Big Data, Big Solutions

X-ray Science Division

Doga Gürsoy[1], Francesco De Carlo[1], Youssef Nashed[3], David Vine[1], Stefan Vogt[1], Suresh Narayanan[1], Vincent De Andrade[1], Sophie-Charlotte Gleber[1], Faisal Khan[2], Arthur Glowacki[2], Nicholas Schwarz[2], Zichao Di[3], Sven Leyffer[3], Stefan Wild[3], Rachana Ananthakrishnan[3], Ian Foster[3], Tom Peterka[3], Young Pyo Hong[4], Rachel Mak[4], Yue Sun[4], Junjing Deng[4], and Chris Jacobsen[1,4]

APS X-ray Science Division[1], APS Engineering Support Division[2] and Math and Computer Science Division[3], Argonne National Laboratory; Dept. Physics & Astronomy[4], Northwestern University

Introduction

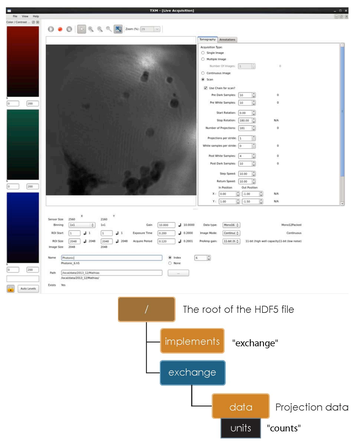
Analysis of large datasets at synchrotron light sources is becoming progressively more challenging due to the increasing data acquisition rates that new technologies in X-ray sources and detectors enable. The next generation of synchrotron facilities that are currently under design will provide diffraction limited X-ray sources and is expected to boost the current data rates by several orders of magnitude stressing the need for the development and integration of efficient analysis tools more than ever.

To continue to fully exploit the rich APS data content and to enable the APS to continue to be on the forefront of science and engineering research, we are developing efficient data management systems, including a Data Catalog integrated with Globus Online for fast and reliable data access, Data Exchange for provenance and data tracking, and tomoPy providing a collaborative framework for the analysis of data intensive synchrotron techniques. We are also developing software to automatically uncover hard-to-find patterns in large image datasets, to rapidly reconstruct images from complicated coherent diffraction data, and to more efficiently close the loop between simulation, materials synthesis, and characterization.

Data Acquisition

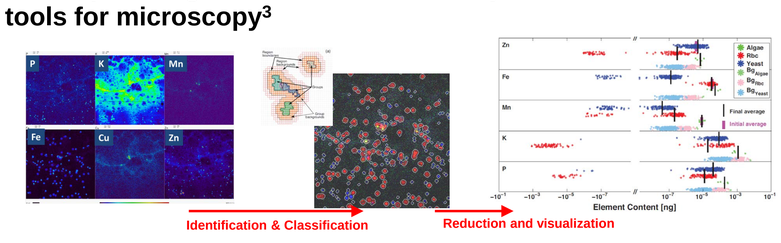
A new experiment control user interface to provide multi-scale nano and micro tomography data integration[1]

The software, written in C++ using Qt, interfaces with EPICS for beamline control and provides live and offline data viewing, basic image manipulation features, and scan sequencing that coordinates EPICS-enabled apparatus. Post acquisition, the software triggers a workflow pipeline, written using ActiveMQ, that transfers data from the detector computer to an analysis computer, and launches a reconstruction process. Experiment metadata and provenance information is stored along with raw and analyzed data in a single HDF5 Data Exchange file.



Visualization and Mining

Develop new generation of data analysis and visualization tools for microscopy[3]

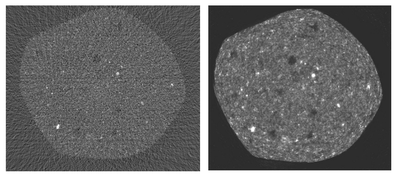


Left: X-ray fluorescence maps of 6 different elements of a sample mixed of 3 different cell types. Center: The software automatically identifies and classifies 3 different cell types, enabling further analysis, taking background around cells into account, and subdividing the sample into independent regions for parallelization – note even overlapping areas can be identified and distinguished. Right: comparison of the extracted average elemental content per individual cell.

Iterative reconstruction methods for incomplete data

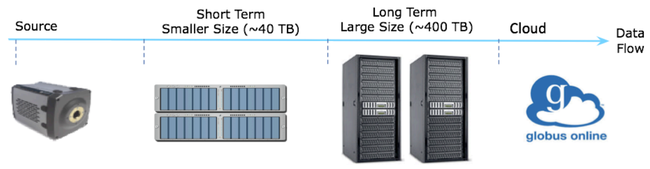
Develop model-based iterative reconstruction methods for dose reduction and fast scanning

Left: Micro-CT reconstructions with 46 projections using direct Fourier method (Gridrec). Right: reconstructions obtained using Maximum Likelihood Expectation Maximization (MLEM) method.

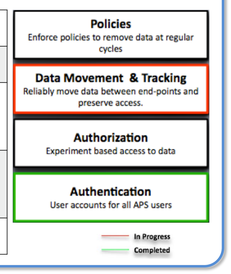


Data Movement and Storage

Automation in data storing, access, archival and distribution

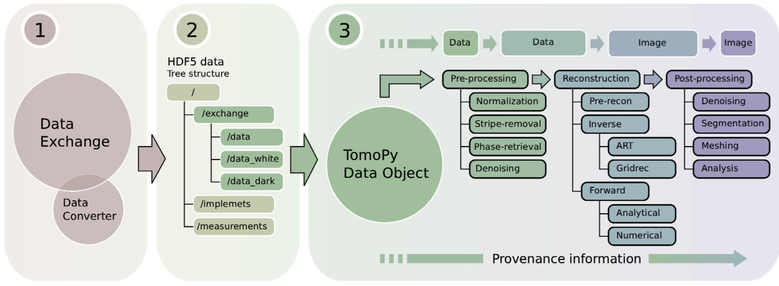


Authentication	Login based on APS user account
Authorization	Access via users on Safety Approval
Data Movement	Can we avoid running 'cp' on every file that gets generated?
Policies	Running low on space. Which dataset can be removed?
Tracking	Where is my data now?



Integration of data analysis tools

TomoPy: a Python/C++ framework for the analysis of synchrotron tomographic data[2]

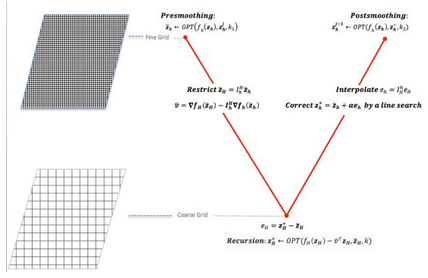


The basic principles of this Python-based open-source framework include ease of collaborative development of scripts, platform and data format independence, modularity.

Numerical Optimization

New mathematical models are under development to integrate X-ray transmission and X-ray Fluorescence data[4]

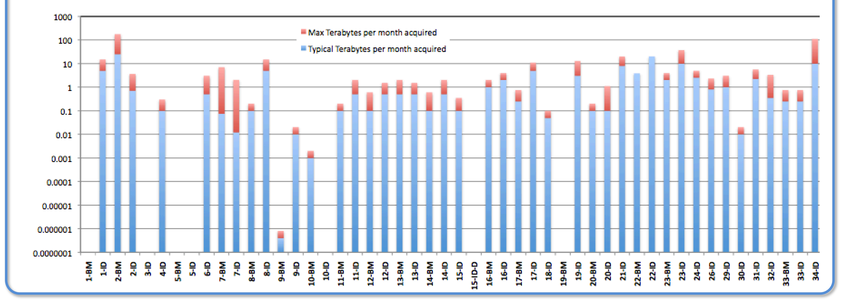
Multi-grid optimization approach (MG/OPT) solves large nonlinear optimization problems using computations on coarser levels to accelerate the progress of the optimization on the finest level.



Current cumulative data volume at the APS

Typical: 100 TB/month
Maximum: 370 TB/month

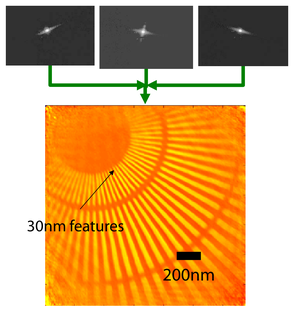
(as of summer 2013)



Real-time Ptychography Analysis

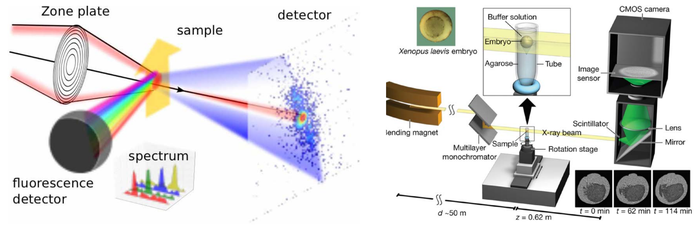
Ptychographic reconstruction of a nano-structured gold Siemens star from data acquired at the 21-ID Bionanoprobe

A newly developed parallel ptychography is able to achieve a 220-fold decrease in the analysis time using a single Graphics Processing Unit (GPU). Further gains can be expected in the future as more GPU nodes are utilized. The spatial resolution in this image of below 10 nm is far beyond what can currently be achieved using focusing optics at this X-ray energy.



Multi-resolution, multi-modal data fusion

Develop data fusion methods to integrate micro, nano and fluorescence tomographic datasets



References (see also `tinyurl.com/n658ssa`)

[1] N. Schwarz et al., Experiment Control and Analysis for High-Resolution Tomography, In Proceedings of ICALEPCS 2013

[2] TomoPy: http://www.aps.anl.gov/tomopy, Data Exchange: http://www.aps.anl.gov/DataExchange/

[3] S. Wang et al, J Synchrotron Radiation, (2013) accepted

[4] A. Borzì et al., Multigrid Methods for PDE Optimization, SIAM Review (2009) 51:2, 361-395

Funding: "Tao of Fusion" LDRD, BES-ASCR postdoc, ASCR ROMPR.

The Advanced Photon Source is funded by the U.S. Department of Energy Office of Science