

NSLS-II Data Acquisition, Management and Analysis (DAMA) Review

June 26, 2015

Review Committee

Mark Rivers (University of Chicago, Chair)
Stuart Campbell (Oak Ridge National Laboratory)
Mark Heron (Diamond Light Source)
Pete Jemian (Advanced Photon Source)
David Skinner (Lawrence Berkeley National Laboratory)

Executive Summary

NSLS-II has taken a bold and broad approach to end-to-end data services. From an operational perspective this means wide involvement from detector to community data lifecycle and comes with proportional resource costs. There are some important aspects of the system, such as authentication and security that have not yet been addressed. Identifying parts of the solution which can be leveraged from the community, doing less in-house, and leveraging resources from the DOE complex are all ways of preventing a resourcing shortfall. Breaking the controls, data acquisition and analysis requirements down in to a number of phased projects, should help manage the risk by ensuring focused delivery on overall project needs.

General Comments

The committee sincerely appreciates the time and effort made in preparing the background document, "NSLS-II Plan for Data Acquisition, Management, and Analysis (15 March 2015)". We also thank the presenters for the effort that went into the presentations. They were very high quality, and the presenters did an excellent job of answering the questions that were posed by the committee. We also thank the beamline scientists who candidly shared with us their experience, perspectives and concerns about the status of the beamline controls and data acquisition software.

We note that the planning and implementation of software for beamline data acquisition, management, and analysis has started very late. Our understanding is that this is largely due to constraints that were placed on the project scope. The controls and data acquisition groups are scrambling at this time to provide software that can be used for commissioning and first experiments, together with defining the overall architecture and implementation.

We have organized this report first according to the four categories of the committee charge: Technical, Scope, Risks, and Resources. Within each category we describe our findings, comments, and recommendations.

Charge category 1. Technical:

Charge questions:

Is the DAMA architecture/framework suitable to the data needs of NSLS-II?

Is the architecture/framework scalable to support expected growth in this area?

Is it flexible enough to support the wide range of techniques/applications and the wide range of user expertise?

Findings:

NSLS-II is a medium energy light source offering best in class of brightness and flux from IR to hard x-rays. The construction project officially ended in March of this year, with all of the Key Performance Parameters being exceeded. There are 7 beamlines currently being commissioned, with a total of 25 to be commissioned by 2017. 20 of these beamlines will be operated by BES.

The beamline staff for controls, data acquisition, software support and information technology is expected to be about 1.1 FTE per beamline.

Organizationally there is a DAQ and Computing group within Photon Sciences. A search for a group leader for this group is currently in progress. Some of the staff working in this group are formally part of the Controls Group in the Accelerator Division. The total number of staff in these two groups is currently 10.3, with plans to expand to 28.2 by FY 2017.

The high brightness of NSLS-II, coupled with the recent developments in detector technology mean that data rates from NSLS-II beamlines are predicted to be very high. Several detectors are capable of data rates in excess of 1 GB/s, and rates of 10 GB/s are not far away. This means that datasets of several TB may be collected every hour.

The data acquisition plan for NSLS-II is based on data residing in a number of separate data stores. These include:

1. Electronic logbook
2. Detector frame store
3. Metadata store
4. EPICS channel archivers of accelerator and beamline information
5. Proposal database
6. Safety database

Access to raw and processed data will be through a Data Broker layer that hides the formats and locations of the files that are used to store the data. The correlation of the data with an experimental scan will rely heavily on the timestamps associated with each data item.

Data are stored as event documents (e.g. a point in a scan). Each event has a timestamp and a unique uuid. Scalar data is stored in the event data, non-scalar data is stored by uuid reference.

Experiment control will be done through a new framework being developed called Ophyd. This is the replacement for SPEC, but written using a modern computing language (Python) and using the EPICS layer for device support. It contains the other major components of SPEC: plotting, configuration management, command line, macros, sequencer, and diffractometer support. While Ophyd was initially targeted at only data collection it has evolved into a larger umbrella encompassing data collection, management and analysis. Ophyd implements a Device Model that abstracts the underlying hardware, using concepts such as signals, signal groups, and pseudo-positioners. Positioners and Detectors are composite objects that participate in scans. For diffractometer support they are collaborating with Soleil on the libhkl tool. Ophyd implements an EPICS Channel Access server (pcaspy) to export its functions to clients. Ophyd provides an abstraction for areaDetector, which is being re-worked because users were not happy with the initial implementation. Ophyd is in active development, with current work on libhkl, refactoring of the sequencer, user session environment and device model enhancements.

The computing infrastructure is planned to initially include only storage and processing resources at each beamline. The facilities will be tailored to the needs of each beamline, and will ensure that data collection depends only on local resources. They will have at a minimum an analysis server with 80 TB of disk storage.

A prototype data center is currently being established in the control system computer room. This will be a modular system with initial tiered storage of 1.08 PB, and a initial computational cluster with 10 nodes. The nodes each have 36 conventional cores and 20 GPU cores, connected with 10 Gb, not InfiniBand. It will use OpenMPI for parallelization and HTCondor for scheduling. This facility is funded at the level of \$2M per year. The beamlines will decide which data will be pushed to this central resource.

There is currently a collaboration with RHIC to explore using their data center for archiving, since they have a large tape system. There are long-term plans (2020's) for a large data center somewhere at BNL, but this is not currently funded. One possibility that was mentioned is converting the old NSLS building 725 for this purpose. There are also plans to explore external or internal cloud-based storage, such as Amazon Glacial.

There is a plan to develop a file exporting tool that will allow the users to define what data they want to write to a file, and the format to be written (e.g. HDF5). This will provide a mechanism for users to take data home with them. However, the primary emphasis is for data to reside at NSLS-II with users remotely analyzing their data, rather than bringing it home. It is anticipated that in many cases the size of the datasets and the sophisticated analysis tools required will be beyond the computing resources available at the users' home institutions.

Comments:

From the presentations and discussions at the review it is clear that the NSLS-II has assembled a very talented and enthusiastic group of software engineers. They are well versed in modern programming methodologies and tools.

From a software perspective the architecture/framework design is well thought out and employs several best-of-kind open source approaches: MongoDB, asynchronous execution, and immutable data, to name a few. The use of Python throughout as well as modular components and a conservatively chosen set of layered dependencies reflect solid scientific software engineering. The design invites a wide range of techniques/applications and user expertise. The design balances ease-of-use notebook approach with close-to-the-metal command line interfaces. The design is scalable to increases in the number of beamlines, and with increases in detector data rates.

The flexible approach to data (nosql, noschema) is a good fit with the diverse set of data at NSLS-II. The metadata server is a key component in the NSLS-II DAMA. Careful attention should be paid to the scalability, and function of this core component.

The proposed framework is quite ambitious and very different from that used at any existing synchrotron facility, which typically use a file-based model for data storage and retrieval. This means that very few existing tools can be re-used at NSLS-II for data collection and analysis, without deviating from the NSLS-II architecture. It follows that developing the new framework is a large task that will require substantial programming resources. There are currently only 2 programmers working on Ophyd, which is a critical tool for beamline operation. Additionally, there will be competition for these resources for the task of commissioning 18 additional new beamlines in the next two years. This is discussed in more detail in the final section of this report on Resources.

The framework is still in its early stages of implementation, and is not yet capable of meeting the needs of beamline scientists, even for many commissioning activities. To overcome this beamline scientists have developed some of their own Python tools outside of Ophyd to allow them to do the necessary commissioning. The relationship of the controls and beamline staff appears to be positive. The controls staff are trying to learn the requirements of the beamlines with on-site observation, and the beamline scientists are not constrained from contributing to the beamline controls software.

The experience at existing facilities is that traditional step scans are being rapidly replaced by on-the-fly scanning, where the hardware does single or multi-axis motion and provides hardware trigger signals for detectors. This is critical for efficient data collection, where the overhead of starting and stopping positioners cannot be tolerated. It appears the Ophyd sequencer was designed for the traditional step scans, and that on-the-fly scanning did not figure prominently in the design. It is important that on-the-fly scanning be supported in a robust and general manner.

A critical part of the framework is the timestamping of the data in each data store. This is what allows the Data Broker to assemble events and return data for a scan. However, as on-the-fly scanning is implemented the scalar data will be buffered in devices like the SIS3820 where the data is stored in a memory array in the device. Each point in the scan does not have an associated time stamp in the SIS3820. How will such array data be stored: as arrays or will it be broken out into individual scalar values? If so, how will each one be timestamped? We heard that there will be an FPGA card with an Event Receiver, and the triggers to the SIS3820 will capture timestamps on the FPGA card. Has this been implemented and tested? This is a capability that is needed very soon in order to use on-the-fly scanning. Similarly it is not practical to accurately time stamp each frame from many commercial area detectors, such as Pilatus or Eiger, as this is handled by the hardware and software developed by the original equipment manufacturer without this requirement. Having custom versions of such detectors appears unlikely and doing in-house designs very resource inefficient.

libhkl is still under development, but successful operation of this is required for many beamlines. Re-use of existing and proven libraries would be preferable to developing those libraries again. Was there a careful review of the diffrac code from Diamond before deciding to develop libhkl? Is it possible to re-use the available SPEC geometry source code in libhkl?

More generally, the committee expressed concern that very little other than EPICS V3 and V4 are being re-used from existing sources. NSLS II's resources are limited compared to other projects. The NSLS-II team seems very 'plugged' into EPICS collaborations but not as aware of beamline computing beyond EPICS.

There are several layers in which pseudo-positioners can be implemented, for example in the Delta Tau controller, in the EPICS database, or in the Ophyd application. We heard about implementing a monochromator energy pseudo-positioner in Ophyd. What criteria are used in deciding at which level to do this implementation? It would appear that implementing this particular pseudo-positioner at a lower level has the advantage of both performance and decoupling from higher level tools.

The Data Broker is intended to hide the physical location of data. As the data is migrated from the beamlines to the central data storage, and ultimately perhaps to tape archive how is this location information made available to the Data Broker?

The experience at Diamond Light Source was that deploying local storage and analysis cluster on the beamlines did not scale with the needs of the data challenging (PX and Tomography) beamlines. If sufficient computing resources are placed at a beamline to meet the peak demand then they will tend to be under-utilized most of the time. It makes more sense to move the large-scale storage and computing to a central facility where the loads are more balanced.

Recommendations:

Below is a short summary of recommendations discussed at the review. The technical/topical area of concern precedes each item.

(software architecture) There is a concern that the various services (Metadata Store, Data Broker, File Store, Ophyd, etc.) are being created without a clear statement of requirements and API for each. Per *NSLS-II-Plan-DAQ-Analysis-Management-v5.docx*, the API for each service should be clearly stated. Where are the interfaces defined to manage and communicate between these services? It should be possible to replace a service if its interface is sufficiently well defined.

(data scaling) MongoDB is designed to support “sharding”, i.e. storing data across multiple machines. This will be needed in the future as the MongoDB database grows in size. NSLS-II should experiment with sharding before they need it.

(modular portability) The architecture is sufficiently appealing that as it matures it may well be implemented at other facilities. The NSLS-II should do what it can to encourage this, since it will likely lead to more resources to develop it, and more robustness in the end product.

(software re-use) The NSLS-II should avoid wherever possible re-writing tools such as ISpyB, where established solutions exist. They should look to join existing collaborations to leverage what already exists and then work with the collaboration to influence future direction.

(network & security) We recommend that a separate network security review be scheduled. Having an ESnet representative at future DAMAs is recommended. Additionally the single 10 GB/s external link from NSLS-II is likely to be insufficient, and should be upgraded.

Charge category 2. Scope:

Charge questions:

Are there areas that have not been identified that are required to deliver the 'end-to-end' data services for the user?

In case of limited resources, what areas can be deemed to be less critical?

Findings:

The scope of the plan presented to us was quite comprehensive. It included data acquisition, visualization, and analysis.

Only a relatively small part of the plan has been implemented so far. In particular the data analysis codes do not exist in a state that can use the proposed architecture.

Comments:

There were several areas that the committee identified that have not received much if any consideration in the current planning.

The first is data access security. At most current synchrotron sources data access is through disk files, for which there are existing robust access controls. The NSLS-II framework uses a set of data stores and a Data Broker to access data, hiding the file implementation. How will data access security be implemented? What are the data ownership concepts? If a user is collecting data at the beamline then presumably the detector data is theirs, and should not be accessible to other users, at least for some period of time. But what about other types of data, i.e. the monochromator energy, sample name, sample positions, etc.? How much of this information is public, and how much is private to that user, since it could be used to allow a competitor to obtain important information? Is there a mechanism to implement data security on such data?

Another area that has not been considered is handling of incorrect metadata. This is a daily occurrence at all synchrotron beamlines. Some examples include:

1. The monochromator energy is found to be slightly incorrect, and is recalibrated. All data collected previously that day should use the new calibration, not the value that was stored with the data at the time it was collected.
2. The user entered the wrong sample name by mistake.
3. The user forgot to change the pixel size for their tomography data when they changed the objective lens.

How will the Metadata Store be updated to use the corrected values for these quantities? How will the Data Broker know which metadata values should be used for a given scan?

Some companies are conducting MX experiments on an international scale, using data from various facilities. They handle their own data pipeline, starting from as early as placing the sample into the instrument. How is this considered? Assuming metadata errors will occur, what scalable systems or algorithms will be used to identify errors?

While the immediate needs are focused towards experiment control, data acquisition and initial analysis, the aspirations of the project will necessitate pipelines to support automated analysis, provision for a suite of science specific applications for analysis during users' visits and post visit (on site) and potentially a post visit via remote analysis services. These will require physical resources e.g. rooms, workstations, storage cluster etc, together with a mechanism for remote analysis.

Recommendations:

Credentialed, authenticated access to stored data must become a requirement. The NSLS-II should consider holding a separate security review as these plans are being developed.

The top priority must be robust and performant operation of these core functions: data acquisition, data broker, and file exporter. This will permit early users to take home useable data. Bringing the data analysis tools in-house will be necessary in the longer run, but should be lower priority until the data acquisition component is complete.

There must be a mechanism to correct errors in the metadata after the data has been collected.

An plan for innovating ahead of emerging NSLS-II detector data rates is advised.

The data retention policy needs to be developed taking into account the cost of retaining information at each level of tiering. This may need to be done in consultation with DOE/BES, and may be resource limited.

Charge category 3. Risks:

Charge questions:

What are the areas of concern/risks with regards to this architecture/framework?

Are there alternative approaches worth investigating?

Findings:

We were provided an estimate that another 12 person-years would be required to finish developing the services in this framework. There is additional effort required to develop each of the science applications.

Comments:

The technical risks of the proposed architecture appear to be small. The framework is flexible, open, and extensible.

The primary risk is that the implementation of a robust system could require more time and resource than is available before the NSLS-II must begin producing science. For example, work has not yet begun on the file exporter. There must be at least a short-term workaround to this problem, so that users can leave with complete data sets, including essential metadata. The reviewers saw some great software talent at the review, but suggest a review of the software team sizes.

The central data facility appears unlikely to be ready in time for the first high data rate experiments. This means that resources must be placed at the beamline, where they may not be as efficiently used.

The various software modules (Ophyd and Data Broker as two examples) each have a small number of developers who understand the source code. These developers have highly desirable skills, and their loss would pose a challenge.

At present the delivery of the data acquisition and analysis appears as one open ended development. Breaking this down to a number of projects with clear deliverables and identified resources against an agreed time line would reduce the open-endedness. It would also serve to protect development effort from unplanned operational support which can easily trump important core activities.

Recommendations:

The architecture has a strong reliance on the central data stores. These need a defined disaster recovery plan. The Research Data Alliance is one of many resources in leveraging community experience in data planning.

The Data Acquisition Group Leader remains to be appointed. This is a critical position in driving the beamline data acquisition and analysis project forward. Not filling this position soon is a risk, and likewise bringing in someone with a different architectural view is a risk.

There has apparently been a prohibition against deploying SPEC on any beamlines, which has led to some frustration of the beamline scientists particularly on the NextGen beamlines. While the controls staff should not spend significant effort on a stop-gap solution, there are many SPEC experts among the NSLS beamline scientists. Configuring SPEC to control an EPICS beamline is not a large task. Deploying SPEC as a stop-gap measure should be considered if Ophyd cannot be made sufficiently robust and usable on the required timescale.

Consider defining a number of phased projects to deliver the overall data acquisition and analysis requirements for NSLS II.

Charge category 4. Resources:

Charge questions:

Are the planned operational resources sufficient?

Are we appropriately leveraging other activities across the DOE complex or elsewhere?

Findings:

There is an aggressive plan to increase the number of beamlines from 7 to 25 in the next two years. During this time the Controls, Data Acquisition and Analysis staffing is planned to increase from 10 to 28.2.

Comments:

The 4-fold growth in the number of beamlines in the next two years itself an very large undertaking. This will be happening at the same time that the core functionality of the data acquisition system needs to be completed. While this is accompanied by a 3-fold increase in the number of controls staff, these staff will require training to come up to speed on the framework. The resources appear low, particularly in the 2 to 3 year timeframe with beamlines both being operated and being constructed. It also appears low compared to other international projects such as ESRF and Diamond.

The planned development of computing infrastructure does not match the expected growth in data rate, data volume and analysis services. \$2M/year may be sufficient in the longer run, but it may not be enough for the initial required investments.

The computing resources required to process NSLS-II data analysis will be substantial and possibly led by the crystallography and tomography beamlines who expect multi TB per sample. Between options for local clusters, new centralized computing at the NSLS building, or distributed approaches it is not clear where the computing power will come from. The ongoing collaboration with RHIC on long-term data retention is positive, but will not be sufficient. Realizing that beamlines vary greatly in their data demands, the right answer may include a mix of local and distributed computing.

Recommendations:

If the core infrastructure for the framework is to be completed in a timely manner more staff need to be added to this effort, and they need to be isolated from the demands of commissioning the new beamlines.

There are a number of high-quality Python data acquisition and data analysis solutions in place at other synchrotron facilities. For example, Matt Newville (the author of PyEpics) has excellent Python software for scanning fluorescence microprobe and x-ray absorption spectroscopy that are

in use at the APS. Clemens Prescher has developed high-performance Python powder diffraction integration software Dioptas (<https://github.com/Luindil/Dioptas>) that is an excellent replacement for the archaic Fit2D software from the ESRF. The NSLS-II should use such types of solutions (at least in the short term) whenever possible.

We advise BNL leadership to take up the issue of centralized computing resources sooner rather than later. (For reference 150TB of LCLS xtal data requires 130k CPU hours to process). BNL leadership should work with other DOE facilities. The BES Facilities Computing Working Group (BESFCWG) may have useful designs as it regards leveraging shared computing.

ESnet is a powerful partner in leveraging remote resources. We recommend ESnet participate in future reviews.

Algorithms and software are also areas to leverage. Not as recommendations but by way of “appropriately leveraging” DOE resources we suggest cross-checking against the following R&D agendas for collaborative opportunities.

Applied Mathematics Research

- Improved methods for data and dimension reduction to extract pertinent subsets, features of interest, or low-dimensional patterns, from large raw data sets;
- Better understanding of uncertainty, especially in messy and incomplete data sets; and
- The ability to identify, in real time, anomalies in streaming and evolving data is needed in order to detect and respond to phenomena that are either shortlived or urgent.

Computer Science Research

- Extreme-scale data storage and access systems for scientific computing that minimize the need for scientists to have detailed knowledge of system hardware and operating systems
- Scalable data triage, summarization, and analysis methods and tools for *in-situ* data reduction and/or analysis of massive multivariate data sets;
- Semantic integration of heterogeneous scientific data sets
- Data mining, automated machine reasoning, and knowledge representation methods and tools that support automated analysis and integration of large scientific datasets, especially those that include tensor flow fields
- Multi-user visual analysis of extreme-scale scientific data, including methods and tools for interactive visual steering of computational processes

Next-generation Networking Research

- Deploying high-speed networks for effective and easy data transport
- Developing real-time network monitoring tools to maximize throughput
- Managing collections of extreme scale data across a distributed network