A079-06-280823

Data Analytics at the Linac Coherent Light Source

J. Thayer, R. Claus, D. Damiani, M. Dubrovin, C. Ford, W. Kroeger, S. Marchesini, V. Mariani, R. Melchiorri, C. O'Grady, A. Perazzo, H. Schwander, M. Shankar, M. Uervirojnangkoorn, C. Wang, M. Weaver, C. Yoon

SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025

jana@slac.stanford.edu

Keywords: LCLS, Computing, Big Data

The increase in velocity, volume, and complexity of the data generated by the Linac Coherent Light Source upgrade (LCLS-II), present a considerable challenge for data acquisition, data processing, data management, and workflow orchestration. These systems face formidable challenges due to high data throughput and intensive computational demand for scientific interpretation. Increasingly large data sets at high velocity open new areas of science but make it more difficult to extract desired physical information. Raw data collected cannot all be saved to disk. Handling the throughput between the front-end electronics and the storage layers and between the storage and processing is a critical element of the system. Conventional human-in-the-loop workflows are too slow to process data on useful timescales. Likewise, current analysis pipelines rely on human-in-the loop decision-making and fine tuning of algorithm parameters that do not scale with increased data rates. Compute-intensive analyses require access to compute resources of the appropriate scale which may be local or remote to the facility generating the data and may include ASCR Leadership Class Facilities.

A further challenge, once the data have been recorded, is the useability of the overall system to access the data and analyze it at scale. Providing fast feedback to experimenters on the timescale of seconds and for complex analyses, minutes, reduces the time to complete the experiment, improves the quality of the data, and increases experiment success rate. Speed and flexibility of the development cycle are critical due to the wide variety of experiments, rapid turnaround required, and the need to modify data analysis during experiments even for complex scientific workflows. Integration of simulations, real-time data analysis, computation-assisted experiment design at large scales, and Machine Learning techniques drive the computation and networking needs of the data system.

The LCLS-II Data System architecture addresses these challenges [1]. To reduce data volume and handle the high throughput, we provide an adaptable, heterogeneous data reduction pipeline (DRP), a compute layer composed of edge accelerators (FGPA, GPU, etc) and CPU that can run experiment-specific algorithms in real time to reduce data volumes by an order of magnitude while preserving the science content of the data. The Fast Feedback layer offers dedicated processing resources to running experiments for the purpose of data quality feedback within minutes. LCLS-II's data management system handles automated data movement through various storage layers such as offline storage and long-term data archiving as well as providing transparent data transfer between local and external compute facilities such as National Energy Research Scientific Computing Center (NERSC). A web-based portal allows users to manage their experiments. The data management system integrates with HPC workload managers and SLAC, NERSC and other facilities to automatically trigger analysis jobs providing easy access to local and remote offline processing capabilities. A real-time analysis framework provides visualization and graphically-configurable analysis of data on the timescale of seconds. The LCLS-II Data System mitigates bottlenecks in computing, storage, and network resources and enables tighter integration between data collection and analysis. An overview of the LCLS-II Data System architecture will be presented. New innovations to accommodate AI/ML at the edge will be described [2].

[1] Thayer, J., Damiani, D., Dubrovin, M., Ford, C., Kroeger, W., O'Grady, C., Perazzo, A., Shankar, M., Weaver, M., Wennger, C., Yamajala, S., and Zohar, S *Data Processing at the Linac Coherent Light Source.* IEEE/ACM 1st Annual Workshop on Large-scale Experiment-in-the-Loop Computing (XLOOP), Denver, CO, 18 Nov 2019.
[2] 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), St. Louis, MO, USA. 2021, pp 15-23.

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. 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.