LCLS Data Analysis Strategy

February 24th 2016

Table of Contents

1 INTRODUCTION.......................................................................................................................2

2 CORE APPROACH..................................................................................................................3

   2.1 DRIVERS..........................................................................................................................3

   2.2 GUIDELINES......................................................................................................................3

   2.3 THE NEED FOR ADVANCED ALGORITHMS.....................................................................6

3 IMPLEMENTATION..................................................................................................................7

   3.1 INFRASTRUCTURE CHALLENGES....................................................................................7

   3.2 INFRASTRUCTURE PROJECTS........................................................................................8

   3.3 TOOLS................................................................................................................................13

   3.4 TOOLS INITIATIVES........................................................................................................13

   3.5 ALGORITHMS..................................................................................................................15
1 Introduction

The LCLS facility has been in operation since 2009 and has had a profound impact on a broad cross-section of scientific fields. Advanced computing systems are playing an increasingly important role in facility operation, data interpretation, and overall scientific productivity. Demands for computing at LCLS are driven by the properties of the LCLS source itself, advances in detector technology, advances in data analysis algorithms, and the need to provide fast feedback to users in real-time while balancing flexibility and ease-of-use.

The intrinsic pulsed nature of the FEL source requires experimental solutions that acknowledge that every shot is different and that a broad suite of information needs to be recorded to interpret a single-shot event. Further, LCLS can provide ultrashort pulses, which introduce the challenge of recording and indexing all scattered photons that emerge from a sub-100 fs pulse. Imaging detectors developed for the FEL require careful signal correction (cross-talk, nonlinearity) in order to correctly interpret the science data. These requirements have resulted in the current LCLS data acquisition, data management, and data analysis infrastructure.

To maintain a world-leading capability for advanced research in energy, materials, biology, and chemistry, the US Department of Energy recently initiated the LCLS-II project. This upgrade, due to come online at around the turn of the decade, increases the repetition rate from 120 pulses per second to ~1 million per second. LCLS-II will also provide for a factor of ~2 increase in experimental capacity. To take full advantage of this investment, upgrades in the areas of data acquisition, data management, and data analysis are required.

The LCLS data analysis strategy describes how the computing needs of LCLS and LCLS-II will be achieved and informs the wider facility requirements [1].
2 Core Approach

2.1 Drivers

The LCLS data analysis strategy is motivated by the following aspects:

1. **LCLS-II Upgrade**: The high repetition rate (1-MHz) and, above all, the potentially very high data throughput (100GB/s) generated by LCLS-II will require a major upgrade of the data acquisition system and increased data processing capabilities [2], [3].

2. **Fast feedback**: Experience has shown that a capable real-time analysis is critical to the users’ ability to make informed decisions during an LCLS experiment. Powerful fast feedback (~minute or faster timescales) capabilities reduce the time required to complete the experiment, improve the overall quality of the data, and increase the success rate of the experiments.

3. **Time to science**: Sophisticated analysis frameworks can significantly reduce the time between experiment and publication, improving the overall productivity of the LCLS science community.

4. **No user left behind**: Most of the advanced algorithms for analysis of the LCLS science data have been developed by external groups with enough resources to dedicate to a leading edge computing effort. Smaller groups with good ideas may be hindered in their ability to conduct science by not having access to these advanced algorithms or forced into collaborations with larger groups. LCLS support for externally developed algorithms and the development of in-house algorithms for some specific science domains, would alleviate this problem.

This strategy will be organized into a three-pronged approach aligned with the following areas:

1. **Infrastructure**: This area includes the systems for processing and managing the LCLS data: computing farm, disk and tape storage, network, data movers, experiment portal.

2. **Tools**: This area includes all the core software needed by the LCLS users to access and analyse their data: build tools, documentation tools, version control, visualization, calibration, data persistency, and basic data analysis algorithms like fitting and filtering.

3. **Algorithms**: This area includes the adoption, development and support of advanced algorithms specific to the various LCLS scientific areas: hit finding, indexing, diffraction patterns integration, macromolecular structure determination, etc.

Where possible, LCLS will pursue partnerships with other facilities and make every effort to leverage the work and expertise of the community to solve common problems. LCLS personnel will encourage the collaboration between software engineers, instrument scientists and users to ensure that code and algorithms are shared. Open source tools will be used where feasible, and shared repositories will be used to store and distribute code to users. LCLS personnel will work closely with the users and scientific staff to provide resources to install, understand, and use the analysis software. By participating in the development of these algorithms and actively communicating with users, LCLS staff can ensure that these algorithms are portable and reusable around the world.

2.2 Guidelines

LCLS recognizes the central role advanced algorithms and computing have in x-ray science and facility operations. The facility, however, does not have adequate resources to maintain a fully staffed in-house research team. Our strategy will be to maintain a core set of personnel who will perform three functions:
1. Remain aware – at an expert level – of algorithmic advances that could expand the scientific capabilities of the facility and work to ensure that these advances are available in usable software to the entire LCLS user base.

2. Work with users and scientific staff to understand problems facing the user community that could be addressed by computational advances, and communicate these to the broader computational community, for example academic computer science departments.

3. Build and maintain relationships with outside research organizations in computational science whose interests and expertise overlap with the computational challenges facing LCLS.

This strategy aims to ensure that first, novel algorithms and software that enable new science are distributed as quickly and broadly to the user community as possible, and second that academically interesting computing problems facing LCLS are rapidly communicated to research communities with the expertise and resources to develop solutions to those problems. A simplified flow of these interactions is shown in Figure 1.

We realize that some software is lab-specific while other software can be reused across facilities. It’s critical to identify areas where we can reuse standard tools across facilities, as well as contributing code to those efforts. This reuse will not only reduce the required effort level, but it will help the field because scientists will be familiar with the tools when experiments are performed at different labs. We believe that the reuse of low level tools is most likely to be effective, since higher level frameworks tend to be lab-specific (see Figure 2).

A key driver in the prioritization of LCLS data analysis improvement tasks described later on in the implementation part of this document, was a 2015 survey where we polled the LCLS users about their view of the impact of various data related projects. A summary of this survey is shown in Figure 3 and the results analytics are presented in reference [4].
**Figure 1:** Outline of the core approach for LCLS engaging algorithmic experts to support user science.

**Figure 2:** Software that is reusable (lab independent) and not reusable (lab specific).
2.3 The Need for Advanced Algorithms

Historically, algorithmic and mathematical advances have enabled new x-ray science: phase retrieval, computed tomography (slice theorem), and auto-indexing are canonical examples. Each enables a new experimental mode — coherent lensless imaging, 3D x-ray microscopy, and serial crystallography — that would not have been possible or even conceivable before algorithmically demonstrated. We expect further algorithmic advances will enable new science at the LCLS, and we will position the facility to maximize the yield from these advances.

Additionally, we will work to develop advanced algorithms for operational improvement, including data triage, reduction, and fast feedback as we look towards the LCLS-II upgrade. LCLS-II will result in a data rate increase outpacing standard Moore’s law hardware improvements, meaning it will be increasingly expensive to rely on hardware alone to handle the increased data rate. Streaming algorithms, for instance, capable of providing reliable vetoes or pulling data of particular interest for fast feedback, could significantly reduce the data rate burden. We will aim to prototype these algorithms now on LCLS, where they can impact current operations in a positive but modest manner, for use when LCLS-II sees first light where they may be an essential component of data collection and analysis.

The academic scientific model provides strong incentives for the development of novel algorithms, but not for the wide distribution of technology to other academic groups. Distribution often involves many tasks not typically considered in the scientific domain: writing production-quality software, documentation, tests, and tutorials. In order to maximize the impact of algorithmic advances on LCLS experiments, we will commit resources to translating published computational research into software that the entire LCLS user base can employ. We believe facility-provided and maintained codes implementing the latest algorithmic advances will result in a reduced lag between experiment and publication and ensure no user is at a disadvantage due to limited computational expertise or resources.
3 Implementation

This part of the document describes how we plan to implement the data analysis strategy introduced in the first part of the document. Each area is split into two sections: the first one describes the challenges we identified in that area, the second one describes the projects and initiatives planned to tackle those challenges.

3.1 Infrastructure Challenges

The main infrastructure challenge will be developing high-throughput, high density, peta-scale storage systems that allow concurrent access from thousands of jobs. Another critical feature is the deployment of a trigger/veto system to veto the readout process for uninteresting events thus reducing the data throughput. Additional critical capabilities include upgrading the SLAC network connection to ESNet and expanding bandwidth and capacity of the tape archive. Specific challenges are noted here, and targeted projects to address these are outlined in the following section.

- **Data Acquisition (DAQ):** Two main modifications to the current system will be required for operating at high repetition rates: moving the event builder from online to offline and developing the ability to aggregate contributions from multiple events in the readout nodes. These changes are required for running at 1-kHz pulse rate or above, independent of throughput. The deployment of a trigger/veto system for LCLS-II will be required for large area detectors if reading out images at full rate turns out not to be feasible. Changes dictated by the increase in throughput are a network upgrade (from 10 Gb Ethernet to Infiniband or to 40 Gb Ethernet) and the online cache upgrade.

- **Real Time Data Access:** The LCLS experience has shown that the most effective way to perform real time analysis is allowing users to run their code against the data on disk (fast feedback storage layer). Fast feedback will become even more important with the deployment of a trigger/veto system for LCLS-II. The existing storage technologies are too slow for the LCLS-II fast feedback layer. Spindle based systems will become cheaper and more dense, but not much faster or easier to manage, and they do not handle concurrency well. Solid state based systems will also become cheaper, but the current trend for commercial systems is to optimize IOPS (input/output operations per second) versus throughput and scalability, the key aspects for a system hosting the LCLS-II science data. In addition, current commercial systems come with a significant premium on the the cost of the flash memory, making a multi-petabyte system prohibitively expensive.

- **Data Storage:** The SLAC tape archive system is approaching limits in overall storage capacity (~20+PB) and throughput. Such limits are already observed at LCLS when archiving data from on-going experiments while serving concurrent user requests to restore files from tape. Based on the current storage requirements and the estimated increase in the amount of acquired data, it is expected that LCLS-II will require around 100 PB of fast storage (see Table 1). Deploying and maintaining these levels of storage at SLAC would require a significant increase in the capabilities of the existing LCLS and/or SLAC IT groups. A more cost-effective solution would be to offload part of the LCLS-II data storage to larger computing facilities like NERSC.

- **Data Management:** LCLS has developed a powerful data management system that handles both the automatic workflows of the data through the various storage layers (e.g. long term data archival) and the users’ requests through a web portal (e.g. restoring data from tape). Some
aspects of the current system, such as checksum calculations, HPSS interface, and lack of prioritization, will become limitations at higher data volumes and will need to be upgraded.

- **Data Processing**: Based on the current computing requirements and the estimated increase in the amount of data to process, it is expected that LCLS-II will require between 200 TFLOPS and 1 PFLOPS. As with data-storage, deploying and maintaining very large processing capacity at SLAC would require a significant increase in the capabilities of the existing LCLS and/or SLAC IT groups. A more effective solution would be offloading part of the LCLS-II data processing to larger computing facilities like NERSC.

- **Data Network**: SLAC recently upgraded its connection to ESNET from 10 Gb/s to 100 Gb/s. The primary reason for upgrading this link is to gain the ability to offload part of the LCLS science data processing to NERSC, as current LCLS data acquisition rates are up to 5 times a single 10 Gb/s link. The 100 Gb/s link will not be enough for LCLS-II, and terabit capabilities will be required if LCLS relies on NERSC for processing LCLS data. In regard to the local network, Infiniband would be superior to Ethernet for building a high throughput network for LCLS-II, especially under high congestion conditions. However, Infiniband has cost-consequences in non-localized installations: it is more expensive than Ethernet to connect devices that are not within 10m of each other, and significantly more expensive to connect devices that are more than 300m apart.

- **Data Format**: The LCLS DAQ is currently writing the raw data in XTC format. Users can request that their data be translated to HDF5. The translation step will become a bottleneck in the future and LCLS-II should adopt a single data format. HDF5 is becoming the de-facto standard for storing science data at light source facilities, but in order to effectively replace XTC in LCLS, a couple of critical features are required. These features, namely the ability to read while writing and the ability to consolidate multiple writers into a consistent virtual data set, are currently missing in HDF5. However, the HDF group claims these features could be added if enough resources are made available.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>LCLS-I</th>
<th>LCLS-II 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average throughput</td>
<td>0.1 - 1 GB/s</td>
<td>2 - 20 GB/s</td>
</tr>
<tr>
<td>Peak throughput</td>
<td>5 GB/s</td>
<td>100 GB/s</td>
</tr>
<tr>
<td>Disk storage</td>
<td>5 PB</td>
<td>100 PB</td>
</tr>
</tbody>
</table>

*Table 1:* Data requirements scaling between LCLS I and II. The disk storage numbers assume the adoption for LCLS-II of the current data retention policy (6 months unlimited, 2 years with quota and 10 years on tape). Note that not all contributions to the science data will increase linearly with the pulse rate, eg some of the diagnostic data, so the overall throughput increase will be less than the increase in the pulse rate.

### 3.2 Infrastructure Projects

We currently envision an evolution of the LCLS data system where the fast feedback storage layer is built on flash memory and where the offline processing and storage capabilities can be offloaded to multiple facilities. NERSC is ideally suited to becoming one of these facilities. See Figures 3 through 6.
This is a summary of the critical projects required to build a data system able to handle the LCLS-II requirements:

- **Online Data Reduction**: This includes both the ability to generate a trigger signal for good events (i.e. a veto system) and the ability to extract the key features from the data (lossy compression). A veto signal could be delivered to the front-end electronics (EuXFEL approach), to the readout nodes, to the online cache or in the fast feedback layer. In general, a veto in the front-end electronics reduces the throughput requirements on the DAQ components, while a veto in the following layers provides cheaper/larger buffers and more time to reach a decision.

- **Flash Storage**: Develop a custom, solid state based, online cache (DAQ recorders) and fast feedback storage layer (users data analysis) to solve the storage challenge. SLAC has previously worked on peta-scale flash-based systems and determined it is possible to build a scalable, peta-scale, solid state storage by aggregating commercial off-the-shelf components. The same technology could be used to build custom recorders for both the online cache and the fast feedback system.

- **Upgrade local area network**: The current system uses 10Gb Ethernet from the readout nodes to the online cache and to the fast feedback and IB from the fast feedback to offline and within the offline nodes and the storage. SLAC will investigate introducing IB or 40Gb Ethernet from the readout nodes to the online cache and to the fast feedback. The final solution will be based on actual space constraints.

- **Add data management capabilities at high throughput**: This includes upgrading the SLAC tape system, since we already see limitations when handling data from on-going experiments and concurrent users requests to restore files from tape, and upgrading the data management framework.

- **Deploy new timing system**: The existing EVG/EVR system is not scalable to MHz operations, therefore a new system will be required. The new LCLS-II timing system is a frame based fiber optic distribution system delivering synchronous control information at the maximum beam rate of 929kHz. Sections of the control frame are dedicated to experiment use for coordinating control of endstation devices with the delivery of beam. In addition, the frame delivery may be re-mastered in the distribution tree to allow peripheral data acquisition systems to make central trigger decisions in their branch of the distribution tree. The fiber optic distribution also allows for feedback information to flow upstream allowing those data acquisition trigger masters to (1) collect endpoint readout status for asserting controlled deadtime, (2) tag a subsample of events for special handling or analysis, and (3) receive fast event feedback information for making higher-level veto decisions on triggered events, which would allow the quick rejection of events not needing the slower detectors’ readout.

- **HDF5 upgrade**: Some critical features are missing from the HDF5 API: ability to read while writing and the ability to consolidate multiple writers into a consistent virtual data set. The HDF group thinks these features are useful in general, not just for LCLS, and that they could, and should, be added to the API. The group needs additional resources to work on these features.

- **Increase data processing capabilities**: Data centers built towards data intensive systems could help offload the LCLS/SLAC offline computing system. General support for LCLS-II offline analysis would require > 100 PB tape storage, a dedicated 100 PB of disk storage and a processing farm in the 0.2-1 PFLOPS range with an aggregate throughput to the storage > 10 GB/s per PB. These capabilities could be achieved by using dedicated resources to extend one of the large NERSC machines (CORI for LCLS-I and the next generation supercomputer for
LCLS-II). Other key requirements are the ability for LCLS users to manage their data through the LCLS tools and workflows, the ability to access accelerators-based computing (GPUs and Xeon Phi), and, ideally, the ability to use their SLAC user-account (or a federated account).

➢ **ESNet link upgrade**: The ability to offload computing capabilities relies on a faster connection between SLAC and ESNet (100Gb/s for LCLS and >1Tb/s for LCLS-II).

Figure 3: Current LCLS Dataflow. The real time monitoring is accomplished via a network based framework (AMI) which observes the multicast packets between the readout nodes and the recorder nodes (Online Monitoring Nodes). The data cache layer and the fast feedback layer are separate to minimize backpressure from the users activities (Users Fast Feedback processing Nodes).
Figure 4: Evolution of the LCLS Dataflow. The two online storage layers are merged into one system and so are the Online Monitoring and the Users Fast Feedback: this approach is made possible by the adoption of flash based technologies for the online storage.
Figure 5: Current LCLS Data Systems Architecture.

Figure 6: Evolution of the LCLS Data Systems Architecture: the data management system will allow to transparently integrate external computing facilities like NERSC.
3.3 Tools

The main elements of an LCLS data analysis are noted here, and targeted initiatives to address these aspects are outlined in the following section.

- **Setup**: Users should be able to setup the analysis environment in a very short time and even with limited UNIX expertise. Ideally, the analysis environment doesn’t create conflicts with other setups on the same machine, e.g. with the controls or DAQ environment, and can be ported to non-LCLS computing systems. Users should be allowed to contribute back to the community if they decide to do so.
  
  Another key aspect of the analysis setup is the documentation, which should be at the same time simple and complete.

- **Data Access**: This aspect covers the movement of data from the persistent layer to memory and the interface to the data in memory. The data access layer must allow for high throughput and for a simple, intuitive and complete API for browsing, with minimal computational expertise or learning time required of user groups, who turn over frequently.

- **Data Reduction**: Users often split the analysis in multiple steps, each step processing the output of the previous step. This aspects includes the algorithms needed to extract features from the data and the ability to move data across steps, either in memory or through a persistent layer.

- **Visualization**: Displaying the data is a key feature of any real-time or interactive analysis. Data can be displayed locally, i.e. by running the analysis on the local machine, or remotely, i.e. by forwarding the screen to the local workstation.

3.4 Tools Initiatives

These are the main initiatives for improving the LCLS analysis tools and for reducing the effort required to maintain these tools:

- **Simplify development/release tools and environments across all systems**: The LCLS analysis environment has become much more python-oriented, and there would be significant gains if the software release system (and perhaps build system) took advantage of that. Also, currently each system (DAQ/Controls/Analysis) has its own build mechanism, its own code repository, its own set of external packages and its own runtime environment. This initiative would investigate the possibility of unifying tools across the controls, data acquisition and data analysis systems as much as possible. This would not only reduce the effort required to maintain these tools, but also simplify the environment for the users who wouldn’t need to worry, for example, about running two different versions of python when operating the DAQ or when analyzing the data offline.

- **Documentation and training**: LCLS users have diversified backgrounds and come with a wide range of computing skills. We must provide documentation which can help all users in being productive with minimal training. This was rated as a high priority in the users survey. This effort will determine a coherent approach to analysis documentation and training and it will define how users can interact with the core analysis team to maximize the productivity of the experiments which require a sophisticated analysis of their data.

- **Data format**: LCLS currently supports two custom data formats: XTC (eXtended Tagged Container, an object oriented data format inherited from the BaBar high energy physics experiment) and an HDF5 based format supporting variable length arrays and compound types. This effort will investigate the adoption of an HDF5 data format to replace both XTC and the existing HDF5 structure. This format could be optimized for some of the key features required
LCLS Data Analysis Strategy

Implementation

by LCLS users: small data selection, virtual data set, real-time analysis. This effort will also investigate the feasibility of adopting existing HDF5 schemas like Nexus: providing the users a familiar data format is one of the most important steps in simplifying the data analysis activities of the groups who come to LCLS.

➢ Event building: The association of data from different detectors belonging to the same LCLS shot (“event building”) is currently done online by the DAQ system. There is common agreement that for the higher data rates/volumes of LCLS-II this will have to be done by the offline system (psana). We will investigate different options for this software (e.g. infiniband RDMA, or NERSC burst buffer based on Cray’s DataWarp technology). The online monitoring framework (currently AMI) will need to be adapted to the new paradigm.

➢ Intermediate data persistency: It could be beneficial to provide a standard mechanism for saving reduced-size intermediate data generated by each step in the analysis. These data can be processed detector data or higher level objects like histograms and ntuples. This effort would choose a technology for making these data persistent (e.g. HDF5), a definition of how the data would look like and an API for storing and retrieving these objects.

➢ Calibrations: This effort would optimize the processing of the most common LCLS detectors used for diagnostics and/or science analysis (e.g. CSPAD, XTCAV, timing tool, phase cavity, monochromators) in order to provide calibrated, reliable, high level information so that users don’t have to worry about interpreting the raw data. It would include better support for detector geometry, non-linear corrections, photon counting. This work was rated as being very important in the 2015 user survey.

➢ Portable, reusable, user-friendly software: This effort will focus on identifying areas where we can reuse standard tools across facilities, as well as contributing code to those efforts. Since we believe that python is becoming the main standard for light-source analysis, we will initially focus on reusable python software. With the frequently changing LCLS experimenters it is important that this software be as simple as possible, while still being able to express a wide variety of complexity. This was rated as a high priority by the users in the 2015 survey.

➢ Visualization: Data can be displayed locally, i.e. by running the analysis on the local machine, or remotely, i.e. by forwarding the screen to the local workstation via some standard X11 protocol (ssh, VPN, NX, etc), via a messaging library (see, for example, http://zeromq.org, http://redis.io) or via web (e.g. iPython notebook, http://www.lightning-viz.org). This effort will investigate which visualization technologies should be adopted and supported for LCLS analysis.

➢ Real-time monitoring: The Analysis Monitoring Interface (“AMI”) is currently the primary tool for real-time analysis in LCLS. It’s part of the DAQ system and it works by observing the multicast packets between the readout nodes and the recorder nodes. It is a very powerful tool: it provides a GUI where users can add some basic capabilities. It also allows users to introduce more advanced analysis by providing a shared library with a well defined interface, however this approach is not widespread because of the required learning curve and short available user-development times can lead to software instabilities. Also it is not integrated with the other analysis tools. This effort will investigate the possibility of integrating the core AMI capabilities in the general analysis framework. The use of python in the analysis framework could allow for easier customization of the online monitoring and the expression of more powerful ideas.

➢ High performance analysis: We believe that python is the becoming the standard language for light-source analysis. We will work to ensure the analysis tools perform well when scaled to the large number of CPU cores that be used for LCLS-II. Also, currently when there are multiple
analyses to be run on LCLS data, the data is typically accessed multiple times. This may be prohibitive at LCLS-II, especially for real-time monitoring. We will investigate techniques for bringing multiple analyses into the same executable to avoid multiple fetches of the same data.

➢ Hutch standard configuration support: To minimize the effort level associated with frequently changing experiments, most hutches have adopted a “standard configuration” that will remain relatively static for periods of time. We will work to support an associated “standard analysis” for that configuration, as much as possible (it may not be possible in cases where the physics being studied is significantly different from experiment to experiment). This was rated as a high-priority in the 2015 user survey.

3.5 Algorithms

We have identified five projects where collaborations with outside computational experts may result in significant advances for the LCLS user community. This list is neither comprehensive nor final. Projects have been divided into those that will either improve operations or enable new science.

Operational Improvements

➢ Automatic tuning of beam parameters. The LCLS accelerator is currently tuned manually. A typical 12-hour shift routinely requires 15 minutes to 1 hour to tune the accelerator to reach desired beam parameters. Automation of, for instance, the quadrupole focusing magnet current settings along the accelerator could result in significantly faster (order minute) tuning with improved results (stability, FEL pulse energy, etc). We’ll be researching algorithms and software capable of tuning the accelerator automatically.

➢ Streaming classification. Classification of images obtained on LCLS area detectors are currently employed by select user groups to separate data of interest from background. For example, in SPI experiments manifold embedding has proven capable of isolating missed shots, single particle images, and multi-particle images. We will research extending classification to a streaming mode, where data of interest could be identified and tagged before reaching disk, enabling vetoes of uninteresting data and rapid analysis of the most promising data first.

Scientific Expansion

➢ Single particle reconstruction. Coherent imaging is a domain where LCLS provides unique capabilities when compared to other light sources. The ability to extend coherent imaging to Angstrom-scale resolution will require algorithms capable of performing near-optimal inference of structure from available noisy, incomplete, and vast (100s TB) diffraction image data. We will continue to work with the Single Particle Initiative (SPI), a collaboration of over 20 institutions led by the LCLS, to develop algorithms for coherent imaging and ensure that these algorithms are made available to the entire LCLS community.

➢ Crystallographic post-refinement. Serial femtosecond crystallography (SFX) has become one of the most common experimental modes performed at LCLS. Most SFX analyses currently conducted rely on Monte Carlo integration to estimate structure factor intensities, which requires order 10,000 to 100,000 good-quality images in order to determine a structure. Monte Carlo integration ignores the fact that Bragg peaks observed on a single image contain significant mutual (correlated) information. Employing this information is known as post-refinement in the crystallographic community, and has been shown by simulation and preliminary experiments to allow for structure factor estimation from order 1000 images, a possible 10-100x improvement in...
operational efficiency. We’ll work towards developing this preliminary technology into a robust tool all SFX users can employ.

➢ *Interpretation of crystallographic diffuse scatter.* Non-Bragg scatter from crystals contains information about correlated motions inside crystal unit cells. Such information has routinely been used in studies of materials, for example to study phonon modes. The ability to study correlated motions in protein crystals -- which could open a new understanding of protein dynamics such as allostery -- via diffuse scatter has been limited by radiation damage at synchrotron sources, but may be routinely possible at LCLS if the diffuse data can be readily analyzed. We’ll work towards developing algorithms capable of interpreting diffuse scatter in SFX experiments into dynamical models of protein structure.

References

[1] LCLS-II Science:

[2] LCLS-II Science Opportunities Document:

[3] LCLS Detector Strategy:
https://slacspace.slac.stanford.edu/sites/reviews/LCLS/sac_oct_2015/backup_docs/Forms/AllItems.aspx

[4] LCLS Data analysis users survey analytics:
https://goo.gl/xnuAeT

Contact

Amedeo Perazzo
Email: perazzo@slac.stanford.edu
Tel: 650-926-5147