National Synchrotron Light Source II response to NITRD RFI for National Big Data R&D Initiative

Paul Zschack, pszschack@bnl.gov, (631) 344-8703
Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 119735000

National Synchrotron Light Source II (NSLS-II) is a state-of-the-art 3 GeV electron storage ring designed to deliver world-leading photon intensity and brightness over a wide spectral range including infrared, ultraviolent, and x-ray. The facility is operated as a DOE Office of Science User Facility for the benefit of widely diverse scientific user communities. When fully developed, the facility will host up to 4000 users each year, and will include over 60 simultaneously operating beamlines, each providing different advanced capabilities. NSLS-II will offer researchers from academia, industry, and national laboratories new ways to study material properties and functions with nanoscale resolution and exquisite sensitivity by providing state-of-the-art capabilities for x-ray imaging, scattering, and spectroscopy. Users will come from all 50 states (and from around the world) to study problems across many scientific disciplines. This research is sponsored through various channels, including NIH, NSF, DOE and others.

High brightness and high flux available at many of the NSLS-II beamlines coupled with advanced high-speed megapixel detectors will drive much higher data rates and volumes. Indeed, all synchrotron disciplines are increasingly using multi-element detectors to efficiently measure and record the results of the experiment, and certain techniques often utilize detectors that can produce multi-megapixel images with frame-rates in the kilohertz range. The volume of data associated with a single experiment may exceed several tens of terabytes, and in the next few years, the facility may produce nearly 20 petabytes annually.

The broad uses of new megapixel detectors together with new, advanced techniques drive a new paradigm for data processing and analysis at NSLS-II and at synchrotron sources worldwide. Individual users are often unable to write the sophisticated algorithms required to process, analyze, visualize, interpret or otherwise extract the important information from raw detector data. Multi-dimensional data sets are intrinsically more difficult and often require specialized data reduction and manipulation algorithms and routines not readily available to individual users.

Building and extending an infrastructure (hardware and software) to improve access for users who are generally not expert in Big Data problems, computing, visualization, etc… is essential to realize the full potential of NSLS-II. While users of the facility are often leading in their respective disciplines, many will not have the necessary resources at their home institution to productively develop knowledge from the data they collect. Further, the ability for users to remotely access, manipulate, analyze and visualize their data will be essential.

Many complex problems span orders of magnitude in spatial or temporal extent. These multi-scale problems are well suited to Big Data approaches that are well adapted to work with heterogeneous data sources. These data sets, together with data collected at one or more synchrotron beamline may be combined to form a more complete picture of a sample or process. Furthermore, users are increasingly applying widely varying and complimentary analytical techniques to study complex materials or processes. Combining these other measurements with data collected at the light source provides an opportunity for a comprehensive understanding and suggests a new pathway for discovery. Tools to facilitate these complex and transformative approaches are needed, as well as research into quantifying uncertainties when combining different data sets.

As synchrotron experiments become more and more sophisticated, there exists an increasingly greater need for adequate theoretical modelling and simulations. Many experimental results have a significant dependence on coherent properties of the beam as determined by beamline optics and beam defining apertures. Therefore, better understanding of how a complete experiment works including the effects of the source and optics is important for correct interpretation of the experimental data. Parameters of the beamline (metadata) are collected together with detector data and can be used for simulation and modelling to greatly enhance the NSLS-II capabilities in experimental planning and design. Improved interpretations of experimental data will provide research results with a greater impact on the research activities in variety of scientific disciplines.

Finally, the value of archiving users' data must be weighed against the costs associated with ensuring the long term sustainability, access, and curation of users' data. Without a detailed process knowledge of the sample, the raw data from any users' experiment does not have particularly significant value to other researchers. The experimental results are largely determined by the particular sample and its preparation or processing. On the other hand, as new data analysis methods evolve, older data may still hold considerable value for new discovery. Well-articulated policies on storage of large data sets should be developed and appropriately resourced.