Massive Scale Data Analytics at LCLS-II

Jana Thayer[1*], *Zhantao* Chen[1], *Richard* Claus[1], *Daniel* Damiani[1], *Christopher* Ford[1] , *Mikhail* Dubrovin[1] , *Victor* Elmir[1] , *Wilko* Kroeger[1] , *Xiang* Li[1] , *Stefano* Marchesini[1], *Valerio* Mariani[1] , *Riccardo* Melcchiori[1] , *Silke* Nelson[1] , *Ariana* Peck[2] , *Amedeo* Perazzo[1], *Frederic* Poitevin[1] , *Christopher Paul* O'Grady[1], *Julieth* Otero[1] , *Omar* Quijano[1] , *Murali* Shankar[1] , *Monarin* Uervirojnangkoorn[1] , *Riccardo* Veraldi[1], *Matthew* Weaver[1] , *Clemens* Weninger[1], *Seshu* Yamajala[1] , *Cong* Wang[1] and *Chun Hong* Yoon[1]

[1]SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA
[2]Chan Zuckerberg Imaging Institute, 2682 Middlefield Road, Redwood City, CA 94063, USA

Abstract. The increasing volumes of data produced at light sources such as the Linac Coherent Light Source (LCLS) enable the direct observation of materials and molecular assemblies at the length and timescales of molecular and atomic motion. This exponential increase in the scale and speed of data production is prohibitive to traditional analysis workflows that rely on scientists tuning parameters during live experiments to adapt data collection and analysis. User facilities will increasingly rely on the automated delivery of actionable information in real time for rapid experiment adaptation which presents a considerable challenge for data acquisition, data processing, data management, and workflow orchestration. In addition, the desire from researchers to accelerate science requires rapid analysis, dynamic integration of experiment and theory, the ability to visualize results in near real-time, and the introduction of ML and AI techniques. We present the LCLS-II Data System architecture which is designed to address these challenges via an adaptable data reduction pipeline (DRP) to reduce data volume on-the-fly, online monitoring analysis software for real-time data visualization and experiment feedback, and the ability to scale to computing needs by utilizing local and remote compute resources, such as the ASCR Leadership Class Facilities, to enable quasi-real-time data analysis in minutes. We discuss the overall challenges facing LCLS, our ongoing work to develop a system responsive to these challenges, and our vision for future developments.

1 Introduction

In 2009, the Linac Coherent Light Source (LCLS) facility, the world's first X-ray free-electron laser (XFEL) began operations at SLAC National Accelerator Laboratory and has had a profound impact on a broad cross-section of scientific fields [1-4]. XFELs provide X-ray beams at wavelengths on the atomic scale and over nine orders of magnitude brighter [5] than a synchrotron source. The ultrafast x-ray pulses (<5-500 fs) from LCLS give researchers the tools to probe complex atomic and molecular structures using

* Corresponding author: jana@slac.stanford.edu

individual pulses, revealing the fundamental processes of biology, chemistry, physics, materials, and other phenomena providing atomic resolution with femtosecond precision and chemical specificity.

The LCLS-II upgrade, coming online in late 2023, increases the repetition rate from 120 Hz to ~1 MHz. The adoption of ultra-high repetition rate imaging detectors will further increase the data throughput from today's 1 - 5 GB/s to >200 GB/s. Future planned upgrades are expected to increase the throughput to multiple TB/s. 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 [6] require careful signal correction (pedestal, gain, crosstalk, nonlinearity, geometry) to correctly interpret the science data. A timing system distributes a global timestamp and unique pulse identifier for each x-ray pulse that comes down the line to interact with the sample in the endstation. Every detector participating in the data acquisition receives this trigger and applies a unique timestamp to the data associated with each pulse. Data from different sensors that belong to the same timestamp, or event, can later be re-assembled and analysed in the same context by the offline processing, a process called event building. The LCLS Data System provides the capacity to reduce the data volume on-the-fly using experiment-specific data compression. Figure 1 shows the anticipated increase in raw and reduced data rates as a function of time.

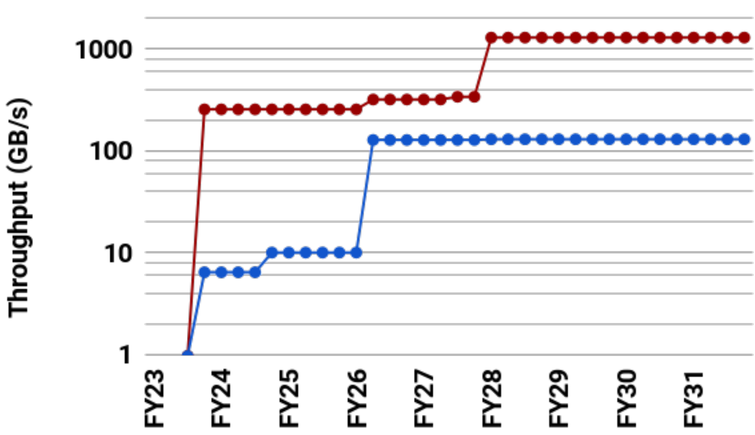

Fig. 1. The red line shows the instantaneous throughput LCLS-II produces as a function of time, showing the increase, due to instrument and detector upgrades from ~250 GB/s in 2023 to over 1 TB/s by the end of the decade. The blue line shows the instantaneous throughput after applying experiment-specific data reduction.

Data acquisition, data transport, event building, data management, and workflow orchestration at LCLS echo the infrastructure needs of high energy physics and other high throughput, computationally demanding endeavours.

2 LCLS-II Challenges

Advanced data and computing systems are playing an increasingly vital role in LCLS operation, data interpretation and overall scientific productivity. The convergence of advances in hardware architectures, development of AI/ML methods, and improvements in computational capabilities provide an opportunity to accelerate science, leverage large datasets, and optimize the use of oversubscribed beamtimes. LCLS-II represents SLAC's largest data and computing challenge, characterized by a variety of time-sensitive and data integration-intensive workflows. The LCLS Data System provides the infrastructure to acquire very high throughput data, transport tens of GB/s of data to disk, cache up to 1 PB of data per shift, manage up to 100 PB of data on disk, provide access to sufficient computing resources for analysis, and provide sophisticated analysis frameworks for data access [7-8]. The increase in repetition rate due to LCLS-II introduces a step change in the size and complexity of data sets and demands a similar advance in computing, algorithms, and analysis to fully exploit the incredibly rich information content contained in this data.

Computing demands at LCLS are driven by the repetition rate of the source, advances in detector technology, advances in data analysis algorithms, and the requirement to provide flexible and easy-to-use fast feedback to users in real time, a challenge given the weekly turnaround of experiments and the one thousand users cycling through the LCLS facility each year. The data life cycle and computational needs of the typical LCLS user vary from experiment to experiment, but on average can be described as follows. A typical experiment is of order 3 – 5 days of 12-hour shifts. Users are responsible for setting the goals of the experiment and work in partnership with beamline scientists and facility staff to operate the beamline and acquire the desired data during their scheduled beam time. Experiments are short, on the order of days, and oversubscription makes it difficult to repeat an experiment at a future date. Likewise, once a measurement begins it cannot be slowed down or stopped if the architecture and data pipelines cannot keep up with the required data rates.

During data collection, it is essential for the users and beamline scientists/staff in the control room to have access to real-time feedback (~1 sec), for example visualization of a camera image to ascertain whether an ice crystal is forming at the end of a sample delivery nozzle. Likewise, users must be able to get data quality feedback within about 1 – 10 minutes, the lifetime of a measurement to optimize the experimental setup for the next measurement. Data quality monitoring is computationally more sophisticated than the real time monitoring requirements, answering questions that require some manipulation of the data and/or the combination of several sources of data. For example, crystallographers may want to know whether a 10-minute acquisition period contained an adequate number of single-hits and indexable crystals. Are there enough statistics available to reconstruct the electron density? Is the signal to noise ratio as expected? Is the resolution of the reconstructed electron density what was expected?

During the 12 hours off shift, the users continue to analyse the data and may run as many as ten passes of the full analysis over the data acquired during the previous 12-hours to optimize analysis parameters and prepare analysis code for the next shift. During the 4 months after the experiment, users analyse the raw and intermediate data on fast access storage in preparation for publication. After 4 months, data are archived to tape where they are kept for a period of 10 years and may be restored to disk at any time, a feature that is increasingly exercised as AI/ML becomes more widely used.

LCLS Data Systems has developed a world-leading capability in high throughput data analytics consistent with the leap to 1 MHz operation, and providing fast feedback capabilities, real-time data visualization, and configurable, cutting-edge data analysis pipelines for researchers. Table 1 shows a summary of the LCLS-II data challenges.

Table 1. Summary of LCLS-II data and computing characteristics.

Description	LCLS-II 2023	LCLS-II 2029+
Readout rate	0.01 Hz - 1 MHz	0.01 Hz - 1 MHz
Wanted fraction of collisions	0.01 to 1.0	0.01 to 1.0
Typical experiment duration (same data-taking conditions)	3 days	3 days
Required turnaround for data-quality checks	Seconds to minutes	Seconds to minutes
Raw digital data rate	200 GB/s	1000+ GB/s
Zero-and-Junk-suppressed rate	20 GB/s	100+ GB/s
Storage need dominated by	Mainly raw data	
Role of Simulation	Growing in science analysis Growing in experiment design	
Analysis, Simulation, and Workflow Software development community	Individuals (in the past) → Organized effort	

LCLS Data Systems provides seamless access to computing, making use of local compute resources as well as ASCR compute and networking capabilities when required. These capabilities shorten the time between experiment and publication and provide researchers easier access to scientific data and the computing required to analyse it.

3 The LCLS-II Data System

The basic architecture of the data system is shown in Figure 2. The LCLS facility provides core hardware and software infrastructure for scalable data acquisition, online monitoring, offline analysis, and data management enabling scientists to efficiently go from measurement to scientific insight. The users provide the last mile and, using their science domain knowledge, develop their own analysis on top of this stack.

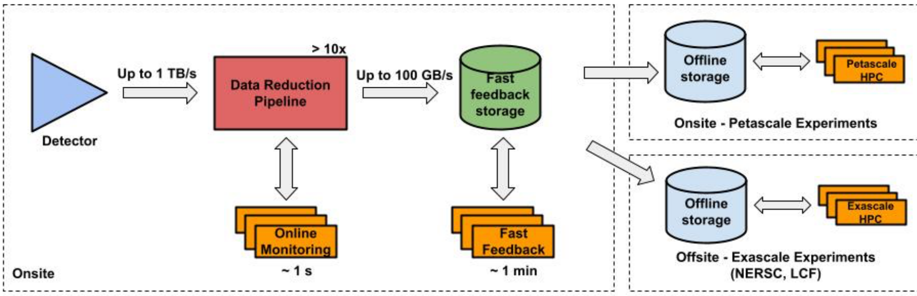

Fig. 2. The basic architecture of the data system shown here illustrates the flow of the data from the front-end electronics on the left to the storage and processing layers on the right.

LCLS instruments shown in Figure 3 offer unique instrumentation to study different areas of science – ranging from atomic and molecular science, ultrafast chemistry and catalysis, advanced materials, structural biology, high-energy density science, to photon science – all instruments use an instance of the same basic data systems architecture for readout, storage, and data processing.

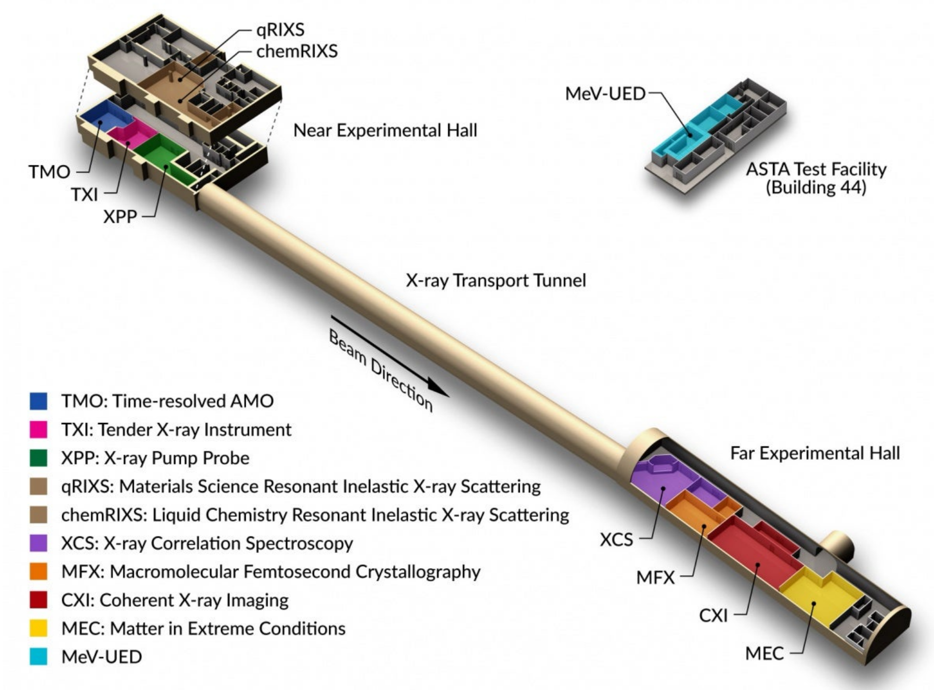

Fig. 3. The suite of LCLS X-ray instruments each have unique capabilities designed to provide a diverse experimental landscape for probing ultrafast dynamics.

The LCLS user facility cycles through ~1,000 users and of order 100 experiments every year. Despite this diversity, the data system can accommodate the sensor quantities and configurations required as well as the various throughput and computational needs of all experiments, such as those shown in Table 2.

Table 2. Examples of compute-intensive workflows showing, for 2023 and 2026, 1) throughput generated at the detector (unreduced), 2) expected computational needs for data reduction to achieve an average reduction in data volume of 10x, and 3) an estimate of computational needs for offline analysis of the data by the user. These are referred to as Level 1, 2, and 3.

Description	LCLS-II 2023			LCLS-II 2026		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
Coherent Scattering (XPCS, XSVS)	20 GB/s	4 TF	34 TF	80 GB/s	34 TF	270 TF
Liquid Scattering	40 GB/s	<1 TF	20 TF	320 GB/s	<1 TF	50 TF
Resonant Inelastic Scattering	20 GB/s	4 TF	1 TF	200 GB/s	40 TF	2 TF
X-ray Emission Spectroscopy (XES)	1 GB/s	<1 TF	2 TF	10 GB/s	<1 TF	20 TF
Coincidence Spectroscopy	200 GB/s	<1 TF	<1 TF	200 GB/s	<1 TF	<1 TF
Nonlinear Spectroscopy	20 GB/s	3 TF	<1 TF	80 GB/s	16 TF	<1 TF

3.1 Data Acquisition (DAQ)

Data flows from a variety of detectors arranged around endstations within each instrument. These endstations may be reconfigured weekly with different sensors added and of order tens of sensors read out through the DAQ. The sensors are a combination of commercial cameras, SLAC-built custom cameras [7-10], waveform digitizers, or point values of temperatures, voltages, pressures, etc. All sensor data flows through a computing layer called the Data Reduction Pipeline (DRP) which reduces the data volume on-the-fly by an order of magnitude on average using experiment-specific techniques. Online analysis of scientific data is vital feedback to the operation of LCLS experiments which require real time tuning of components for beamline operation. The Analysis Monitoring Interface (AMI) [11-13] software framework provides real time (<1 second) analysis of the acquired data and is implemented on a scalable set of monitoring nodes connected to the same network which connects the DRP nodes. AMI provides both a graphical and a scripted interface. Users primarily interact through a graphical user interface (GUI) used to configure and display analysed information on-the-fly, but users may also choose to run their analysis python scripts in real-time on the "live" data feed. AMI operates on data event built from the sensor data received over the InfiniBand data network from the DRP. A statistical subsample of the data passing through the DRP is read from memory and presented to the user for analysis providing a valuable tool for real time feedback and validation of the DRP performance and parameters. Data are then written to Non-Volatile Memory Express (NVMe) devices in the Fast Feedback (FFB) Layer where data are locally cached awaiting transfer to offline storage. Depending on the computational needs of the experiment, data are transferred either to disk storage on the SLAC Shared Science Data Facility (S3DF) [14] for Petascale experiments or disk storage at a remote High Performance Compute (HPC) Facility for Exascale experiments.

3.2 Automated Data Movement and Run Processing

The LCLS data system handles the transparent data movement within several layers of computing in the pipeline from the Detector Edge (FPGA/ASIC) through the data reduction compute layer (CPU, FPGA, GPU), to the data cache FFB Layer where it is accessible to users for fast feedback analysis for approximately 1–5 shifts. From the data cache, the first persistent storage layer, data is automatically transferred by the data system to offline storage using *bbcp* [15] and *xrootd* [16], where the data remain on disk for 4 months post experiment. Data is also automatically archived to tape where it remains for 10 years. A second copy of the data is held on tape at NERSC. LCLS provides space for all experimental data at no additional cost to the user. LCLS-I (120 Hz) datasets are between 1 - 50 TB in size, but LCLS-II datasets are expected to be about 1 PB per shift and up to several petabytes per experiment.

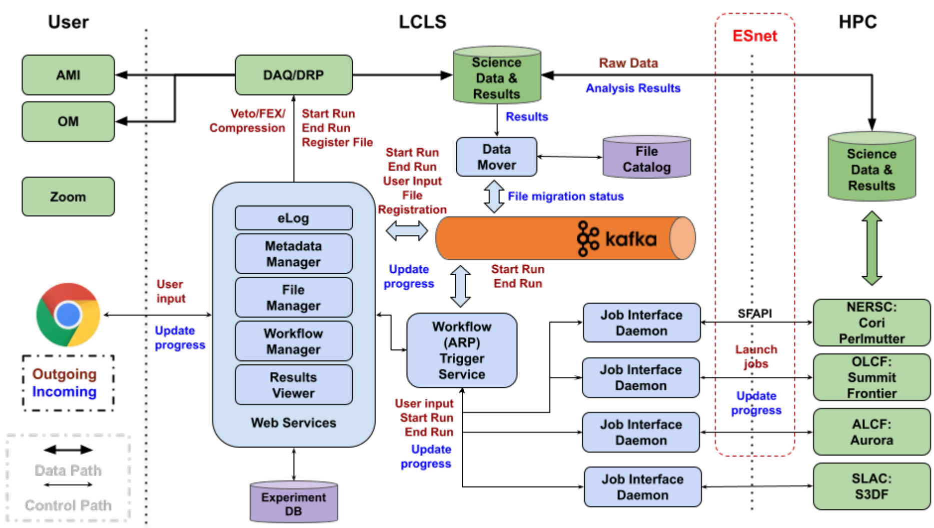

Fig. 4 The LCLS data management system is shown. The user interface (left) consists of a web browser for logging, configuring analysis, and viewing results, online monitoring software (OM and AMI), and optional Zoom for remote participants. The internals of the data management system shown in the LCLS (center) panel are opaque to the user and fully automated. Data are acquired and automatically transferred to the designated HPC where workflows are triggered. Results are migrated back to the web browser on the user (left) side.

Figure 4 shows the key features of the LCLS Data Management System which includes authentication/authorization for users, a file catalog based on Rucio [17], an electronic logbook for experiments, an experiment database with the capacity to automatically capture metadata at the point of data acquisition, a sample database, a workflow manager, feedback and reporting tools for users, instrument operator portals, and automated run processing capabilities. The user interacts with the system through a web browser. The web services allow an experimenter to monitor the flow of data from the DAQ through the FFB to disk storage and tape archival. The user configures their analysis in the browser, and the data management system takes care of the rest, automatically transferring the data, using SLURM to launch thousands of parallel jobs on remote HPC, monitoring the status of the jobs as they run, and returning the results back to the user as they are accumulated.

Analysis is performed by user teams using available computing resources. For each experiment technique, the LCLS Data Systems team, in collaboration with beamline scientists and science domain experts, benchmarked likely techniques for data reduction, fast feedback (data quality monitoring) and offline analysis and factored in user re-processing of data to evaluate computing requirements for LCLS-II experiments. The results are summarized in Figure 5 which shows computing requirements as a function of fiscal year, from FY23 – FY31. The size of the bubble represents the percentage of experiments that will require a particular scale of computing in a particular fiscal year.

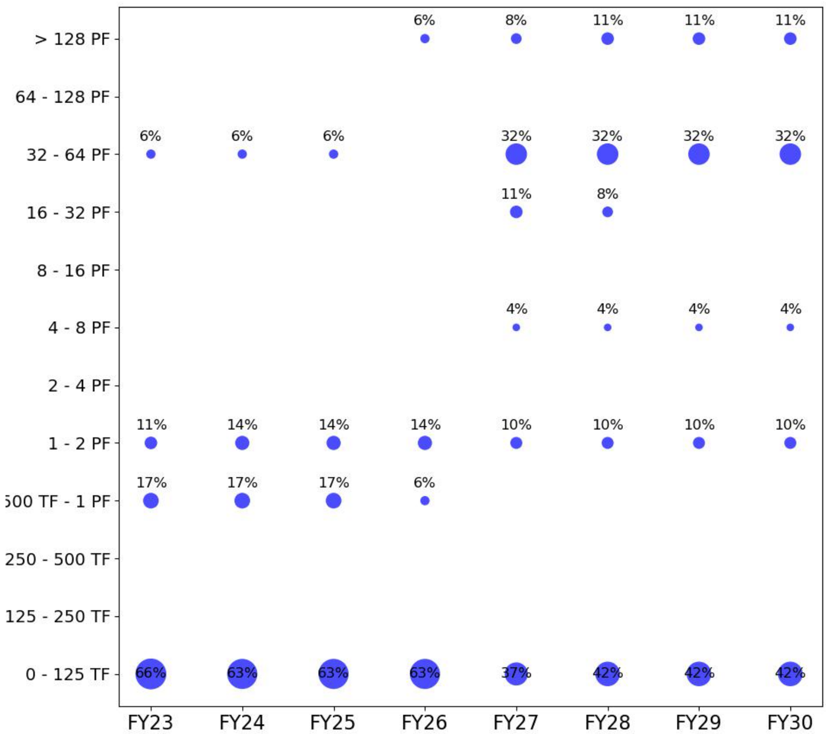

Fig. 5. LCLS computing requirements for experiments as a function of Fiscal Year are shown as a bubble chart. The size of each bubble represents the percentage of experiments that require a specific amount of computing for their analysis. Most experiments require < 5 PFLOPs. A handful of experiment require significantly more compute resources, which are unavailable at SLAC's S3DF. To support these experiments, LCLS will stream data to NERSC for analysis.

For the 80% of experiments requiring Petascale resources, SLAC offers a local, mid-scale computing facility called the Shared Science Data Facility (S3DF) for user analysis. Approximately 20% of experiments have computational requirements that cannot be met by the S3DF and must be offloaded to DOE ASCR facilities such as NERSC. These will heavily leverage ESnet network resources to facilitate data mobility. LCLS to S3DF/ASCR workflows are typically used to analyse raw data for data quality monitoring and experiment feedback or to do post-experiment analysis, archive LCLS data sets, train/retrain AI/ML models, transmit simulation data that can be used during the experiment, and to do multi-modal analysis.

4 Time-Sensitive and Data-Intensive Patterns for Workflows

Time-sensitive patterns for workflows require real-time or end-to-end performance with high reliability for timely decision-making or experiment steering. Data-integration intensive workflow patterns require combining and analysing data from multiple sources such as experiments and computational facilities. LCLS workflows are both time-sensitive and data-intensive, managing the transport, reduction, and analysis of high-throughput data across a heterogeneous pipeline that spans facilities, often incorporating the results of simulations to inform experiment results. LCLS requires transparent solutions that seamlessly link science instruments to compute facilities.

4.1 On-the-fly data reduction: Data Reduction Pipeline

The volume of data produced by experiments is prohibitively expensive to transport and store, making it necessary to reduce the data prior to writing it to persistent storage. LCLS has developed a data reduction pipeline (DRP) using a computing layer composed of FPGAs and CPUs, capable of extracting key features of the data using a toolbox of algorithms targeted to the expected experiment types. On-the-fly data reduction is accomplished using a performant library of algorithms designed to reduce data volume without compromising the scientific results. Lossless compression, generic, all-purpose data reduction, and lossy compression [18] with tuneable error bounds are employed. Feature extraction such as sparsification (peak-finders, region-of-interest selection), calculation such as region-of-interest integration, angular integration, or binning integration, and data transformation into a space where data are sparse (wavelet JPEG style compression) are used to reduce the data volume from imaging detectors. The DRP contains the ability to event build a subset of events enabling the implementation of a software trigger or event veto, useful for experiments like crystallography and single particle imaging that may have a low hit rate. This technique enables the veto of entire events based on user-defined criteria that distinguish between good and bad quality shots.

The DRP pipeline reduces the data volume for many rapidly changing experiments, must be scalable up to the detector I/O maximum capabilities, and is easily configurable to accommodate rapidly changing experimental conditions. The DRP also provides a user-definable fraction of non-reduced events in addition to the reduced data for validation. LCLS has developed the psana2 (based on psana [19]) analysis framework for accessing the data, handling parallelization across multiple nodes and cores and event building. Additional algorithmic tools can be seamlessly integrated on top of psana2 in the DRP to support various user needs, their parameters tuned for a specific experiment and run in highly parallelized fashion. The LCLS data system can run psana-python in the DRP as well as offline making it easier for users to run the same analysis both online and offline. Supporting tools such as robust real-time calibration, beam-position determination, per-shot jet-streak determination, and other per-event backgrounds are also necessary for proper event characterization.

Data reduction confers several advantages beyond cost savings. By applying experiment-specific veto, compression, and feature extraction, the overall throughput and storage requirements for the facility are reduced. This makes it possible to move data more quickly to offsite computing facilities such as NERSC. Furthermore, the offline processing farm required to keep up with the incoming data stream is smaller in size and simpler to maintain. Data reduction mitigates storage, networking, and processing requirements and reduces the time required to go from measurement to scientific insight. Finally, producing feature extracted data earlier in the pipeline provides actionable scientific insight to users or ML agents, enabling experiment steering.

4.2 Intelligent Detector Systems

LCLS needs streaming feature extraction on high-rate data from large imaging detectors to enable smart, autonomous experiments. An architecture capable of real-time feedback built into the ASIC/FPGA tightens the coupling between experimental analysis and data acquisition [20]. Incorporating AI/ML techniques allows the optimization and prioritization of the data acquisition that maximizes the scientific return, but data reduction at the ASIC/FPGA level is challenging. ASICs and FPGAs may only see a fraction of the uncorrected image while the feature extraction algorithms are typically developed on fully calibrated and reconstructed images in the offline analysis.

The SparkPix-RT project develops a new family of experiment-specific detectors that seeks to overcome these obstacles and build a detector capable of extracting information in real-time and operating at the effective rate of 1 MHz, the natural production rate of the data. Specifically, the SparkPix-RT project is developing intelligent auto-correction techniques, provisioning for buffering and deadtime to ensure that the pipeline can absorb some variability in performance, developing triggering capability, determining what kind of information extraction is feasible and developing AI/ML-based workflows for dynamic real-time experiment operation.

4.3 ExaFEL: Compute-intensive analysis of TB/s data

The ExaFEL project has created an exascale-based data analysis workflow for serial femtosecond crystallography (SFX), leveraging exascale computing to reduce, from weeks to minutes, the time to analyse LCLS molecular structure x-ray diffraction data. The high repetition rate and ultra-high brightness of the LCLS make it possible to determine the structure of individual molecules, mapping out their natural variation in conformation and flexibility, but the classification of diffraction patterns into conformational states, and subsequent reconstruction of a series of 3D electron densities, enabling the visualization of how the structure is changing, demand intensive computational analysis. The molecular structure is determined by merging the x-ray diffraction patterns from millions to billions of protein crystals exposed in random orientations.

Exascale computing allows the diffraction pattern to be modelled with greatly enhanced detail, leading to very granular atomic resolution that follows the path of single atoms reacting within a large molecular complex. Additionally, by streaming the experimental data to a supercomputing facility in real time, diffraction quality can be assessed in a matter of minutes [21]. ExaFEL has also created an automated analysis pipeline for imaging of single particles via diffractive imaging. This entails the reconstruction of a 3D molecular structure from 2D diffraction images using the new Multi- Tiered Iterative Phasing (MTIP) [22] algorithm. The LCLS Data Management System is employed to stream data automatically from SLAC to NERSC over ESnet, launch the highly parallel analysis jobs on the supercomputer, and report the results of the analysis back to the experimenters in quasi-real time.

4.4 Connect scientific instruments and HPC to create smart instruments

Autonomous operations at LCLS requires real-time actionable information that drives facility reaction to the experimental environment. The Actionable Information from Sensor to Data Center project [23] integrates ML capabilities with the LCLS-II data flow, from the front-end electronics, to the DRP, to the HPC data center. LCLS will run ML inference in the DRP at three levels of increasing sophistication: in the CPU/GPU, in the FPGAs, and inference-specific ASICs. In addition to the development of the overall infrastructure, the project developed the reconstruction of photo-electron spectra from attosecond angular streaking data, data extraction for SFX and SPI, and in-situ High Energy Diffraction Microscopy using ML. Rapid analysis and closed loop control of experiments at light sources rely on the ability to update ML models rapidly in response to changes in an instrument or sample while an experiment is running. The infrastructure needed to deploy trained ML models on edge devices and leverage HPC to rapidly (re)train AI/ML models on Data Center AI Systems (DCAI) was also developed, taking the turnaround time for a new model from hours to seconds.

5 . Scientific Computing for Facilities

ASCR's Integrated Research Infrastructure Architecture Blueprint Activity [24] has created a vision of a DOE/SC integrated research ecosystem that transforms science via seamless interoperability. LCLS infrastructure is poised to interface to such an architecture as it scales to meet the high-throughput, compute-intensive demand of LCLS-II. Already, the data system is providing real-time data analysis capabilities in the form of data reduction and complex workflow orchestration including on-demand utilization of super-computing environments. LCLS is developing a pipeline that spans from the detector edge to HPC and is strategically developing AI/ML for targeted applications. Intelligent detector systems and real-time analysis will enable autonomous experiment steering and allow users to extract new scientific insight from massive data sets interpreting data in new ways at higher speeds.

Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award Number FWP-100643.

References

1. W. White, A. Robert, M. Dunne, J. Synchrotron Rad., **22**, 472-476 (2015)
2. M. Liang, et al., J. Synchrotron Rad., **22**, 514-519 (2015)
3. B. Nagler, et al., J. Synchrotron Rad., **22**, 520-525(2015)
4. S. Boutet, A.E. Cohen, S. Wakatsuki, Synchrotron Radiation News, **29(1)**, 23-28 (2016)
5. P. Abbamonte, et al., "New Science Opportunities Enabled by LCLS-II X-ray Lasers", SLAC-R-1053, (2015)
6. G. Blaj, et al., J. Synchrotron Rad., **22**, 577-583, (2015)
7. T. van Driel, et al, J. Synchrotron Rad., **27**, 608-615 (2020)
8. P. Caragiulo, et al., "Design and Characterization of a high-rate readout backend for ePix detectors at LCLS II", IEEE Nuclear Science Symposium and Medical Imaging Conference Proceedings, 1–3, (2018)
9. D. Doering, et al., "Readout System for ePixHR X-ray Detectors: A Framework and Case Study". IEEE Nuclear Science Symposium and Medical Imaging Conference. 1-4, (2020)
10. D. Doering et al. "ePixHR10k 2M – A 2M Pixel X-ray Detector at 5,000 Frame Per Second for LCLS-II", IEEE Nuclear Science Symposium and Medical Imaging Conference, (2022)
11. J. Thayer, et al, Advanced Structural and Chemical Imaging, **3**, 3 (2017)
12. J. Thayer, et al., "Data processing at the linac coherent light source", Proceedings of XLOOP 2019: 1st Annual Workshop on Large-Scale Experiment-in-the-Loop Computing: Held in Conjunction with SC19, Denver, Colorado, 32–37. (2019)
13. V. Mariani et al., Journal of Applied Crystallography, **49.3**, 1073-1080 (2016)
14. https://s3df.slac.stanford.edu/public/doc/#/
15. https://www.slac.stanford.edu/~abh/bbcp/
16. https://github.com/xrootd/xrootd
17. https://rucio.github.io/documentation/
18. R. Underwood et al.. "ROIBIN-SZ: Fast and Science-Preserving Compression for Serial Crystallography", arXiv: 2206.11297 [cs.DC], (2022)
19. Damiani, D. et al., Journal of Applied Crystallography, **49,** 672-679 **(**2016)
20. L. Rota et al. "SparkPix-ED: a readout ASIC with 1 MHz frame-rate for rare event experiments at LCLS-II.", iWorID: The International Workshops on Radiation Imaging Detectors, (2021)
21. D. Bard et al., "LBNL Superfacility Project Report". doi: 10.2172/ 1875256 (2022)
22. J. Donatelli, et al., "Reconstruction from limited single-particle diffraction data via simultaneous determination of state, orientation, intensity and phase", PNAS 114 (28), 7222-7227 (2017)
23. Z. Liu, et al.. "Bridging Data Center AI Systems with Edge Computing for Actionable Information Retrieval", 2021 3rd Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP), 15-23, (2021)
24. B. Helland, "Future of Computational Infrastructures: Exascale Computing and an Integrated Research Infrastructure," (2022) [Online]. Available: https://science.osti.gov/-/media/bes/besac/pdf/202212/7-Helland--BESAC-Panel.pdf.