OPEN ACCESS

The SNS/HFIR Web Portal System for SANS

To cite this article: Stuart I Campbell *et al* 2010 *J. Phys.: Conf. Ser.* **247** 012013

Journal of Physics: Conference Series **247** (2010) 012013 doi:10.1088/1742-6596/247/1/012013

The SNS/HFIR Web Portal System for SANS

Stuart I. Campbell, Stephen D. Miller, Jean-Christophe Bilheux, Michael A. Reuter, Peter F. Peterson, James A. Kohl, James R. Trater, Sudharshan S. Vazhkudai, Vickie E. Lynch

Oak Ridge National Laboratory

Mark L. Green

Tech-X Corporation

campbellsi@ornl.gov

Abstract. The new generation of neutron scattering instruments being built are higher resolution and produce one or more orders of magnitude larger data than the previous generation of instruments. For instance, we have grown out of being able to perform some important tasks with our laptops. The data sizes are too big and the computational time would be too long. These large datasets can be problematic as facility users now begin to struggle with many of the same issues faced by more established computing communities. These issues include data access, management, and movement, data format standards, distributed computing, and collaboration with others. The Neutron Science Portal has been designed, and implemented to provide users with an easy-to-use interface for managing and processing data, while also keeping an eye on meeting modern computer security requirements that are currently being imposed on institutions. Users can browse or search for data which they are allowed to see, run data reduction and analysis applications, perform sample activation calculations and perform McStas simulations. Collaboration is facilitated by providing users a read/writeable common area, shared across all experiment team members. The portal currently has over 370 registered users; almost 7TB of experiment and user data, approximately 1,000,000 files cataloged, and had almost 10,000 unique visits last year. Future directions for enhancing portal robustness include examining how to mirror data and portal services, better facilitation of collaborations via virtual organizations, enhancing disconnected service via "thick client" applications, and better inter-facility connectivity to support cross-cutting research.

1. Background

With the advent of the Spallation Neutron Source (SNS) at Oak Ridge National Laboratory, an opportunity occurred to revisit how to provide software and services via the facility. Under the guidance of Ian Anderson, then Director of the Experimental Facilities Division (XFD), a number of Neutron Science Software Initiative (NeSSI) workshops were held in 2003 and 2004 to solicit feedback from the community on its requirements and wishes for next generation software [1]. The primary outcome was the desire to have a portal system which would help manage and process data for users. For easy access, it was decided that the Neutron Science Portal (portal) would have a web interface via a web browser not requiring users to install software. Portal design work commenced in

2004 and 2005 with an operational portal deployed by April 28, 2006 when first neutrons were detected at SNS. These first data were incorporated into the portal and are still available today.

Since the early days, the portal has grown considerably. It holds data from SNS [5] and High Flux Isotope Reactor (HFIR) [6] as well as the Low Energy Neutron Source (LENS) at the University of Indiana Bloomington [7], some example data from the NPDF instrument at the Lujan Center [8], and the entire IPNS [9] data repository. The portal has approximately 370 registered users, contains almost 4TB of data (much of it is compressed) in almost a million files. On an average day, the portal has about 20 individual visitors, receives about 4500 web site “hits”, and performs about 260 compute jobs. Thus for SNS and HFIR users, the portal has established itself as a regular fixture as part of performing experiments and data analysis.

2. What is the portal?

A user accesses the portal via a web browser. In order to do this and to utilize portal computing resources, the user must first obtain an XCAMS computer user account [13], which must be requested and granted. Once approved, users can login to the portal whereupon they will have access to their “workspace”, consisting of pointers to experiment data that they are allowed to view. Data are stored by facility, instrument, proposal IDs in order to support providing a multi-facility data repository.

In addition to providing data access, the portal also provides access to applications and computing resources. These same resources are available both to on-site users at SNS and HFIR, and to users external to the facility. The portal provides a suite of applications which the user is able to launch remotely, but the applications appear locally. This capability is made available via NX remote desktop technology [14] accessed via the browser. The NX remote desktop feature enables the portal to present an application to the user as though it runs locally, while also taking advantage of remote computing resources. The application backend can run on a portal server, create and submit jobs to a job queue, and return results to the GUI and the user can access at their sites.

Segmenting computation from the user interface layer enables utilization of a wide array of computing resources. GUI applications run by users at SNS on analysis computers can run locally. Applications launched by remote users can be run on separate portal servers and computation servers at SNS. Other resources include the Oak Ridge Institutional (OIC) cluster [15] and the NSTG cluster [16]. Both of these resources are not managed by SNS staff and thus have specific rules for accessing these resources which could be difficult for most users to deal with. To facilitate using these resources, community accounts are utilized where a job for a portal user is run as the community user on these resources. The Application Manager is the intelligent subsystem within the portal which handles transparent execution of user jobs on a variety of aforementioned target environments. In addition, it handles seamless data staging in and out of these resources, job monitoring, logging, failover and recovery.

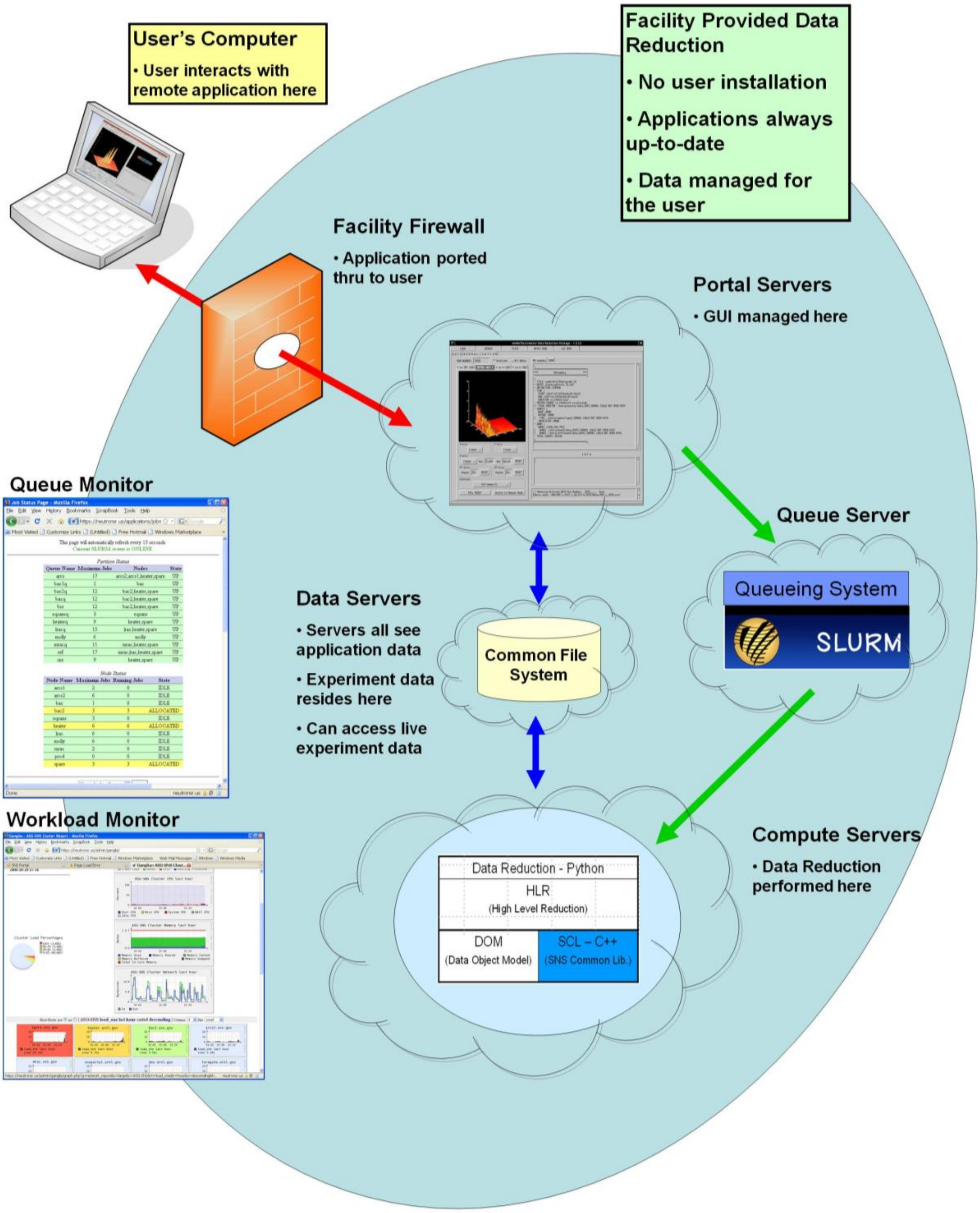

Figure1: Portal Based Data Reduction showing "thin client" remote application access.

Thus the portal includes data browsing and search, but also provides Graphical User Interface (GUI) tools that are connected to powerful computing resources. Applications include experiment data histogramming tools, data reduction tools, data visualization tools, and McStas [4] Monte Carlo simulation. Due to the ability to run applications remotely, the portal also supports hosting of community produced applications such as ISAW [17], DAVE [18], DANSE [19] applications, EXPGUI [20], FullProf [21], GSAS [22]. Database tools such as the Cambridge Structural Database

[23] and the Inorganic Crystal Structure Database [24] are also available. Experiment support tools such as a Sample Activation Calculator are also available which enables users to compute the estimated radioactive dose rate that specified samples will produce for a given instrument.

The portal development effort has produced a significant code base. Currently the composite portal development effort has produced over 650,000 lines of code since the onset of portal development. Data reduction applications account for approximately half of this code base. Software development follows a rigorous process for defining the content of each release. A centralized software build process is utilized to coordinate building correct versions for test and deployment. Testing is done to ensure that features meet requirements, and upon successful completion, the new version is deployed. Software are maintained in a subversion (SVN) [25] repository with a web based issue tracking system called Trac [26] placed atop SVN.

3. Small Angle Scattering Application

Figure 2 shows the Small Angle Scattering Reduction GUI as seen from the portal. It should be noted that the data shown is pre-normalised commissioning data. The vertical lines that can be seen on the plot result from the fact that the detector bank for EQ-SANS at the SNS consists of interleaving detector tubes. The vertical red lines are dead tubes that will be rejected in the data reduction. The bottom right plot displays the counts against neutron time-of-flight of the non-rejected pixels.

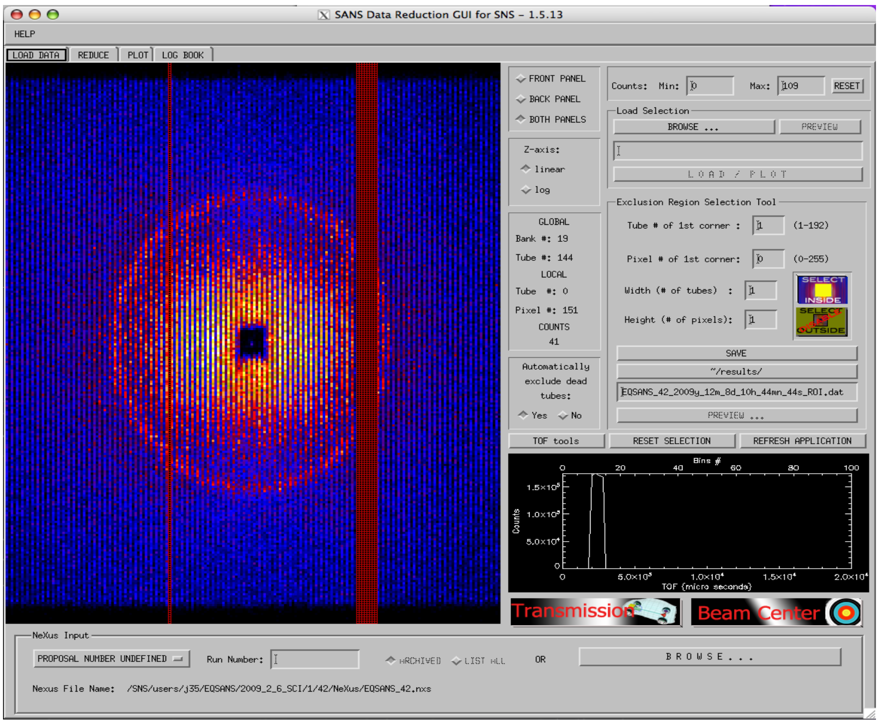

Figure 2: A screenshot of the SANS Data Reduction GUI as seen via the portal. This is the main view of the reduction application.

There has been a great deal of time and effort involved in fitting such a large number of options and functionality into a user interface that is both small and intuitive to use. A tab based design was implemented to try and optimise the use of screen real estate, and arranged / sequenced in a left to right fashion to indicate the order of the operations and user interactions. The experimental data is entered by specifying the run number of the measurement, the system locates the file and passes data onto the compute resources that will be used for the calculation.

The results of the reduction can either be output directly to the users home directory (inside a results directory) or to a shared directory for that measurement. The latter enables users to easily share data and results within the experimental team.

There is a great deal of functionality contained within the application, some examples of which are fitting, beam centre calculations (figure 3), region selection and transmission calculation (figure 4).

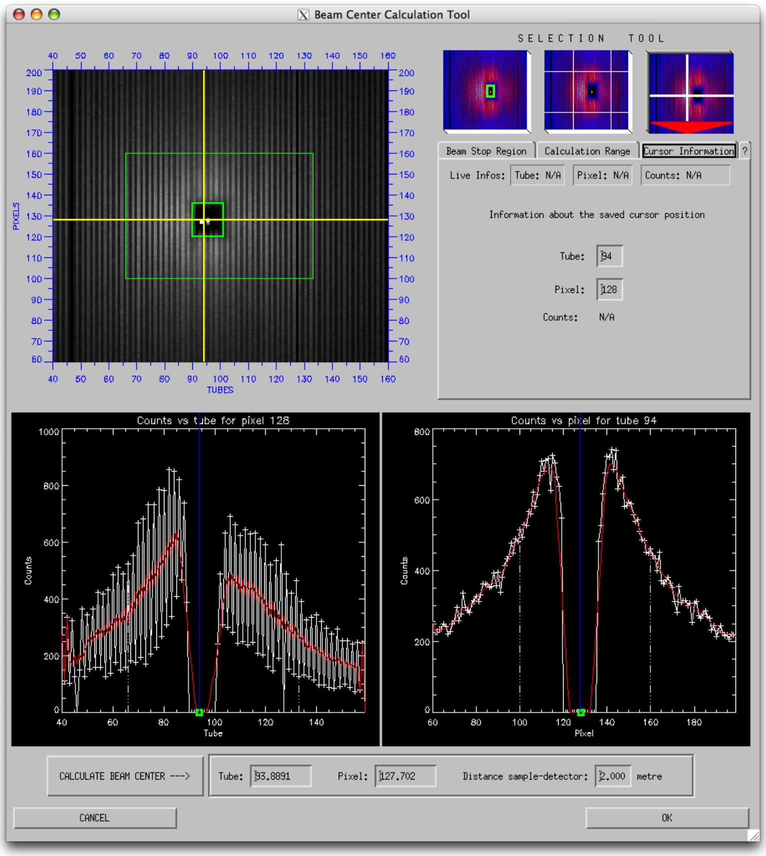

Figure 3: A screenshot of the beam centre calculation tool GUI. This allows the user to determine the beam centre in a user friendly manner.

Every operation that is performed on the GUI is logged to a journal, which can be sent to the developers with a single click from the GUI itself. This helps both the users and developers by easily being able to capture any problems that may arise.

Although at the moment the user runs this application via the portal, it is possible that this GUI forms part of a 'thick client' infrastructure that would allow the user to run the application on their local systems but be able to utilise the computational resources at the facility. Included in this is the intelligence to determine which jobs or calculations should be performed remotely and what is more efficient to execute using local resources.

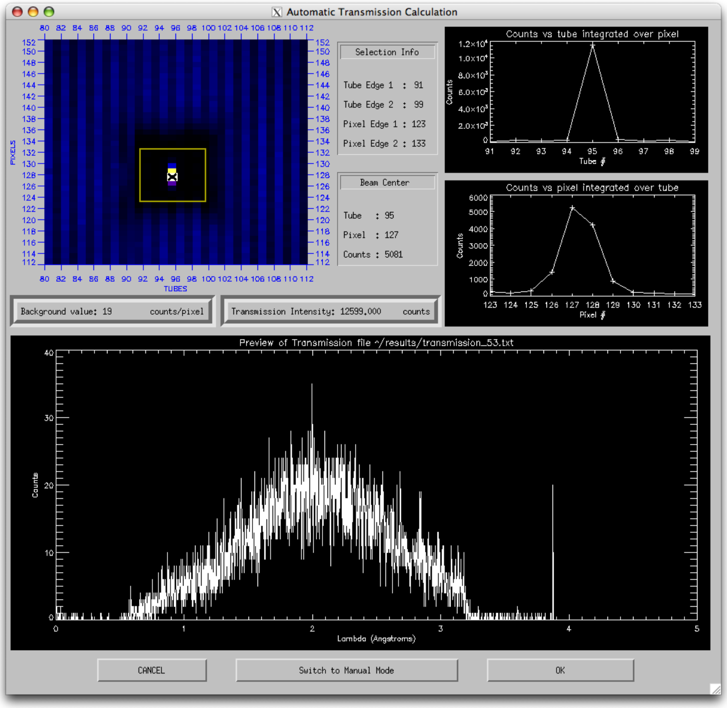

Figure 4: A screenshot of the automatic transmission calculation GUI.

4. Why is the portal usage growing?

A simple answer for why the portal usage is growing could be down to the fact that the data volumes being generated are continually growing and it is becoming difficult for users to move it between systems easily. The facility creates the data and thus has significant influence over how convenient it is for users to acquire, access, and process data. In the past, users collected raw data at their instruments, but this is no longer the primary mode of operation at SNS, which has invested considerable effort into these major capabilities that significantly improve convenience:

1. Experiment metadata are automatically captured and associated with the experiment data. These data are translated into a self-describing file format called NeXus [27], which utilizes the underlying HDF-5 scientific data format [28].
2. The live cataloging process takes experiment data as it is streamed from the instruments, catalogs pertinent metadata, and stores the data for access. The entire process is quick and places data in the centralized repository in a matter of seconds to minutes.
3. There is a high level of integration between user experiment support systems and the data management system. For instance, the data management system integrates with the proposal system (IPTS) [29] in order to capture and associate user and experiment information with experiment data.
4. The facility curates the data on behalf of the user, thus addressing worries users may have for where they store their data, and the integrity of this data.

These functions are all conducted automatically on behalf of the user, in a transparent manner and their data is available for access via the portal using a basic web browser. Thus the complexities of preparing, moving, cataloging, and discovering data are all handled for the user. The portal offers both data browsing and data search for discovering data. A path structure had to be determined in order to facilitate locating data contained within a multi-facility data repository. The generalized path stores data by facility, instrument, proposal ID, collection, and run number. In this way, users can uniquely locate their data across facilities and instruments. The search interface offers users a powerful means for quickly locating data. Search is particularly useful in helping those users who have performed a number of experiments to locate their data. The search interface enables users to locate data by a number of metadata tags such as proposal ID, run number, Principle Investigator (PI) name, sample name, and date. But the user can only locate data which they have permission to access according to the proposals with which they are affiliated. Thus, in addition to the XCAMS authentication system, there is an authorization layer that grants fine-grained access. The facility is capable of providing this data privacy as it has knowledge of the experiment team members via the proposal system. Additional users can be added to the proposal as the PI desires using the proposal check-in utility of the portal.

Once data are located there are a number of things which a user can do such as inspect the data, examine the metadata, or visualize the data. The portal facilitates examining a number of standard file formats such as ASCII text, XML, NeXus, along with a number of imaging formats such as jpg. In particular, one can inspect the NeXus metadata of interest by looking at the attribute groups. The data within the NeXus files can also be visualized. The ISAW visualization tools have been adapted to work via the portal, but needed to be adapted to handle the case where the data could be too large to send over the internet. Thus the visualization tool operates in a client-server mode where only the necessary data are sent over the network link. To enhance data exploration, data can be viewed in tabular form, 1D or 2D plot. The 2D plot can present the data in either linear or log color scales.

The portal also provides users the ability to both upload and download data, and users can freely move data to and from the portal. This begs the question why users do not simply take all of their data home with them to reduce and analyze. They can certainly do so. however, users do not usually take the data with them out of convenience, but out of necessity. One feature of the NeXus files is that they utilize internal compression for the data much like a ".zip" file. The difference being that the metadata within the file are still readable. The HDF-5 utilities perform compression and decompression on behalf of the user. It is common to achieve 90-95% compression with the NeXus files. Thus copying a NeXus file to one's laptop and reading the data can produce an out-of-memory error. For instance, a typical 40MB to 50MB NeXus histogram file expands to approximately 4GB. The largest instrument at the SNS, TOPAZ [30] the data file will expand to almost 15GB. Thus the data are simply too large to process efficiently via typical user laptops, especially when one considers that data reduction typically requires processing a number of files concurrently – perhaps as many as six.

Realizing that the data are large, the facility also provides another convenience – facility produced and maintained data reduction applications are also made available via the portal. Integrated within the portal is a mechanism which enables GUI applications to be displayed locally on one's own computer, however the application is actually running back on facility computers. Note that paging techniques do exist which would make it possible to process such data on a laptop, however there is a question of the feasibility for processing very large data sets this way. Utilizing this web based "thin client" approach has a number of significant advantages:

1. The user always has access to the most recent version of the application without having to perform installation.
2. Data are kept proximal to computing thus the amount of time to move data is minimized.
3. The facility can provide access to substantial computing resources which users may not have available to themselves otherwise.

The portal is useful for providing individual researchers convenient access to computing and data, however it is already providing more than that by facilitating collaboration. In the past, data sharing required the interested parties to figure out where to place the data in a mutually accessible, and safe location. Depending upon the size of the data, finding such a location is not always easy else we would simply email it to our collaborators. Within the portal data hierarchy, each experiment proposal subdirectory has within it a 'shared' subdirectory to facilitate collaboration. These are the only read-write subdirectories within the repository as all others are read-only. In addition, only the experiment team that can access the experiment data can access these shared areas. Thus experiment team members can place data in an area common among themselves in order to share data. The shared areas are the fastest growing regions in the data repository. All of the above features ensure that the portal attracts a steady influx of users.

5. Future Plans

There are several notable growth areas for the portal ranging from data, computing and networking, and new application development. Integration with X-ray sources to make X-ray data available for co-analyses with neutron scattering data has good scientific impact potential. Working with the X-ray facilities is one step towards developing a user facility network (UFNet) which could be used to facilitate cross-cutting scientific research. Other portals exist in the X-ray and neutron scattering communities, and new portals are likely on the horizon. Inter-operating with these portals will be important but will bring significant challenges in federating their respective user authentication systems.

The future also looks to hold a new way of working by enabling users to utilise the ever-increasing portable computing power available to them. To illustrate, the web browser thin client approach does not require the user to install applications and relies upon the server-side for most of the work. But a "thick client" would enable a user to perform work on their local computer while also providing a remote interface to portal infrastructure enabling data movement and remote computation at the discretion of the user. Thus computing via a thick client would shift work away from a remote server onto a user's own computer based upon what makes sense to be done locally. Work has begun in producing a thick client via an Small Business Innovation Research (SBIR) project lead by Mark Green of Tech-X Corporation. Similarly, the ISAW package is also being adapted to leverage remote data and computing resources.

Successful next generation applications will enable scientists to more fully focus on their research and less on the nuances of computing, yet will provide access world class cyberinfrastructure and scientific computing applications. Users will have more freedom of choice and flexibility for how and where they perform data analysis and computing. The future of computing for facility users looks bright.

Acknowledgements

The authors would like to thank J.K. Zhao for the data used in the figures.

References

[1] NeSSI: http://neutrons.ornl.gov/workshops/nessi/NeSSI_agenda_linked.pdf
[2] SQRL: http://sqrl.eecs.utk.edu/
[3] NSTG: http://www.teragrid.org/gateways/projects.php?id=56
[4] McStas: http://www.mcstas.org/
[5] SNS: http://neutrons.ornl.gov/facilities/facilities_sns.shtml
[6] HFIR: http://neutrons.ornl.gov/hfir/hfir_facilities.shtml
[7] LENS: http://www.iucf.indiana.edu/lens/
[8] Lujan: http://lansce.lanl.gov/lujan/
[9] IPNS: http://www.pns.anl.gov/
[10] Tech-X Corporation: http://www.txcorp.com/
[11] UTK SEELAB: http://www.cs.utk.edu/~seelab/
[12] Diamond Synchrotron Facility: http://www.diamond.ac.uk/
[13] ORNL XCAMS: https://xcams.ornl.gov/xcams/
[14] Nomachine: http://www.nomachine.com/

[15] OIC Clusters: https://oic.ornl.gov/
[16] NSTG Cluster: http://www.ornl.teragrid.org/guides/userguide.html
[17] ISAW: http://www.pns.anl.gov/computing/isaw/index.shtml
[18] DAVE: http://www.ncnr.nist.gov/dave/
[19] DANSE: http://danse.us/
[20] EXPGUI: http://www.ncnr.nist.gov/programs/crystallography/software/expgui/expgui.html
[21] FullProf: http://www.ill.eu/sites/fullprof/php/programs.html
[22] GSAS: http://www.ccp14.ac.uk/solution/gsas/
[23] CSD: http://www.ccdc.cam.ac.uk/products/csd/
[24] ICSD: http://icsdweb.fiz-karlsruhe.de/
[25] SVN: http://subversion.tigris.org/
[26] Trac: http://trac.edgewall.org/
[27] NeXus: http://www.nexusformat.org/Main_Page
[28] HDF: http://www.hdfgroup.org/HDF5/
[29] IPTS: http://www.ornl.gov/sci/iums/ipts/
[30] TOPAZ : http://neutrons.ornl.gov/instrument_systems/beamline_12_topaz/index.shtml