

BES Lightsource Facilities ASKD White Paper

Recent improvements in detector resolution and speed and in source luminosity are yielding unprecedented data rates at BES's national light source and neutron source facilities. In the past 2 years, Advanced Light Source (ALS) scientists have seen their data volumes grow from 65 TB/year to 312 TB/year, and in another 1-2 years will be generating 1.9 PB/year. These data rates exceed the capabilities of data analysis approaches and computing resources utilized in the past, and will continue to outpace Moore's law scaling for the foreseeable future. The national BES software landscape is largely ad-hoc and relies heavily on a limited number of experts to handle analysis. The result is that BES beamlines' efficiency and ability to address important scientific questions are diminished.

Next Generation Light Source (NGLS) data volumes and the associated challenges will be orders of magnitude larger than seen today at the ALS. The growing consensus within light source scientific communities is that scientific insight and discovery at BES facilities are now being limited by computational and computing capabilities much more than by detector or accelerator technology. "Past limitations of detector technology have been largely solved through recent DOE investment and that the present bottleneck for research throughput is the lack of availability of appropriate analysis software and modeling tools."¹

The trend of large data sets and high frame rates in light sources is however just the beginning of the massive data volumes that are anticipated. The brightness of synchrotrons, not including free electron lasers, is expected to grow many orders of magnitude over the next years; schemes are also being developed to upgrade existing light sources by a factor of up to 10,000 in brightness. In addition detectors are rapidly evolving into very high frame rates with large pixel count and high bit rates. Currently commercially available photon counting detectors for hard X-ray science produce up to 500 images a second with each frame at 1 million pixels and 20 bit dynamic range. The next generation is anticipated to run in the kilohertz range. With potential data rates of up to 1.5GB per second per beamline and ten of beamlines per synchrotron, a very sophisticated solution has to be developed.

ASKD Challenges for BES Facilities include:

- **Big Data management, access, and sharing for large, and growing datasets:** Scientists need an end-to-end solution for access, management, and collaborative sharing of light source data. Such a system would deliver data in realtime to compute resources large enough to handle large-scale analysis of the data. The system would need to handle very large data sets such as are generated at BES national light sources, and allow us to scale to next generation data rates.
- **Real-time feedback during beamtime for data quality assurance, and for in-situ, time-resolved experiments:** Scientists conducting experiments at

the ALS require a real-time analysis environment to answer fundamental questions of performance, scalability, and accessibility. Such a system would allow scientists the fastest feedback possible on the integrity of their scientific samples, the detector and accelerator performance, and software suitability. It would also permit realtime steering of time-resolved, in-situ experiments.

- **Diversity of data, science, detectors, and methodologies at light sources:** Problems addressed by modern day light sources are very diverse. The science questions range from biology, material science, physics to earth science and archeology. A system with broad applicability and that accommodates that diversity will most effectively leverage efforts from multiple communities and have the greatest scientific impact.
- **Multimodal data, metadata, and analysis for lightsource science:** Domain scientists routinely conduct experiments at multiple accelerators and/or beamlines and utilize different experimental techniques (eg. Tomography and GISAXS) to fully understand their samples. A single framework that can support best-of-breed analytics for diverse experimental techniques, data, and metadata will allow researchers to quickly extract scientific knowledge.
- **Small science teams and usability:** BES facilities (Light Sources and Neutron Sources) serve ~10,000 researchers per year. Most light source experiments are conducted by small groups of 10 or fewer researchers in concert with beamline scientists at each facility. Software for data management, analysis, and simulation which will have the greatest impact on the quality and quantity of scientific productivity must be easily learned and mastered by many users who are non-experts in both computing and the particular beamline technique and detectors being used.
- **HPC ready software and algorithms:** Domain scientists today are routinely able to generate 10,000's to 100,000's of images in a few days of running. Using current analysis techniques, software, and resources only a small percentage of usable data are analyzed. To take full advantage of the scientific information in the data requires an environment for data analysis on HPC resources and that alleviates the need for beamline operators to manually intervene and support each users' individual analysis.

References:

1. "Advanced Data Analysis and Modeling Tools for Scattering Methods." Report of the Workshop on Advanced Data Analysis and Modeling Tools for Scattering Methods: Gaithersburg, MD (September 21, 2010).
2. R.S. Canon et al. "Defining an Ecosystem to Support Data-Intensive Science."