

Reference ID: 11227061061\_Thayer

---

**Reference ID:** 11227061061\_Thayer

**Submission Date and Time:** 12/16/2019 4:50:26 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, <https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp>. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

*Consent Statement:* "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

**Consent answer:** I consent to NSF's use and display of the submitted information.

#### **Author Names & Affiliations**

Submitting author: Jana Thayer - SLAC National Accelerator Laboratory, on behalf of the LCLS facility

**Additional authors:** Amedeo Perazzo, SLAC National Accelerator Laboratory; Robert Schoenlein, SLAC National Accelerator Laboratory; Paul Fuoss, SLAC National Accelerator Laboratory

**Contact Email Address** (for NSF use only): (hidden)

#### **Research domain(s), discipline(s)/sub-discipline(s)**

Jana Thayer, data acquisition systems and high-performance software in the fields of HEP, astrophysics, and photon science; Amedeo Perazzo, architecture of data acquisition and data analysis systems for HEP, astrophysics and photon science experiments, pattern recognition, trigger/veto systems, silicon detectors, low noise electronics, Montecarlo simulations, large scale software development, advanced controls systems, data management and HPC; Robert Schoenlein, Solution Phase Chemistry

#### **Title of Response**

Data Processing Challenges at the Linac Coherent Light Source

## Abstract

The Linac Coherent Light Source (LCLS) is an X-ray Free Electron Laser (XFEL) generating intense ultrafast coherent X-ray pulses to address grand challenge problems in materials science, condensed matter physics, chemistry, atomic and molecular science, and structural biology. LCLS is a user facility that serves a few thousand users across a wide range of material science disciplines. By 2025, the data throughput at LCLS will increase by 1,000-fold, enabling a qualitative advance in science impact. Users of LCLS require an integrated combination of data processing and scientific interpretation, where both aspects demand intensive computational analysis. The high repetition rate and uniform/programmable time structure of LCLS-II (and LCLS-II-HE) will provide a transformational capability to collect 10<sup>8</sup>-10<sup>10</sup> scattering patterns (or spectra) per day. By exploiting revolutionary advances in data science, developing and applying advanced computational algorithms to massive datasets (e.g. kinetic inference, Bayesian analysis, pattern recognition, manifold maps, and machine learning algorithms) it will become possible for the first time to characterize heterogeneous ensembles of particles, map stochastic dynamics, or extract new information about rare transient events from comprehensive data sets of X-ray scattering patterns and/or spectra. These capabilities promise to revolutionize forefront areas of chemistry and materials science.

**Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s).** Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

Studies at LCLS elucidate the detailed electronic and atomic structural dynamics of multi-electron photocatalysts. A prototypical example is the Mn<sub>4</sub>CaO<sub>5</sub> cluster that underpins the water-oxidation cycle of photosystem II (PS-II). Understanding PS-II will inspire the development of artificial photosynthetic molecular assemblies for capturing, transforming, and storing solar energy. Recent results (Nature 563: 421, 2018) provide the first high-resolution structures of all the intermediate states of the photocycle - under operating conditions, at room temperature.. Recent advances in our understanding of PS-II exploit a powerful new tool enabled by XFELs: Serial Femtosecond X-ray Crystallography (SFX). SFX reveals transient metastable molecular structures at the atomic scale, and will be instrumental in many areas of science, ranging from the determination of intermediate structures during chemical reactions and biological processes, to understanding and controlling of materials nucleation pathways. A second example is in the area of heterogeneous catalysis, where the stochastic evolution of atomic and electronic structure, the making and breaking of chemical bonds, and the exchange of vibrational energy ultimately determine functionality. These interactions further lead to dynamic restructuring of catalyst materials during reaction cycles. Knowing the time evolution of the atomic and electronic structure of molecules and substrates, particularly near transition states, is critical to developing a predictive understanding for design of new catalysts. LCLS-II-HE will enable completely new approaches for simultaneously following both the atomic and electronic structure of heterogeneous catalysts in

operation. Single-Particle Imaging (SPI) is a promising emerging technique for such applications. Diffraction images are collected from individual particles, and are used to determine molecular (or atomic) structure, and their evolution, in operating conditions that are inaccessible through other methods. These techniques will be instrumental in many areas of science, ranging from understanding and controlling nano-materials self-assembly, to the chemistry and morphology of combustion aerosols to the coupled electronic and nuclear dynamics in heterogeneous (nano) catalysis. A key science opportunity for LCLS-II-HE is the potential for capturing “rare events” at the atomic scale. Specific problems include: dielectric breakdown in polymers (i.e. how do capacitors fail), spatio-temporal defect dynamics in metals (i.e. how do materials break), and phase transitions including nucleation and martensitic transformations. Part of the challenging nature of these problems is that the interesting events happen at random, stochastic times that make observation difficult.

**Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s).** Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

LCLS-II-HE, coupled with advanced computational methods applied to massive datasets, will enable completely new approaches for simultaneously following both the atomic and electronic structure of dynamic chemical and materials systems in operation. The combination of the data volume produced by LCLS-II-HE and data analysis needs will require exascale computing capabilities. Future offline computing systems will require a mix of local capabilities and access to DOE High End Computing (HEC) facilities for the highest demand experiments. SFX experiments provide ultrafast time resolution without radiation damage under operating conditions. These advanced methods present challenges for data processing and interpretation. In SFX, the structural information is derived from the scattering data through four main steps: (1) identifying Bragg diffraction spots, (2) deducing the lattice geometry, (3) refining the model, and (4) summing the X-ray signal in each spot for further analysis. Based on the current requirements for SFX experiments, peak processing needs are estimated to be 1 PFLOPS in 2020 and 60 PFLOPS in 2025. Beyond the existing pipeline, advanced analysis techniques including powerful new algorithms, the analysis of imperfect crystals, and the extraction of information from diffuse X-ray scattering, will require between one and three orders of magnitude increase in processing capacity. Thus, future computational pipelines will require exascale processing. Single Particle Imaging enables the ability to serially interrogate single molecules with ultrafast XFEL pulses to determine the three-dimensional structure of biomolecules without crystallization. The complexity of retrieving atomic structure from noisy and incomplete scattering data will require computing capabilities beyond that

required by nanocrystallography. Capturing rare events with high fidelity requires the acquisition of a significant number of high-resolution images at closely spaced times triggered by the event. For example, high rate, low resolution detectors may be used as a trigger to read out low rate, high resolution detectors (and associated massive data sets). Many XFEL science areas will have similar computing requirements. Multi-dimensional Spectroscopy (X-ray wavemixing) incorporates time-ordered sequences of X-ray pulses to generate a signal that is a function of multiple time delays and/or photon energies. X-ray pulses are used as both a pump, to prepare specific near equilibrium states of matter, and as a probe of these evolving states. X-ray Photon Correlation Spectroscopy studies the stochastic near-equilibrium processes occurring in condensed matter systems. Fluctuation X-ray Scattering collects a very large number of diffraction images from unoriented particles, and determines the structures from higher order correlation data. The number of acquired images and the computational complexity will ultimately require advanced algorithms that are incorporated into large-scale parallel codes. The capabilities required for these science opportunities are driven by the twin needs of high-throughput data handling and improved mechanics and useability of the light source/HEC Superfacility: real time monitoring of data quality with sub-second latency, extracting science content from the data before saving to persistent storage, data quality monitoring with latency smaller than the typical length of a measurement, and sufficient resources to run the full analysis with no stringent time requirements. The gaps fall into two areas: performance and simplicity. Performance capabilities address the challenges of high throughput data streaming, fast feedback, and data management. The gaps compared to current plans are as follows. Quasi-real-time analysis results from HEC today require advanced block reservations that do not pair well with the bursty nature of the FEL. Allocation of reserved data paths, ESnet resources, and facility resources are needed to allow transparent, facility agnostic data management. Interactive analysis capabilities built on efficient, scalable, user-friendly visualization tools that filter and display information to the user during data-taking or analysis are necessary. Data storage and access become a bottleneck as throughput and concurrency increase several-fold.

**Question 3 (maximum 300 words) – Other considerations.** Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

In order to accommodate end to end workflows that span light sources and HEC facilities, it will be necessary to coordinate resources at multiple levels. The data flow across facilities will need to be coordinated, using resources such as ESnet to stream data. On the supercomputer, it will be necessary to coordinate the execution of sophisticated workflows and large compute jobs on supercomputers as well as the exchange of data between the internal layers of the supercomputer. This will require a certain amount of infrastructure development and a corresponding shift in HEC facility policies to accommodate streaming data analysis and bursty jobs, a contrast to the large simulations that dominate HEC facility usage today. The policies and processes that define the boundaries and areas of responsibility across this geographically distributed tool chain will need to be refined. LCLS-II-HE, by its nature, demands a computationally intensive workflow, however users and user groups vary in their

levels of computing expertise. There is a gap between the tools needed to do the science and a user-friendly system that is capable of accelerating the process of data analysis. This can be achieved by educating users, providing users with additional resources to assist with data analysis, or by investing in the tools' ease of use. The work done at LCLS impacts many PIs from the NSF community. The LCLS facility is interested in engaging with these PIs in the development of the capabilities described here to ensure maximum impact and scientific output.

-- End Submission --