as during the experiment to confront theory with observation directly.

Key to the success of such an effort is the ability to leverage ASCR funded advances in visualization for simulations to visualization for experimental data that will be able to handle real space, reciprocal space (k-space, dual-space), and spectroscopic analysis. Future analysis software should be open source, platform independent (from the laptop to high performance parallel computers), able to have both a simple GUI interface as well as an expert mode and take advantage of emerging computing hardware (such as graphical processor units (GPUs), multi-core, ARM (advanced reduced instruction set computer microprocessor). Of course such an effort will require standardized data formats so that one package can be used across beamlines and facilities and will likely involve considerable research into high-performance/parallel I/O.

Data Processing and Management

Today, users accessing BES user facilities are routinely able to generate 10,000's to 100,000's of images (in real or reciprocal space) in a few days of run time. Such data volumes cannot be analyzed individually, but rather must rely upon automated methods that translate, for example, materials science descriptions to input for modeling and simulation, and that quantitatively compare output of simulation with beamline data. To maximize both the functionality and robustness of these kinds of end-to-end analysis systems, discussions in this section centered on a community-wide, open-source effort to meet the needs for data acquisition, communication and processing for user facilities of the future. Similar large scale, automated and customized systems have been developed and deployed for other science communities. Although none are directly adoptable by BES user facilities and scientists, many principles, approaches, and lessons learned are directly applicable. Discussions focused on the characteristics these tools must possess in order for them to be widely adopted by the BES community.

Ease-of-use and extensibility. These features both ensure widespread adoption within a scientific community and maximize the reusability of software components developed by research teams at different beamlines and/or facilities. As an example, a graphical modeling interface similar to those used by solids modeling programs would provide researchers a natural method of describing the microscopic structure of material samples, and a common format for input to simulations.

Deployment of advanced algorithms using state-of-the-art computer hardware. In order to develop the fastest algorithms and robust codes, we need to exploit and leverage the resources provided by the ASCR office, including the expertise in Applied Math, Computer Science and the High Performance Facilities, such as the National Energy Research Scientific Computer Center (NERSC) and the Leadership Computing Facilities (LCFs). In particular, parallelization of these algorithms on multiple central processing units (CPUs), GPUs, and hybrid CPU/GPU multicore

architectures will dramatically decrease the analysis time by more than several orders of magnitude while simultaneously permitting larger data sets to be treated.

Leveraging of advanced computer technologies. As enabling computer technologies (such as data I/O and formats, ontologies, FFT libraries, etc.) and architectures (such as GPUs, multi-core, or heterogeneous architectures) evolve and improve, a common framework must allow for graceful evolution to accommodate and take advantage of the latest improvements while insulating users from the underlying details. This provides two advantages: The perturbative effects of such changes to scientists' research are minimized and the advantages are more quickly and widely available.

Standardized Data Formats. Given the diversity of science and facilities, having standardized data formats across user facilities will enable fast access and easy handling of very large and/or complex datasets. It will also allow users to easily share and exchange data across a wide variety of computational platforms using applications written in different programming languages.

In addition to these points, one also needs to highlight the need for secure, robust archival and retrieval/transfer of data to users at their home institutions and other computing facilities. This will require investment in data transfer systems and operational expertise to support them by the BES facilities. This can be done in collaboration with the ASCR facilities, which have developed the necessary expertise. There are successful pilot projects of this kind today, but they will need to be generalized. Note that future data rates of 1-10 Petabyte/day will add increased capacity requirements to the Energy Services Network (ESnet) and to the site networks of the institutions that support the BES facilities An integrated strategic planning process is needed that considers both DOE-supported and local network capabilities to facilitate data transfer to user institutions.

The current use of a diverse set of data formats for individual experiments result in bottlenecks that span the entire chain of experiment to science. This prohibits a common set of analysis tools at all facilities, it forces the users to acquire much more “facility knowledge” than is necessary to accomplish their goals, and it hinders a tighter coupling between theorists and experimentalists. Furthermore the archiving of data and issues related to data provenance are difficult, if not impossible, to solve and several of the more arcane formats, designed for computing systems a decade or more ago, will not scale well as facilities generate orders of magnitude more data.

Workflow Management

Within the context of this workshop, Workflow Management was discussed in terms of the process of taking experimental data to scientific publication. As an overall process view, the discussions included many of the specific points brought out in the other discussion sections noted above. The 'concept to conclusion' aspect of this session brought out many of the general issues involved – responsibilities of experiment teams versus facility scientists; advantages and potential pitfalls of standardization; implementation of near-term fixes versus consideration of longer-term solutions. Given the holistic charge to the working group, there was also considerable discussion of responsibilities and prospective mechanisms for improvements. It was found useful to discuss the various aspects of workflow management in terms of a simplified diagram as sketched in Figure 3.

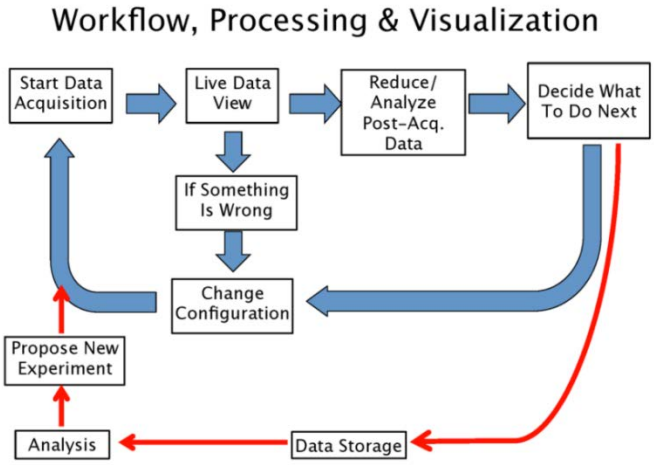


Figure 3. Simplified workflow diagram for experiment control, data acquisition and analysis, and feedback paths.

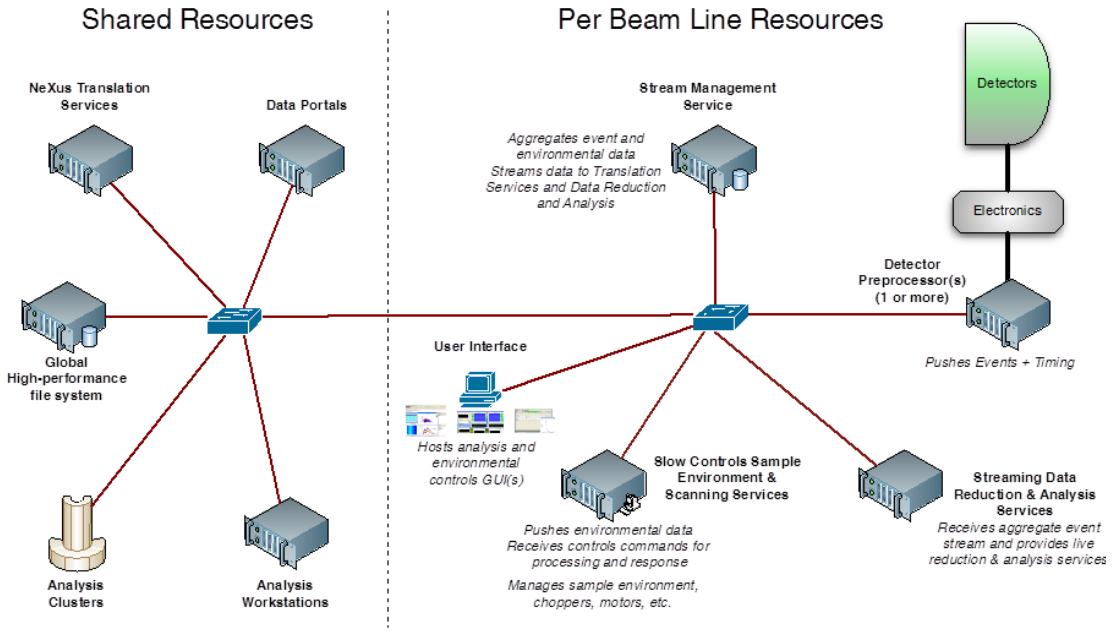
As a group of scientists begin an experiment, they need real-time analysis capability in order to ascertain the success of the experiment and to determine if it is time to move on to another run, or repeat the present run with changes to the experimental configuration. Such real-time (*in situ*) visualization and analysis requires cutting-edge algorithms to compress and analyze the appropriate data on meaningful time scales. Significant improvements in efficiency and overall scientific impact from user facilities are expected to be gained from this key addition to the workflow. It is anticipated that full analysis of the data acquired will continue to require post-experiment processing, either at the experimental facility itself or off-site at a data/computing facility. Improving the effectiveness of this step in the process requires the direct involvement of network researchers to enable data transfer, and the availability of appropriate computing capability and specialized analysis software. Several individual advances were identified as important to this overall process, including the use of standardized data formats (including associated metadata); further development of robust, open-source, community-based analysis software that can handle the large data volumes; and a tighter integration with theory to enhance the scientific value of individual experiments and facilitate direct comparison or combination with complementary results.

The consideration of these needs moved in many cases toward discussions of areas of responsibility and prospects for support of near- and longer-term development of

workflow enhancements. A number of BES user facilities have recognized these needs and are beginning efforts to address them through local collaborations of BES facility scientists and ASCR-supported researchers. An example illustrating the scope of the challenges and the range of expertise needed to address the issues is provided by an ongoing project to enhance workflow within the SNS.

Example: Streaming data flow for fast local analysis

The ADARA (Accelerating Data Acquisition, Reduction and Analysis) project at ORNL is a laboratory-supported joint effort between beamline and computational scientists at the SNS and the Oak Ridge Leadership Computing Facility (OLCF). The direct problem being addressed is that the 'batch' design of the SNS data system is not meeting the needs of the users. The solution being developed is based on a streaming data workflow rather than the current batch system. Using the widely-used NeXus data format and the Manipulation and Analysis Toolkit for Instrument Data (MANTID) data processing system, the core of the approach lies in streaming workflow from the instrument to nearby (~3 km) computational facilities through high-speed connection.



"Local" resources at the SNS manage the data flow and the incorporation of metadata for the experiment ("slow controls") along with the user interface. The resulting data stream is communicated to the computational facility for translation and file services, data analysis, and communication to external networks. As compared with the present system, ADARA is expected to give SNS users nearly "instant" access to data, and to provide the foundation for future live data analysis and steering/optimization of experiments. The overall scope of the project involves a laboratory investment of approximately $2M over two years, and builds on infrastructure and expertise developed in OLCF with ASCR support. Success in this effort will provide tremendously enhanced speed, access, and control within the SNS instrument suite and user interfaces.

Conclusions: Addressing Data Needs with Collaborative Efforts

Pathways to Success

Source and instrumentation advancements coupled with new concepts for experiments generated by the scientific community have led to unprecedented capabilities for transformative science using BES user facilities. The ability of the scientific community to reap the maximum benefit from experimental science, and to contribute maximum benefit toward solving societal problems, is linked to the effective, efficient use of these forefront facilities. Current systems are producing a tremendous amount of data, making it increasingly difficult to carry on "research as usual" – the involvement of small teams of users with facility scientists to carry out experiments producing data for future (remote) analysis. In the near future, experimental systems have the potential to overwhelm current analysis pipelines. Theory, modeling, and simulation sciences have become key components of planning and guiding the progress of experimental research, in addition to their historic roles in interpreting, understanding, and making predictions based on experimental results.

Faced with these challenges, some scientific communities have responded through the formation of large, multidisciplinary teams focused on concept-to-publication pursuit of individual experiments or program goals. With the wide range of disciplines and scientific problems impacted by BES user facilities it seems appropriate to consider how the needed advances in data acquisition, communication, management, and analysis can be made in a way to enable the continued effective access to these forefront capabilities, in a timely and effective way, by small teams of investigators studying a broad range of problems.

As experimental capabilities at BES user facilities have leapt forward, computational sciences capabilities supported in ASCR research programs and user facilities have made similar great advances. Leadership-class computational capability sustained at levels greater than a petaflop and made available to the user community through ASCR-supported centers have driven the development of advanced storage, manipulation, communication and visualization tools and techniques. A primary purpose of this workshop was to bring these two communities closer together, exchanging ideas, and opening an effective dialog toward the efficient, effective application of forefront data systems to new experimental instruments. In this light, this workshop and the present report serve mainly to highlight areas identified as particular needs in BES user facilities and particularly valuable capabilities and experience gained in ASCR-supported computational research. Three recommendations were identified by the workshop as clear areas that would energize paths forward.

Theory and analysis components should be integrated seamlessly within experimental workflow.
Achieving this goal will make it much more straightforward to couple theoretical guidance into experimental design. Combined with on-the-fly comparison between prediction and experimental results, the feedback could improve the effectiveness of a unit of experimental facility time by steering experiments in progress. Adopting common data formats and community toolsets for analysis and workflow would further enhance experimental productivity by simplifying data transmission, analysis, and archiving. In these areas it was clear from all discussions that ASCR's investment to date in visualization and analysis tools applied to current and emerging experimental work in BES user facilities can have a significant near-term impact on facility effectiveness and scientific impact.

There is a need for advances in both theoretical and computational research to reach the goal of seamless analysis. There are varying architectures that could support enhanced workflow. Streaming architectures are being investigated and applied for these applications, as noted in examples above. Local computational power associated directly with a beamline (or beamlines) has been applied historically for acquisition and analysis of user data. Further research in computational sciences, linked directly to the problem of an experimental workflow, is needed to clarify the relative strength of these approaches, and to offer clear guidance on appropriate architectures on existing and planned systems. It may be that one approach will not provide the best solution for all facilities or users, but it should be possible to make reasoned choices for analysis components based on consideration of current and future needs.

Advances in theory are leading to more accurate predictions of physical and chemical behavior, but often at the cost of being computationally intensive in their own right. Research leading to faster calculations based on existing models, or to fast approximations that can help to guide ongoing experiments would be of value in enhancing the overall efficiency of experimental workflow. Close integration of theory with actual experimental results is a key factor in this process, as discussed further below.

Many experiments in spectroscopy and dynamics are inherently four-dimensional in space (or momentum) and time (or energy). Research into optimal ways to visualize and quickly analyze experimental results for interesting features or phenomena will strongly enhance the ability of experimentalists to focus on areas of the phase diagram of particular interest for a given sample.

Move analysis closer to experiment.
While this presents a greater challenge more appropriately addressed in an intermediate time frame (2-5 years), the potential of this approach to simultaneously reduce data volume and increase experimental productivity is profound. Success in this effort can make possible real-time, streaming analysis at

the beamline, including such local data reduction capabilities as zero suppression, hierarchical filtering, baseline and background subtraction, and other core data analysis processes. The goal of a live visualization of ongoing experiments will improve the efficiency of individual experiments and overall facilities and greatly simplify and streamline subsequent off-line analysis.

Research supporting this goal builds naturally on the areas suggested for workflow problems. Moving 'closer' is in time, not necessarily in physical location. Thus research in computational sciences aimed toward understanding advantages of 'local' computer capability versus streaming-data concepts using multiple nodes is needed to enable informed decisions on future data analysis architecture.

Predictive theory and modeling has a central role in achieving this goal. Taking modeling beyond the prediction of structure and properties to a prediction of the expected experimental result on a particular instrument could close the loop between theory and experiment during (or near) the time of the actual measurement. Tackling this problem requires not only sound fundamental theory and fast algorithms, but a direct consideration of instrument capabilities (e.g., Q-range and resolution for diffraction experiments. Close collaboration (e.g. research teams) involving beamline scientists and theorists is essential to the experiment/theory feedback process.

Match data management access and capabilities with advancements in detectors and sources.

Progress toward the ultimate goal of maximizing scientific and societal impact of user-facility research will be greatly enhanced by close future coordination of computational, data-management, and experimental capabilities. In the near-term, ongoing and planned efforts are aimed at removing the bottlenecks related to data communication, storage, and manipulation by applying existing data transport and mobility toolsets. In the mid-term, greater use can be made of forefront mathematical techniques to more efficiently extract science from individual experiments. In the long-term, the goals focus on the seamless interoperability of data systems across different facilities/beamlines, the ability to combine multiple data sets and legacy data in a comprehensive analysis framework, and to establish and maintain the integrated teams of engineers, scientists, and computer scientists needed to solve emerging problems and challenges.

Workshop participants support a concerted effort to match data capabilities with experimental facilities. Teams with a critical mass of experimental, computational and theoretical capabilities are best able to take advantage of forefront facilities to produce timely, high-impact science. Within and across DOE laboratories, computational and experimental facilities have strong drivers for collaboration in research, and joint efforts are expanding as the advantages become more apparent. Targeted joint research programs aimed at expanding and cementing these relationships are beginning; programmatic support specifically for this purpose

could provide important additional stimulus for these collaborations, to the benefit of the entire user community.

Summary: This workshop was by design broad in its consideration of data challenges facing BES user facilities and the possibilities for BES-ASCR cooperation and collaboration. A number of potential paths forward to build on these areas warrant consideration. Programs such as SciDAC aimed toward specific areas identified here could make significant contributions toward near- and mid-term advances in data handling. Future workshops could profitably focus on single areas or issues identified here, with an eye toward a more detailed description of challenges and approaches to solve individual problems. Additional attention to the importance of the role of data specialists within experimental facilities could help to foster communications between experimental and computational facilities and programs, and with the experimental user groups who increasingly must depend on facility support for data needs.

Finally, success in opening and pursuing this dialog is expected to lead to the identification of new science grand-challenge questions best addressed at the leading edge of combined experimental and computational science and incorporating the best ideas of both user communities. The workshop laid the foundation for continuing discussion and useful actions in addressing data needs and challenges in DOE SC user facilities.

Agenda

Sunday, October 23, 2011

6:00 - 9:00 pm	Working Dinner and Organization Meeting	Organizers and Chairs Timberlawn Room

Monday, October 24, 2011

8:00 - 8:30 am	Registration Open Continental Breakfast	Salon D
8:30 - 9:00 am	Welcome and Goals: BES and ASCR	Harriet Kung, BES Dan Hitchcock, ASCR
	Conference Chairs	Michael Simonson, ORNL
9:00 - 9:40 am	Workflow Management	Brent Fultz, Caltech
	Theory and Algorithms	Thomas Schulthess, CSCS
10:20 - 10:40 am	Break	Salon D
10:40 - 11:20 am	Visualization and Analysis	Dave Pugmire, ORNL
11:20 - 12:00 pm	Data Processing and Management	Quincey Koziol, The HDF Group
12:00 - 12:15 pm	Charge to the Participants	Peter Nugent, LBNL
12:15 - 1:30 pm	Working Lunch with Speaker	Adam Riess, JHU & STScI - 2011 Nobel Laureate in Physics
1:30 - 3:00 pm	Parallel Sessions	
	Workflow Management	Middlebrook Room
	Theory and Algorithms	Salon D
	Visualization and Analysis	Timberlawn Room
	Data Processing and Management	Great Falls Room

3:00 - 3:15 pm	Break	Salon D
3:15 - 4:30 pm	Resume Sessions	
4:30 - 5:30 pm	Reports of Session Progress	Chairs
5:30 - 7:00 pm	Poster Session and Light Refreshments	Salon D
7:00 pm	Adjourn for evening	

Tuesday, October 25, 2011

8:00 - 8:30 am	Continental Breakfast	Salon D
8:30 - 10:30 am	Resume Breakout Discussions	
10:30 - 11:00 am	Break	Salon D
11:00 - 12:00 pm	Plenary Reports from Discussions	Chairs Salon D
12:00 - 12:15 pm	Closing Remarks and Paths Forward	Organizers Salon D
12:15 - 1:30 pm	Workshop Adjourn: Working lunch for organizers and chairs	Salon D
1:30 pm	Report preparation	Organizers and Chairs Salon D

Participants

Name	Institution
J. Michael Simonson	ORNL
Peter Nugent	LBNL
Workflow Management	
Brent Fultz (Plenary)	Caltech
Mark Hagen (Chair)	ORNL, SNS
Ross Miller (Chair)	ORNL
Ilkay Balatsky	LANL, SNL
Charles Bouldin	NSF
William Johnston	LBNL, ESNET
Raj Kettimuthu	ANL
Scott Klasky	ORNL
Kerstein Kleese-Van Dam	PNL
Gabrielle Long	NIST
Robert McGreevy	ORNL, SNS
Razvan Popescu	BNL
Andreas Roelofs	ANL, CNM
Kurt Schoenberg	LANL
Dean Williams	LLNL
Kathy Yelick	LBNL, NERSC
Dantong Yu	BNL

Participant List (continued)

Theory & Algorithms	
Thomas Schulthess (Plenary)	CSCS
Bobby Sumpter (Chair)	ORNL, CNMS
Chao Yang (Chair)	LBNL
David Brown	LBNL
Jaquelin Chen	SNL, CRF
Alok Choudhary	Northwestern
Ashley Deacon	SSRL, SLAC
George Fann	ORNL
Ian Foster	ANL
Ioannis Kevrekedis	Princeton
Bala Krishnamoorthy	Washington State University
Alex Lacerda	LANL
Stefano Marchesini	LBNL, ALS
Normand Modine	SNL, CINT
Jeff Nichols	ORNL
Raymond Osborn	ANL
Amedeo Perazzo	SLAC
John Rehr	University of Washington
Robert Ryne	LBNL
Sean Smith	ORNL, CNMS
Michael Sternberg	ANL

Participant List (continued)

Visualization & Management	
Dave Pugmire (Plenary)	ORNL
Dean Miller (Chair)	ANL, EMC
James Belak (Chair)	LLNL
Wes Bethel	LBNL
Jim Ciston	LBNL, NCEM
Francesco de Carlo	ANL
Alex Hexemer	LBNL, ALS
Graham Heyes	Jefferson Lab
Chris Jacobsen	ANL, APS
Chandrika Kamath	LLNL
Karren More	ORNL
Kenneth Moreland	SNL
Ronald Nelson	LANL
D. Frank Ogletree	LBNL
Eric Stach	BNL, CFN
Guebre Tessema	NSF
Craig Tull	LBNL
Jon Woodring	LANL

Participant List (continued)

Data Processing & Management	
Quincey Kozoil (Plenary)	HDF Group
Amber Boehnlein (Chair)	SLAC
Eli Dart (Chair)	ESNet
Alexander Balatsky	LANL, SNL, CINT
Michael Banda	LBNL
Shane Cannon	LBNL, NERSC
Peter Denes	LBNL
Thomas Devereaux	SLAC
Steve Dierker	BNL, NSLS-II
John Maclean	ANL, APS
Michael Miller	ORNL, SHARE
Ruth Pordes	FNAL
Nagi Rao	ORNL
Lauren Rotman	ESNet
Arie Shoshani	LBNL
Piotrek Sliz	Harvard
Rick Stevens	ANL
Lee Ward	SNL
Paul Whitney	PNL
Kevin Yager	BNL

Participant List (continued)

Observers			
Name	**Program**	**Name**	**Program Office**
Laura Biven	SC-2	Sandy Landsberg	ASCR
David Boehnlein	HEP	Peter Lee	BES
Rich Carlson	ASCR	Eliane Lessner	BES
Lali Chatterjee	HEP	Natalia Melcer	BES
Jim Davenport	BES	Thomas Ndousse	ASCR
Susan Gregurick	BER	Van Nguyen	BES
Bill Harrod	ASCR	Karen Pao	ASCR
Robin Hayes	BES	Thiyaga Pappanan	BES
Dan Hitchcock	ASCR	Mark Pederson	BES
Linda Horton	BES	Katie Perine	BES
Helen Kerch	BES	Walt Polansky	ASCR
Dorothy Kock	BER	David Price	BES
Phil Krashaar	BES	Andy Schwartz	BES
Jeff Krause	BES	Wade Sisk	BES
Harriet Kung	BES	Lane Wilson	BES