# Data Analytics

## ExaFEL: Data Analytics at the Exascale for Free Electron Lasers

The SLAC Linac Coherent Light Source (LCLS) facility uses x-ray diffraction to image individual atoms and molecules to observe fundamental material processes. Near-real-time interpretation of molecular structure revealed by x-ray diffraction will require computational intensities of unprecedented scales coupled with a data path of unprecedented bandwidth. Detector data rates at light sources are advancing exponentially: with the LCLS-II-HE upgrade, LCLS will increase its data throughput by three orders of magnitude by 2025. The objective of the ExaFEL project is to leverage exascale computing to reduce, from weeks to minutes, the time to analyze molecular structure x-ray diffraction data generated by LCLS.

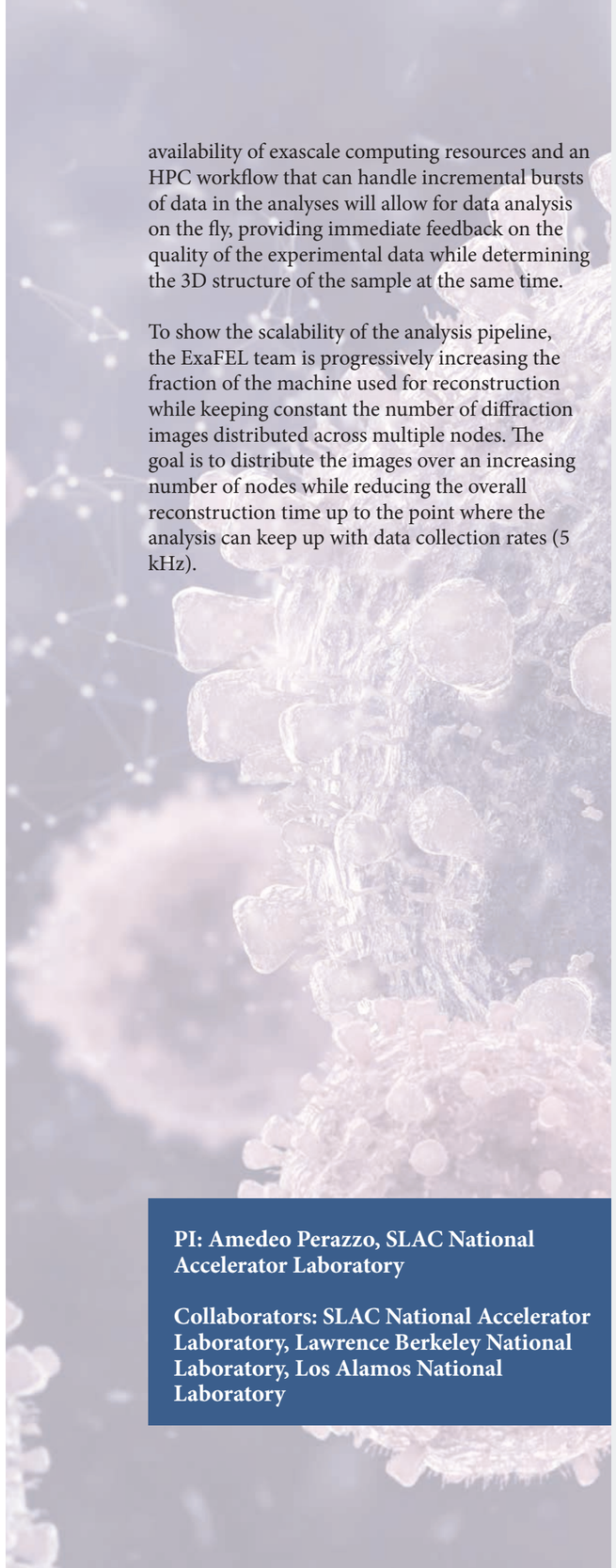
Users of the LCLS require an integrated combination of data processing and scientific interpretation, where both aspects demand intensive computational analysis. The ultrafast x-ray pulses are used like flashes from a high-speed strobe light to produce "stop-action movies" of atoms and molecules. The analysis must be carried out quickly to allow users to iterate their experiments and extract the most value from scarce beam time. Enabling new photon science from the LCLS will require near-real-time analysis (~10 min) of data bursts, requiring commensurate bursts of exascale-class computational intensities.

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. Structural dynamics and heterogeneities, such as changes in size and shape of nanoparticles, or conformational flexibility in macromolecules, are at the basis of understanding, predicting, and eventually engineering functional properties in the biology, material, and energy sciences. The ability to image these structural dynamics and heterogeneities using noncrystalline-based diffractive imaging, including single-particle imaging (SPI) and fluctuation x-ray scattering, has been one of the driving forces behind the development of x-ray free-electron lasers. However, efficient data processing, classification of diffraction patterns into conformational states, and subsequent reconstruction of a series of 3D electron densities, which allow for visualization of how the structure is changing, are vital computational challenges in diffractive imaging.

The ExaFEL challenge problem is the creation of 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) algorithm. In SPI, diffraction images are collected from individual particles and are used to determine molecular (or atomic) structure, even from multiple conformational states (or nonidentical particles) under operating conditions. Determining structures from SPI experiments is challenging because orientations and states of imaged particles are unknown and images are highly contaminated with noise. Furthermore, the number of useful images is often limited by achievable single-particle hit rates, currently much less than 1. The MTIP algorithm introduces an iterative projection framework to simultaneously determine orientations, states, and molecular structure from limited single-particle data by leveraging structural constraints throughout the reconstruction, offering a potential pathway to increasing the amount of information that can be extracted from single-particle diffraction.

Rapid feedback is crucial for tuning sample concentrations to achieve a sufficient single-particle hit rate, ensuring that adequate data are collected and available to steer the experiment. The availability of exascale computing resources and an HPC workflow that can handle incremental bursts of data in the analyses will allow for data analysis on the fly, providing immediate feedback on the quality of the experimental data while determining the 3D structure of the sample at the same time.

To show the scalability of the analysis pipeline, the ExaFEL team is progressively increasing the fraction of the machine used for reconstruction while keeping constant the number of diffraction images distributed across multiple nodes. The goal is to distribute the images over an increasing number of nodes while reducing the overall reconstruction time up to the point where the analysis can keep up with data collection rates (5 kHz).



**PI: Amedeo Perazzo, SLAC National Accelerator Laboratory**

**Collaborators: SLAC National Accelerator Laboratory, Lawrence Berkeley National Laboratory, Los Alamos National Laboratory**

### Progress to date

- Implemented Psana tasking, a port of Psana to the Legion programming system for exascale computing, that demonstrated performance comparable to MPI when running on up to 2,048 nodes of the Cori supercomputer.
- Demonstrated that using Legion in Psana tasking enables GPU use, a critical step toward readiness for upcoming exascale architectures.
- Optimized the merging step in the nanocrystallography pipeline to first read in the reduced data files in parallel and then to merge all duplicate observations together. Critically, the new algorithm allows the project team to merge a data set of $>10^9$ observations, which was previously intractable with a single process.
- Extended LUNUS diffuse scattering data processing to handle $10^3$–$10^4$ diffraction images in parallel across multiple nodes. Developed MPI LUNUS for performance analysis and optimization and defined image processing requirements for SFX diffuse scattering data.

ExaFEL will enable real-time analysis of the thousandfold planned increase in x-ray diffraction data at LCLS, which will greatly increase the facility's ability to answer fundamental science questions about the nature of matter.