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

April 20th 2017

Review Committee
Simon Billinge (Columbia) – Chair
Alun Ashton (Diamond Light Source)
Thomas Proffen (ORNL)
Jonathan Taylor (European Spallation Source)
Shelly Ren (ORNL)
Nicholas Schwarz (APS, Argonne)

Executive Summary

The review committee would like to thank the DAMA group staff and management, and the controls, IT and NSLS-II senior management for giving their time and energy to prepare for, and present at, the review.

The DAMA group has managed to produce a software stack of the highest quality under the most difficult of conditions, and despite staff attrition and the absence of permanent line-management, has produced software for experiment control and data acquisition that is stable, flexible, maintainable and scalable, and universally admired by the beamline scientists. This is quite an accomplishment and deserves high praise. The preceding staffing situation was not a stable or sustainable one, which was appreciated by NSLS-II management, and concrete steps have been taken to steady the ship with the successful recruitment of permanent leaders for the controls program and the DAMA group (Richard Farnsworth and Stuart Campbell, respectively). Initial indications are that the situation has significantly stabilized with these hires. However, our view is that this is just the start of a fairly long journey that must be undertaken for the group to reach its stated goals of 'end-to-end' data collection to analysis software so that NSLS-II can deliver the high quality science that is essential for its success, and extreme vigilance needs to be maintained. With the arrival of the new management, the relative maturity of the Ophyd rewrite and the Bluesky product, and the impending hiring of 2 new staff to fill vacant roles (with the expectation for two additional hires to bring the group back to its former strength), we feel strongly that it is time for the group to undertake a serious planning exercise, resulting in a detailed, prioritized, and resourced plan that defines scope that will be delivered and a plan for delivering it, based on NSLS-II scientific priorities and realistic resources.

Risks

Because the DAMA mission is not currently clearly defined it was not easy for the committee to identify specific project risk areas. However, we list below some of the largest risks that we believe the group faces, broken out into subcategories, in its broad mission to provide software to deliver science to NSLS-II

- Objectives
    - Lack of clarity on objectives and priorities
- Staffing
    - Retention of core competencies
    - Correct blend of personnel to achieve objectives
    - Provision of basic needs with currently allocated resources
    - Achieving objectives with limited IT and controls support
- Expectation management
    - Top level management expectations
    - Instrument scientist expectations
    - Between groups within the controls program
    - Developer career expectations
- Security authentication on data management
    - Integration with PASS
- Working in isolation
    - Between groups in control program
    - With rest of the world
- Scientific producitivity of the facility will depend heavily on M and A
    - Data graveyard

Recommendations

The committee summarizes here its recommendations. Context and more detail around these points may be found in the findings and comments sections below.

- Create a short term staffing plan to assess expertise gaps and evaluate current roles versus expertise and staff expectations. This can be used to ensure that the impact of the current open recruitments are maximized. Our recommendation is that in the short term the group recruit the 4 FTEs needed to bring the resources available back to 2015 levels.
- Together with the DAMA group stakeholders, evaluate what scope can be delivered in the near term:
    - What scope can be effectively covered with the current resources.
    - What scope can be covered with +2 and +4 FTE.
    - What level of operational support can be effectively delivered with +2 and +4 FTE.
    - This evaluation should be used as a tool to allow prioritization of future developments and tensioning between development and support.
    - Involve program level management to ensure that there is no scope creep between groups
    - In the evaluation, collect and sift requirements from operational and construction phase beamlines.
    - Review the current estimated data rates in light of experiences at other facilities. They currently seem very high and certainly unsupportable with current resources.

    - Revisit the recommendations of the 2015 review in light of current progress, expectations and resource availability.

- Prepare a 5 year strategy, with a bottom up estimation of the required resources that match the prioritized scope and schedule of the NSLSII instrument buildout:
    - Development and support for acquisition
    - Development and support for data management
    - Development and support for data analysis

  A multidisciplinary project approach including controls, DAMA, user office, beamline staff and IT is required to ensure that all needs are covered. The plan should outline what core staff resources should be provided by NSLSII, what resources could be provided by external areas at BNL and what resources could be provided from either external funding or collaborative developments. Prioritize the plan and make the tasks transparent to the entire organization
- NSLSII put in place a mechanism that will allow scope and priorities to be assessed and agreed for the group's activities.
    - Prioritization of scope delivery should be governed by the scientific priorities of the facility.
    - The facility and group should remove scope areas (or shift scope to a later stage in the facility's development) if resourcing and priorities dictate.
    - The group and its stakeholders should when possible ensure that common tools be developed in areas such as data analysis and data collection and experiment controls GUIs. In so doing the amount of resources required will be minimized.
- In the long term the NSLSII leadership recognize the desirability of, and work towards, a DAMA staff level of average 1FTE per beamline. This does not include controls and IT staff. Our view is that this level will result in large gains in scientific scope and impact.
- The group assess, agree and communicate what level of development defines a completed task or feature, This will allow limited resources to be moved into other areas of development.
- Leverage external resources to the extent possible, for example, controls and IT groups, and CSI

Technical Aspects

Findings

The technical software stack being developed by DAMA has evolved since 2015 with priority being given to ensuring robust capabilities for data acquisition and data capture. This objective has been successful based on feedback from the beamline staff who praised the stability. The software has also evolved to be as flexible as possible. This has (correctly in a resource limited environment) been at the cost of ease of use. Data analysis has not significantly evolved, beyond giving users their data to analyze at home..

The current software stack comprises of:

Ophyd – device abstraction layer above EPICS that creates a reusable hardware layer for transferable experiments (Python)

Bluesky – end user environment for experiment specification and execution that sits above Ophyd (Python)

Data Acquisition UI - currently a command line interface working in an iPython session. Because the chosen environment is iPython, a browser-based implementation using Jupyter notebooks has also been demonstrated but ipython running in a terminal is currently the standard UI at the beamlines.

Databroker – a python based interface to saved experiment data and metadata that could be from multiple sources. The database underpinning the databroker is currently MongoDB with instances distributed around the ring one per beamline.

Document model/event model – Experiments are modeled as a series of synchronous and asynchronous events that can be defined by users as event streams. The document model bundles these into documents with a schema for metadata and data. Currently little pre-defined structure is imposed, though a minimal required dataset is being agreed.

Analysis Tools - some rudimentary python tools were presented and some initial investigations into more domain specific tools.

The software stack is well engineered using current good practice and appropriate technologies that have not currently been exceeded by experimental demands or changes in implementations. Two members of the DAMA group were clearly expert in these technologies though it was not clear to what extent they understood the detail for each stack or just specific areas they worked on.

Although not presented, the tools that were being used for development, including GitHub for open source community development and a stool/platform testing, continuous integration, deployment and monitoring were described.

Most of the software stack was reviewed in 2015 but no planning or process was outlined on how the current environment was reached.

As currently implemented there is still no security model on any of the data or the metadata once on the NSLS II network. There was no plan outlined. The platform supports rich metadata capture, but there is currently no 360° experiment metadata capture, as largely speaking, captured metadata is specified by the user. A critical component underpinning successful experiments, data acquisition and analysis is the underling IT landscape, hardware architecture and implementation. The current NSLS II status was presented to the review by Robert Petkus who was open about the current fragmented situation of a mix of local, domain specific and centralized infrastructure. The technologies being used e.g. a standard linux distribution and deployment, GPFS file systems were sound. A development plan for this area was mentioned but no indication of how realistic this would be given current resource constraints. A mixed

vision of local and centralized resources may not currently be considered best practice in the domain.

Little work on data analysis was presented which is indicative of the current priorities and lack of domain expertise in the group. Collaborations such as those for live tomographic reconstruction were highlighted though not utilizing NSLS II data given there is no imaging beamline currently operational.

No technical details of the relationship with CSI were outlined. No technical details of long term data storage or preservation were outlined other than the policy of retaining experimental data for 100 days.

Remote access to NSLSII is currently challenging and only available through a VPN service with a high overhead for setting up user access.

Comments

Ophyd – the ophyd layer is a critical component that was well reviewed in 2015. However, the 2015 implementation was not sufficiently stable and flexible and this layer has received significant refactoring since then. Some discussion took place on support for this layer being taken on by the controls teams since it is the interface layer to the hardware; however, the general feeling was that this is sufficiently user-facing that significant oversight should be retained by the DAMA group.

Bluesky – this layer was not reviewed in 2015 and has been developed in the intervening years. It is a minimal level acquisition tool with excellent and robust functionality well respected by science staff. However, there is little control and methodology behind how bluesky scripts are captured, versioned and shared and this was recognised by the developers. Training in both bluesky and python has been previously given to science staff.

Data Acquisition UI - one of the priorities for the science staff was an easier interface to Bluesky as the current Bluesky command line interface required too much staff intervention and involvement in a user's beamtime.

Databroker – in 2015 some concerns and recommendations were raised into the underlying technology used for the databroker database and some of those concerns e.g. security, scalability and speed seem to have come to fruition with changes in implementation being used as workarounds.

Document model – one of the strengths of the data model is its flexibility but the lack of standardisation will become a problem. The development of base/standard metadata set is a good start.

Analysis Tools - Analysis at a synchrotron facility is a vast domain with extensive legacy and specialised developments given the variety and evolution of hardware and techniques. Although highlighted by the group, there is little prioritization or a plan to exploit or develop this area based on NSLSII science program.

Expectations and Objectives

Findings

The NSLS-II draft strategic plan outlines an end-to-end vision for data acquisition and control, data management, data analysis and modeling and simulation. This is a compelling vision and in line with changing user expectations when it comes to data collection, analysis and modeling support. The short term focus, as you would expect from a new facility, is on data acquisition and instrument control using EPICS and managing data and metadata.

The DAMA group does not seem to have a vision or plan related to a complete instrument workflow, user interfaces and data analysis and modeling in general.

Staffing especially in light of new beamlines coming online in 2017 is not adequate to deliver on end-to-end user software needs from data acquisition to data analysis.
As part of the DAMA strategy and plan, it is important to have a blueprint to outline the systems to be built, and how they are integrated with each other. The BNL PASS system is currently disconnected from the NSLS II systems such that users have to authenticate multiple times and enter same metadata into various systems. Because the proposal system is segregated, the proposal team members are not known to the DAMA systems, and experiment data is wide open to all who have access to the beamline computers. This presents a major security concern and a scientific integrity issue because researchers should only have access to their own data unless the data becomes public per facility policy or PI's authorization. Authentication and authorization are critical part of the data management. They are not easy tasks, but neither could be missed. Is it feasible to leverage the resources from BNL IT to implement a BNL wide authentication and authorization service? BNL clients like NSLS II could redirect the users to the BNL hosted authorization server for authentication and authorization.

GUIs, Sample databases and hkl libraries were items brought up by the beamline scientists as their top priorities.

Once captured, the data analysis needs of the beamlines could be clearly categorized into a tiered approach to support e.g.:

Tier 1: fully supported within the facility with expertise, a development program with or without collaborators

Tier 2: partial support with deployment of domain software with little in house knowledge but recommendations

Tier 3: no support.

Resources and Scope

Findings

The scope of the DAMA group broadly encompasses everything from device abstraction from EPICS to data analysis, including data acquisition, computer experiment control, data treatment and correction simple data analysis and complex data modelling. The group has the scope of

installation and integration of experiment control and acquisition for beamlines in the construction phase and support of beamlines in operations.
It is clear that for NSLSII to succeed in its mission the bottleneck of data interpretation through data analysis must be minimized. The scope for this requirement lies within the DAMA group. In other works the scope that the group has is mission critical for the scientific success of the beamlines and therefore of the facility.
The group's vision is to provide an end to end user experience from acquisition to analysis. It is therefore reassuring to see that the group's vision matches its scope. That is to say that the group and all of its members understand the mission and its importance for scientific impact.

With that said the resources available to the group are not currently commensurate with the end to end vision which has led to some operational issues. The group has seen a fairly high staff attrition rate since the last review. 4 members have left. The group has authority to recruit 2 staff with a possibility of a further 2 positions, which would see the resource level back to that of 2015.

The low staff numbers has led to issues associated with remaining staff having to cover areas previously covered by others, which in some cases is away from staff expectation of their role. It is reassuring that despite of these operational issues the group has continued to develop the required data acquisition and visualization software stack in a way that matches the build out of beamlines. That is to say that in an environment of limited resources the group has prioritized the correct area of experiment control and data acquisition and produced a software stack that is admirably stable, well tested and flexible. This fact is one contributing factor for the high regard the group's work is held by the beamline staff.

It was excellent to see the team spirit that was evident between the beamline staff and the DAMA group. However with limited resources the scientific program is already starting to suffer, it was noted by some beamline scientists that whilst data can be collected in large volumes the tools (and time to develop those tools) is limiting the rate at which publishable analyzed data can be produced. At this early stage in the operations of NSLSII, and in light of the aggressive / impressive beamline buildout this should be seen a warning light to the senior management.

The low level of resources has led to an implicit contraction of scope, the area of data analysis is not being resourced in a manner that is commensurate fashion with respect to the existing operational beamlines and those that are expected to be built in the near future.
Overall the group covers the existing beamlines with .3FTE / beamline. *This is simply too low*. The work that is done on analysis at this stage seems to be uncoordinated with the instrument schedule.

Whilst there is a plan to recruit possibly 4 staff, even at this staffing level there the group will be resourced limited. The existing staff are working in areas, for example, low level controls development, which is not best suited to their skill set. Whilst this is maintainable in the short term it will be a push factor for staff.

The plan to leverage external lab sources for aspects of data analysis is sensible, as is the plan to collaborate with other photon sources to minimize resources required for implementation of new functionality, that said this strategy will take time planning and a significant amount of management effort to set up, thus the lead time may be longer than expected.
As operations continue to grow the amount of staff must grow in order to support the user program, irrespective of the amount of effort available from outside sources.

In respect of maintaining delivery and development of experiment control and data acquisition the group cannot afford to lose further experienced staff.
Regarding staff recruitment the management should carefully plan what skills are essential to ensure the stability of the group is maintained.

Overall the findings are that the resources for the group do not match the scope. However the group is *managing* to cover the key priorities of DAQ and controls for operations in a way that is efficient, maintainable and extensible.

Comments
The DAMA group (all levels) were fully aware of the resourcing issues, the effect of which had led to a focus on experiment control and data acquisition. There was an expectation from the staff that the resources will be increased.

One requirement that was of high priority from the instrument scientists was development of GUI's for data acquisition, areas of scope such as GUIs require a great deal of time to develop into production ready systems. The group does not have an experienced GUI developer which is something that would ideally be addressed, either by implementing a hiring strategy or by finding external expertise.

One of the many challenges for the new management will be retaining expertise , building a coherent group, and maximizing everyone's potential. Therefore, the next two hires will be very critical. The DAMA group members and beamline scientists wanted helpers to handle the day-to-day data acquisition for the operating and commissioning beamlines so the developers who are currently working on data acquisition can be released from operation support to data analysis, which is what they are better trained for and more passionate about. The data acquisition software tools are relatively mature. A nice GUI is desired by the beamline scientists. A streamlined release plan and process will make the release more efficient and robust.

Data management is an extremely critical component as data is ultimately the final product of the facility. Data alone is useless without the associated metadata. Data Broker is implemented to store metadata and locate experiment data. A central storage is desired to provide load balance and to avoid single point of failure. As the experiment data is migrated from the beamline computer to central storage, and eventually to the tape archive, the Data Broker needs to be upgraded to indicate where the data resides. In addition, does NSLS II have a disaster recovery plan that would enable not only backup data but also efficient recovery?

Management and Organization

The newly on board NSLS II Controls program manager and DAMA group leader were on-target for the challenges that the facility is currently facing. The group morale is high, and the group members are excited to finally have the stability, guidance, and support they deserve. The DAMA group leader did an excellent job on working closely with the DAMA developers, conducting management observation, communicating with beamline scientists, reaching out to the Brookhaven National Laboratory (BNL) business units, and seeking collaboration opportunities with Computational Science Initiative (CSI) at BNL and other synchrotron and neutron facilities around the world. While the acquisition software tools (BlueSky and Ophyd) received very high praise from the instrument scientists, data management and data analysis work should start very soon in order to support the user program and boost science productivity. The high staff turnover rate and limited resources presented a huge risk to successfully delivering the scientific solutions, especially in data management and analysis. Some concerns were also expressed about a high dependency on single point of failures within the groups.

Strategy

Findings

The DAMA group and its parent program at NSLS-II have developed a comprehensive high-level vision addressing the current and anticipated computing needs for beamlines at NSLS-II in the coming years. This vision is documented in the facility's five-year plan, and associated presentations made to the NSLS-II's Scientific Advisory Committee (SAC). At a high-level, the vision touches on all key aspects of the data lifecycle from acquisition through data management, data analysis, and the publication of results. This is a problem common to many such facilities.

The facility-wide adoption of a single yet flexible tool in this space has already begun to reap benefits in terms economies of scale and reuse, and will continue to in the future. This effort is commendable as other facilities have strived to achieve this to no avail. For the near-term, the group has planned the roll-out of experiment control software for beamlines that will be commissioned and brought into production, in alignment with facility-wide plans for the build-out of beamlines.

At present the group (and facility) has a limited data retention policy, stating that the ultimate responsibility for data retention falls on the user group performing the experiment. The facility will attempt to provide best-efforts to keep data for at least one year.

The DAMA group plans to rely heavily on collaborations the BNL's Computational Science Initiative (CSI), and with outside organizations such as CAMERA and software efforts at the APS and other facilities to achieve its data analysis objectives. Strategizing in this area has only just begun.

The development of good working relationships with the internal users of the DAMA group's tools is very important to realizing the group's and facility's vision. The group should again be commended for developing excellent working relationships and trust with key scientific staff and management.

Comments

The high-level vision of the group to provide "end-to-end" or full data lifecycle management is a vision very much needed by modern user facilities such as NSLS-II. The group's vision as stated is very ambitious and too high-level (with the exception of experiment control/acquisition) to constitute a reasonable strategy needed in order to plan or justify resources needed to achieve the desired end state. Technical roadmaps and plans for the deployment of BlueSky at new NSLS-II beamlines are well defined for the near-term. There is no strategy and there are no plans in place for the development of data analysis and reduction tools, and data management distribution tools needed to ensure the facility will be scientifically productive. This is an area of great concern that should be addressed as soon as possible.

As stated above, the group's work on BlueSky is commendable. The development of a single, modern experiment control package, and its continued support and deployment has shown to be a good strategic decision and should continue. The group's desire to collaborate with other facilities and with BNL's Computation Science Initiative (CSI) will be critical to helping the group achieve its goals. These collaborations are only in the planning stages at present, and with time and cultivation, will hopefully pay off.

It is apparent to the reviewers that the boundaries of the DAMA group are unclear with relation to other support groups at NSLS-II and BNL. It is also unclear to many as to the group's future plans. This causes confusion and frustration for both the group's staff as well as for those the group is attempting to serve.

It is would be desirable for the group to be exposed to more activities in the scattering science domain and that beamline staff also play a part in sharing their expertise of activities and software developments at other facilities.