APS Scientific Computation Seminar Series

Speaker: Ray Osborn
Senior Physicist, Materials Science Division
Argonne National Laboratory

Title: A Flexible Approach to Big Data at the APS

Date: Monday, November 23, 2015

Time: 11:00 a.m.

Location: 401/A1100

Hosts: Nicholas Schwarz and Brian Toby

Abstract:
One component of the Grand Challenge LDRD on Data Driven Science, Discovery Engines for Big Data, has been to develop flexible methods of handling the data volumes that can now be generated by the new generation of fast area detectors. Several APS beamlines have developed highly successful methods for handling high data rates in the analysis of, for example, tomography and high-energy diffraction microscopy, but these involve workflows that are specific to a particular experimental technique. We have been developing computational tools for the analysis of single crystal diffuse x-ray scattering, with the goal of prototyping a framework that can be applied to other measurement techniques at the APS and other large scale x-ray and neutron facilities. These tools are designed to be used by both instrument scientists and facility users to allow them to collect, visualize, and analyze "big data" without requiring specialized expertise, other than some basic knowledge of Python. We have now performed three experiments on Sector 11 using the Pilatus 2M detector to generate data rates of several GB/s amounting to ~30 TB in total. For example, three-dimensional measurements of S(Q) over the entire phase diagram of a compound (e.g., 6 samples x 25 temperatures) produce 5 TB of data in two or three days. Raw images in TIFF or CBF format were streamed to a remote server and automatically stacked in HDF5 files by Python scripts, which also harvested the relevant instrumental and sample metadata. By registering these files in the Globus Catalog, the data were immediately available for remote visualization and analysis, using an extensible Python GUI, NeXpy (http://nexpy.github.io/nexpy). All of the data and metadata are accessible at a granular level using Python Remote Objects, so it is possible to write simple scripts to process the results in real time during the experiment from any location, even without a fast network. We are also testing the framework on data stored at the Spallation Neutron Source and the CHESS facility, and we believe that the experience gained on this project should be of use in the design of remote data facilities. I will give a demonstration of the ways we have used this framework in our own experiments and explain how it could be deployed in other contexts.

This work is a collaboration with Stephan Rosenkranz, Matt Krogstad (MSD), Guy Jennings (APS), Justin Wozniak, Michael Wilde, Ben Blaiszik, Kyle Chard, and Ian Foster (MCS).