Addressing the Data Challenges at the Advanced Light Source

Alexander Hexemer
Senior Staff Scientist
Program Lead for Computing
Advanced Light Source
Center for Advanced Mathematics for Energy Research Applications (CAMERA)
Lawrence Berkeley National Lab
We can’t do it alone!!!
Acknowledgment

Lawrence Berkeley National Lab
- Ronald J Pandolfi
- Dinesh Kumar
- Dylan McReynolds
- Singanallur Venkatakrisnan (now ORNL)
- Luis Barrosso-Luque
- Holden Parks
- Austin Blair
- Dilworth Parkinson
- Shuai Liu
- Nathan Melton
- Andrew Wiedlea
- Debbie Bard
- Hari Krishnan
- Krishna Muriki
- Dani Ushizima
- James A Sethian

Argonne National Lab
- Zhang Jiang
- Doga Gursoy
- Francesco De Carlo
- Xianghui Xiao
- Ian Foster
- Nicholas Schwarz
- Ryan Chard

SLAC
- Amanda Fournier
- Fang Ren
- Yury Kolotovsky
- Apurva Mehta
- Chris Tassone
- Amedeo Perazzo

Brookhaven National Lab
- Masafumi Fukuto
- Kevin Yaeger
- Thomas Caswell
- Stuart Campbell

Funding Acknowledgments
- LBNL LDRD Program
- DOE Early Career Award
- CAMERA
Center for Advanced Mathematics for Energy Research Applications

Include latest theory and math, take advantage of latest architecture: Multi CPU/GPU, open source, everything shared, many collaborations

www.camera.lbl.gov
Challenges for User facilities

• >20% new users
• Provide very fast feedback and/or experiment combining data from different modalities
• Applying custom workflow to many data sets
• Need for new math and algorithms
• Make things easy and faster for users
• Data access across facilities
• Large amounts of data
New Sources are on the Horizon

- **Advanced Light Source (ALS)**
- **New Sources**
- **Photon Energy [eV]**: 10^1 to 10^5
- **Brightness [photons/sec/mm^2/mrad^2]**: 10^16 to 10^23
- **Coherent Flux [photons/sec/0.1% BW]**: 10^13 to 10^16
- **Soft x-rays**
The Light Source Data Working Group

The purpose of the DWG is to serve as a resource for the 5 BES light source Facility Directors to call upon to provide information and recommendations on working together in the areas of data and computing.

- Development of the standards for data storage and file formats
- Create reports and outline ways for the light sources to work together.
- Explore the topic of real time computing at the 5 light sources and into the future.
- Data access across facilities
- Regular meetings online and in person.

ALS (Alexander Hexemer), APS (Nicholas Schwarz) LCLS II (Amedeo Perazzo), NSLS II (Stuart Campbell)
Development of the Data Movement in Collaboration with other Divisions of LBNL

- **NERSC** (Debbie Bard et al.): HPC API specs are currently developed, GPGPUs

- **Lab science IT** (Andrew Wiedlea et al.): GPGPUs, data storage, infrastructure, Mongo and many other ideas

- **ESNet** (Kate Mace et al.): Network challenges

- **Data Acquisition Group** (Kevan Anderson et al.): Data and Metadata access (NEXUS), development of a python/scripting interface to allow for analysis driven control of the beamline

- **CAMERA** (Jamie Sethian et al.): Analysis, Xi-CAM, cam link, HPC API, sharp and many more
GIWAXS scattering on printed OPVs
Full data lifecycle: Xi-CAM-Link [Data]

Execution steps
1. Identify and setup resources
2. Launch services
3. Connect network
4. Execute graph

Remote Storage/Tape Archive
- Compute
- Publishing
- Archiving

Edge Services –
DataBroker: Event Based
- Metadata
- Provenance
- Data Tracking

MongoDB
- Workflow State
- Data Cataloging
- Data Tagging
- Logging

CAM-Mart - Algorithm Marketplace
Full data lifecycle: Xi-CAM-Link [Analysis]

Execution steps
1. Identify and setup resources
2. Launch services
3. Connect network
4. Execute graph

- Compression
- Preprocessing
- Analysis close to detector

Python-based
- Remote Workers
- Support Docker/Singularity (WF Provenance)

Graphical User Interface (GUI)

User / local computer

Compute cluster
Deployment of Mathematical Algorithms: Xi-cam

Result of DOE Early Career Award

Scientific Achievement
Development of a community-maintainable platform for new analysis and visualization techniques for synchrotrons.

Research Details
– Remote processing with HPC for real-time high data rate analysis
– Remote data access for high-volume data retrieval
– Highly interactive design

Xi-CAM plug-ins

<table>
<thead>
<tr>
<th>Tomography</th>
<th>Remote Computing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time resolved SAXS</td>
<td>GISAXS</td>
</tr>
<tr>
<td>Electron Microscopy</td>
<td>HIPGISAXS</td>
</tr>
<tr>
<td>NEXAFS</td>
<td>GIWAXS simulator</td>
</tr>
<tr>
<td>iPython</td>
<td>Reverse Monte Carlo</td>
</tr>
<tr>
<td>Globus Online</td>
<td>CD-GISAXS</td>
</tr>
<tr>
<td>Databroker</td>
<td>Plug-in Store</td>
</tr>
</tbody>
</table>

Xi-cam: A versatile interface for data visualization and analysis

Scientific Aims
Efficient platform for rapid parallel screening and materials discovery.
Data-driven feedback directly drives material synthesis in-situ.

Significance and Impact
Prototype system at GIWAXS/XRD beamline at SSRL for development.
Material properties and characterizations computed on-the-fly; ~1 second per sample.

http://pubs.acs.org/doi/pdf/10.1021/acscobmsci.7b00015
Remote Execution

GISAXS Workshop in Bayreuth in collaboration with CAMERA

File Plugins Testing Active Session (localhost)
Batch | HipGISAXS | IPython | Log | Timeline | Tomography | Viewer | 3D Viewer

Welcome to Xi-cam

Please cite Xi-cam in published work:
Pandolfi, R., Kumar, D., Venkatakrishnan, S., Hexemer, A. (under preparation)
The High-throughput NEXAFS workflow collaboration with Materials Project

Materials Project website

Processing/reduction

Quantities

soooon

Results

Analysis

Phase diagrams

MPContribs Framework

Public

bl 6.3.1
Deep learning for X-ray Scattering

Example 1: Classify nanostructures by packing
Input: 2D image of crystal diffraction (cubic, hexagonal ... packing)

Example 2: Classify between SAXS/GISAXS data
Automatize the data classification on beamlines (7.3.3 beamline)

Example 3: Using Fourier transform to improve learning
Truncate frequencies in the Fourier space to retain more information

Generate a model which predict with 80% accuracy the different space groups
Collaboration K. Yager and M. Fukuto (NSLS2)

<table>
<thead>
<tr>
<th>Method</th>
<th>Original Data Set</th>
<th>Noisy Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convolutional Neural Networks</td>
<td>92.3%</td>
<td>91.6%</td>
</tr>
<tr>
<td>Histogram of Oriented Gradients</td>
<td>92%</td>
<td>80%</td>
</tr>
</tbody>
</table>

Queries
Best matches

Reduce image size for classification in Fourier space to limit the loss of information

Rippel et al. (2015)

Use Case in Development:
X-ray Photon Correlation Spectroscopy (XPCS)

New Beamline development

Interface: Xi-CAM

Control:
Bluesky and DataBroker

HPC Code for XPCS

Future of Creative Discovery

Next Generation of Creative Tools for Analysis
- Visual and Interactive
- Multi modal
- Physics and Math engine by design
- Full integration of ML

Rick Perry Explores Tomography Data Spring 2018

KAREN ROSS Secretary of the California Department of Food and Agriculture
Thanks
"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

**Mailing Address:** NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

**Physical Address:** 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: [https://www.nitrd.gov](https://www.nitrd.gov)