

## Light Source Data Analysis & Simulation: The Data Challenge

*Alexander Hexemer (ALS), Xiaoye Li (CRD), Stefano Marchesini (ALS), Dilworth Parkinson (ALS), Nobumichi Tamura (ALS), Craig E. Tull (CRD)*  
*Advanced Light Source, LBNL; Computing Research Division, LBNL*

BES facilities (Light Sources and Neutron Sources) serve ~10,000 researchers per year. Recent improvements in detector speed and light source luminosity are yielding unprecedented data rates which exceed the capabilities of most data analysis approaches utilized in past experiments. In order to unlock their tremendous potential for scientific discovery and technological advances, this new generation of facilities and detectors must be complemented with state-of-the-art networking and computing facilities<sup>1</sup> and a new generation of analytical methodologies and tools.

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.

"The development of a robust data analysis and modeling software package is absolutely critical for maximizing the impact of ALS scattering infrastructure and addressing the DOE Grand Science Challenges. 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."<sup>2</sup>

Most light source experiments are conducted by small groups of 10 or fewer researchers in concert with beamline scientists at each facility. While the aggregate size of scientific communities is comparable to that of other, large-science communities (eg. LHC experiments), BES science does not have large, structured groups that can absorb the kind of computational effort that HEP has been able to mount within LHC experiments.

## Data Themes

The problems addressed by modern day light sources are extremely diverse. The science questions range from biology, material science, physics to earthscience and archeology. While the science challenges are diverse the techniques are connected by common themes.

### Real Space

X-ray imaging beamlines have been generating TBs of raw image data per month, and with the development of new and faster cameras data rates will double or triple in the near future. As the data rate and volume have increased, the need for on-site analysis has increased. The

---

<sup>1</sup> Issues of data movement, management, and storage and of computational resources are beyond the scope of this whitepaper. We will address these issues in a separate paper.

<sup>2</sup> Advanced Data Analysis and Modeling Tools for Scattering Methods Workshop - Sept, 2010

imaging data analysis involves reconstruction of large 3D images, segmentation of the images into sub-regions, discrimination of multiphase solid materials, identification of microstructures, calculation of statistical correlation functions (e.g. surface, pore-size) and extraction of channel networks. As part of the analysis, visual representations of structures are often mandatory. For example, during modeling of structures, a theoretical optimum may not lead to the optimum problem solution, and scientific visualization tools can be indispensable in such cases. The amount of data demands both automation and re-factorization of algorithms into software that can use multiple cores, several nodes to solve a problem that we can characterize as a high performance analytics (data centric). Issues of disk I/O, efficient threading and distributed communications will play a major role on scaling algorithms to provide the necessary tools.

### **Reciprocal Space**

Just as in imaging beamline, scattering beamline have seen a revolution with respect to flux and detectors. High brilliance beam, efficient x-ray focusing optics and fast 2D area detector technology allow the collection of thousands of diffraction patterns occupying terabytes of disk space. The need for real time analysis also stems from the need of being able to modify experiment on the fly in cases the next action depends on the outcome of the previous scan. Additional computational needs for the light sources is the integration of modeling and simulation as tools accessible to the users in real time. Clever visualization tools are also needed for displaying multidimensional data in a practical way. Techniques such as coherent diffractive imaging, nanocrystallography and ptychography under development at light sources are also the fruit of advances in reconstruction techniques. In these techniques, the need for algorithms capable of solving large scale ill-conditioned, underdetermined, noisy inverse problems has never been so clear. The vast majority of data in diffractive imaging is almost never looked at. Reconstructions fail, often.

### **Spectroscopy**

The data rates for spectroscopy are traditionally not as high as the previous described techniques, but just as before it relies very heavy on usually very complicated analysis algorithms and simulations.

## **Computational Tools**

In each of these light source data themes, domain scientists are routinely able to generate 10,000's to 100,000's of images in a few days of running. Such data volumes cannot be analyzed individually, but rather must rely upon automated methods that translate Materials Science Descriptions to input for modeling and simulation, and that quantitatively compare output of simulation with beam line data. To maximize both the functionality and robustness of these kind of end-to-end analysis systems, a community-wide, open-source project must be initiated, and nurtured at the agency level (ie. DOE-BES).

These same kind of large scale, automated, and customized systems have been developed and deployed for other science communities. Although none are directly adoptable by BES light source scientists, many principles, approaches, and lessons learned are directly applicable. These tools must possess the following essential features in order for them to be widely adopted in the materials science community.

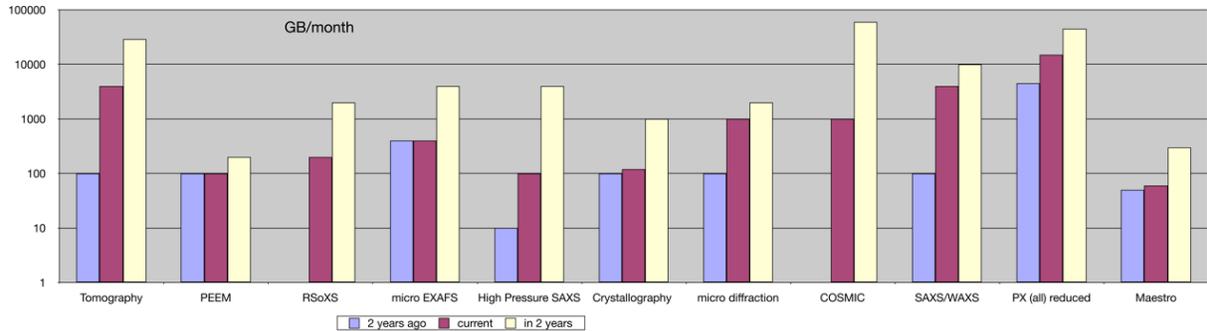
1. **Ease-of-use and extensibility.** These features both ensure widespread adoption within a scientific community and maximize the reusability of software components developed by research team by others. As an example, a graphical modeling interface similar to those used by solids modeling programs would provide researchers a natural method of describing the microscopic structure of material samples, and a common format for input to simulations.
2. **Deployment of advanced algorithms using state-of-the-art computer hardware.** In order to develop the fastest algorithms and robust codes, we need to exploit and leverage the resources provided by the ASCR office, including the expertise in Applied Math, Computer Science and the High Performance Facilities, such as NERSC. In particular, parallelization of these algorithms on multiple CPUs, graphical processor units (GPUs), and hybrid CPU/GPU multicore architectures will dramatically decrease the analysis time by more than several orders of magnitude while simultaneously permitting larger data sets to be treated.
3. **Quantitative, interactive visualization.** Visualization tools which allow quantitative comparison of simulation and experiment, including whole-image comparison or feature extraction, will aid both large-scale processing, and improve the quantity, quality, and reproducibility of scientific results.
4. **Leveraging of advanced computer technologies.** As enabling computer technologies (such as data I/O and formats, ontologies, FFT libraries, etc) and architectures (such as GPUs, multi-core, or heterogeneous architectures) evolve and improve, a common framework must allow for graceful evolution to accommodate and take advantage of the latest improvements while insulating users from the underlying details. This provides two advantages: The perturbative effects of such changes to scientists' research are minimized and the advantages are more quickly and widely available.

## Conclusion

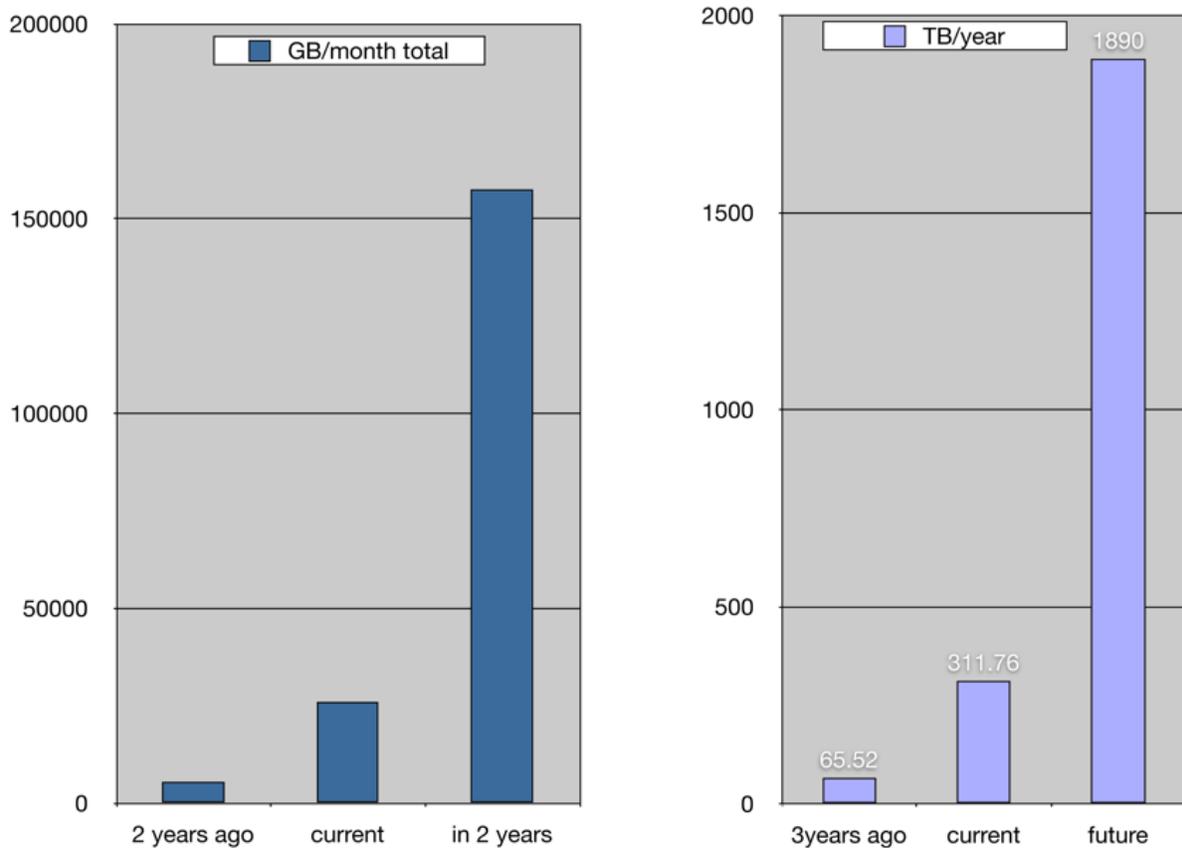
In conclusion, BES light source science has reached a defining moment of both challenge and opportunity, where recent advances in accelerators and detectors must be matched with commensurate advances in computing and software. *The challenge:* These computational advances are necessary to keep pace with the wealth of data produced by BES national facilities. Significant changes in data management, data analysis, and simulation will be required to analyze new light source data and to realize important scientific and technical advances. *The opportunity:* In responding to the challenge, rather than solve problems on a case-by-case, beamline-by-beamline, or facility-by-facility basis, a community-wide open-source solution will increase scientific output (both quantity and quality) by enabling researchers, postdocs, and students to move between beamlines, facilities, and experiments while retaining and disseminating analysis and software expertise and experience that will advance the field.

# Appendix

We show here results from a recent user survey at the Advanced Light Source (ALS). **Figure 1** summarizes the average data volume for a variety of beamlines collected during a single month. The light blue bar highlights the average data volume two years ago, red the current volume and yellow the projected volume due to beamline and detector updates in two years. Currently, the 18 ALS beamlines in the survey generate about 300 TB of data per year. We expect this to rise to almost 2 petabytes, as can be seen in **Figure 2**.



**Figure 1. Past, present, and future average data size in GB per month for ALS beamlines.**



**Figure 2. Total average data volumes for 18 ALS beamlines.**