Real-time ML at the Linac Coherent Light Source

Fast ML for Science at ICCAD 2023 Workshop

Jana Thayer, Zhantao Chen, Cong Wang

November 2, 2023

SLAC NATIONAL ACCELERATOR LABORATORY

Stanford University | U.S. DEPARTMENT OF ENERGY

Contents

Introduction to the Linac Coherent Light Source (LCLS) - the most powerful X-ray Free electron Laser in the world

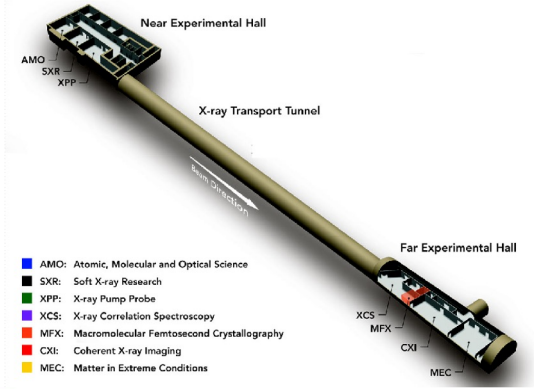
Challenges

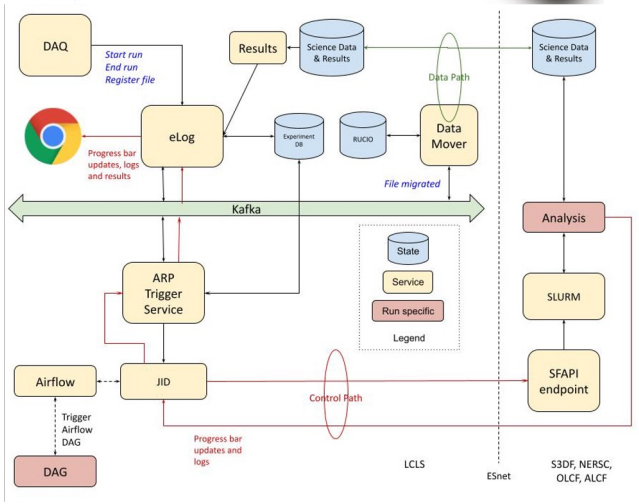
- Variety of workflows
- Data Volume and Velocity
- Computing

Opportunities

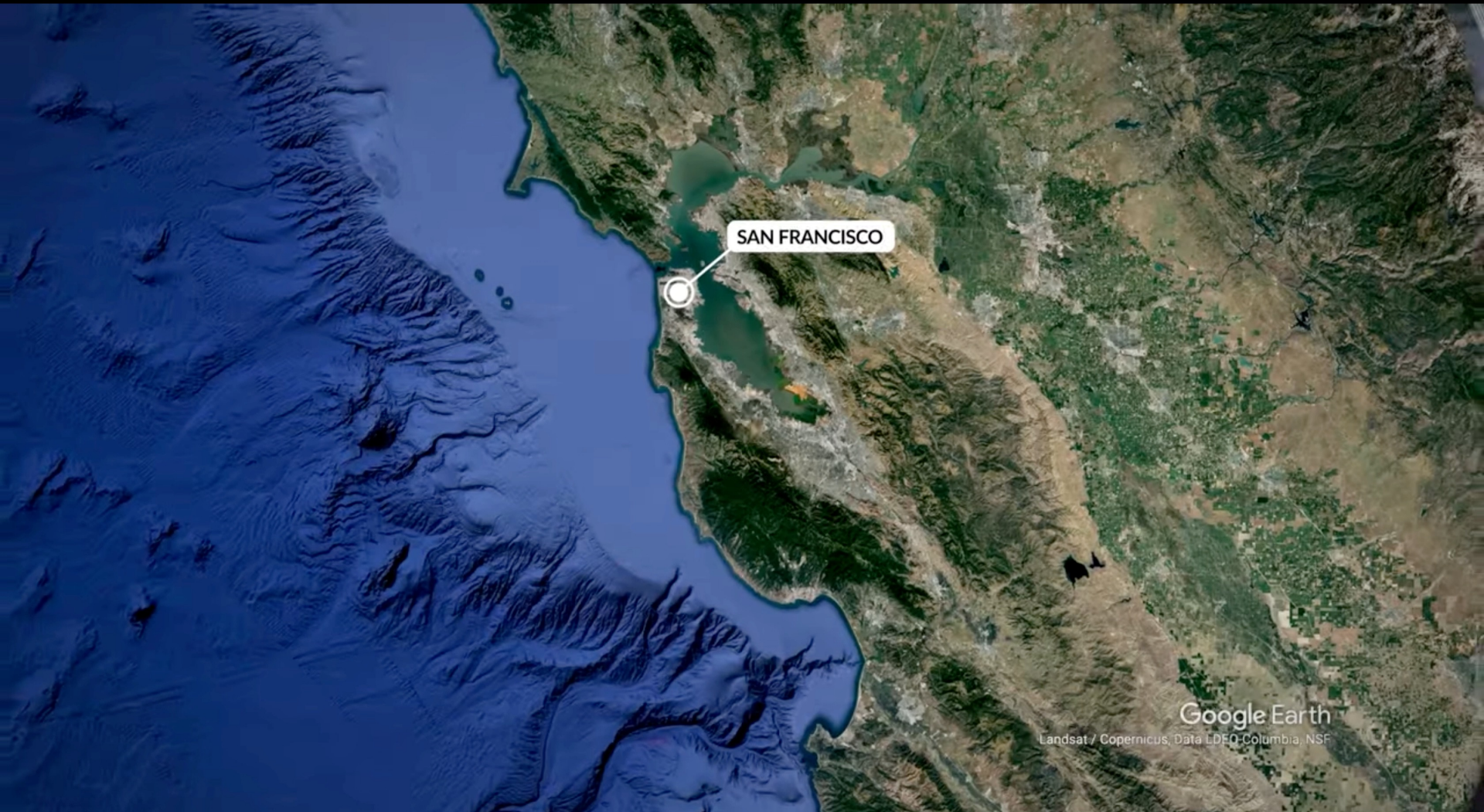
- AI/ML at the Edge
- Better Science through ML
- Edge to HPC workflows
- Experiment Steering

New infrastructure and analysis methods that leverage massive data quantities will maximize the science output from the Linac Coherent Light Source





SLAC

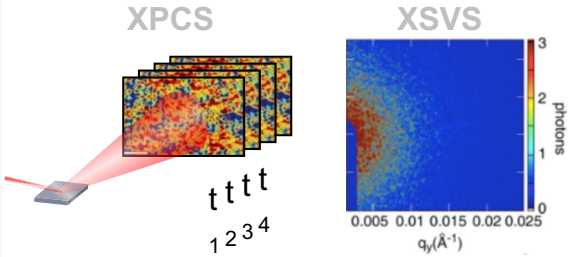

SAN FRANCISCO
Google Earth
Landsat / Copernicus, Data LDEO-Columbia, NSF

Linac Coherent Light Source Challenges

Linac Coherent Light Source: 20+ Experimental Techniques with Unique Workflows

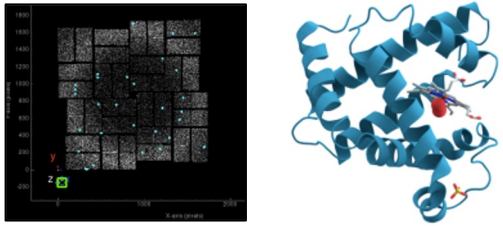
Each workflow with different throughput and compute needs; need flexible development cycle

Coherent Scattering



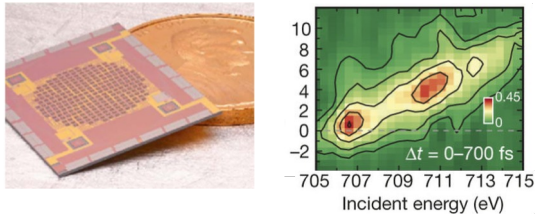
2022: 20 GB/s, 4 TF (reduction), 34 TF (analysis)
2026: 80 GB/s, 34 TF (reduction), 270 TF (analysis)

Nanocrystallography



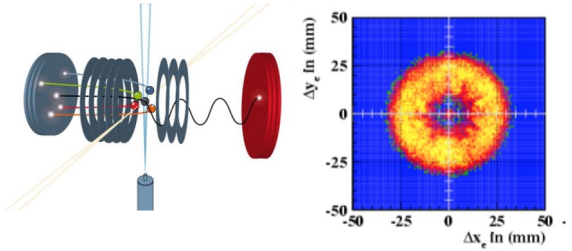
2023: 64 GB/s, 3 TF (reduction), 4 TF (analysis)
2026: 1.2 TB/s, 16 TF (reduction), 20 TF (analysis)

Resonant Inelastic Scattering



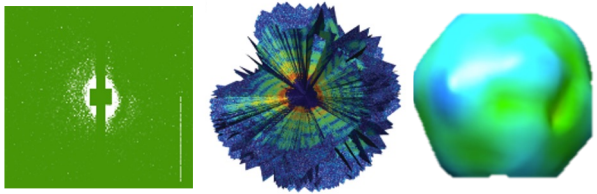
2023: 20 GB/s, 4 TF (reduction), 1 TF (analysis)
2026: 200 GB/s, 40 TF (reduction), 2 TF (analysis)

Coincidence Spectroscopy



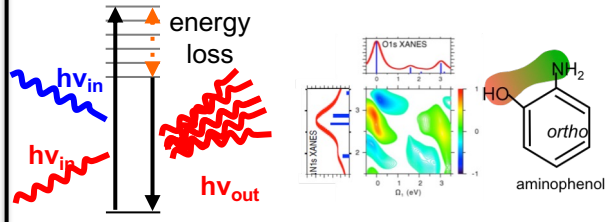
2021: 200 GB/s, <1TF (reduction), <1TF (analysis)

Coherent Imaging



2022: 64 GB/s, 3 TF (reduction), 270 TF (analysis)
2026: 1.2 TB/s, 16 TF (reduction), 1340 TF (analysis)

Nonlinear Spectroscopy



2023: 20 GB/s, 3 TF (reduction), <1 TF (analysis)
2026: 80 GB/s, 16 TF (reduction), <1 TF (analysis)

SLAC

Challenge: High Throughput, Large Data Volume

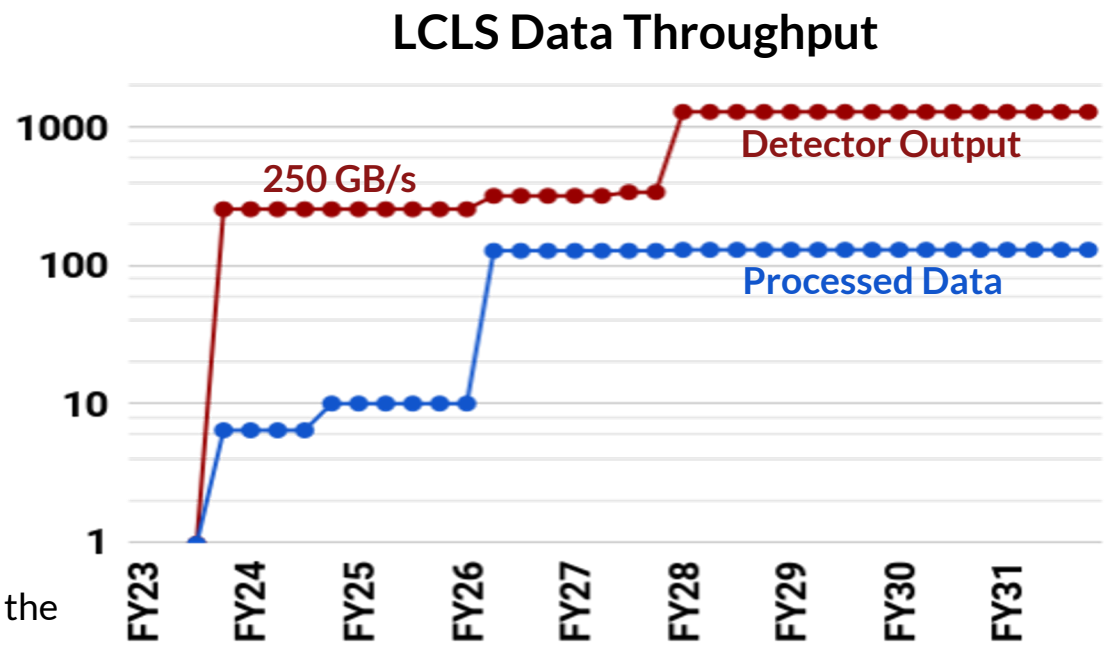
LCLS-II Upgrade: greater data *velocity*, *volume*, and *complexity*

Data Rates:
120 Hz to 1 MHz (10000x)

Raw Data Rates:
2 GB/s to 200 GB/s (100x)

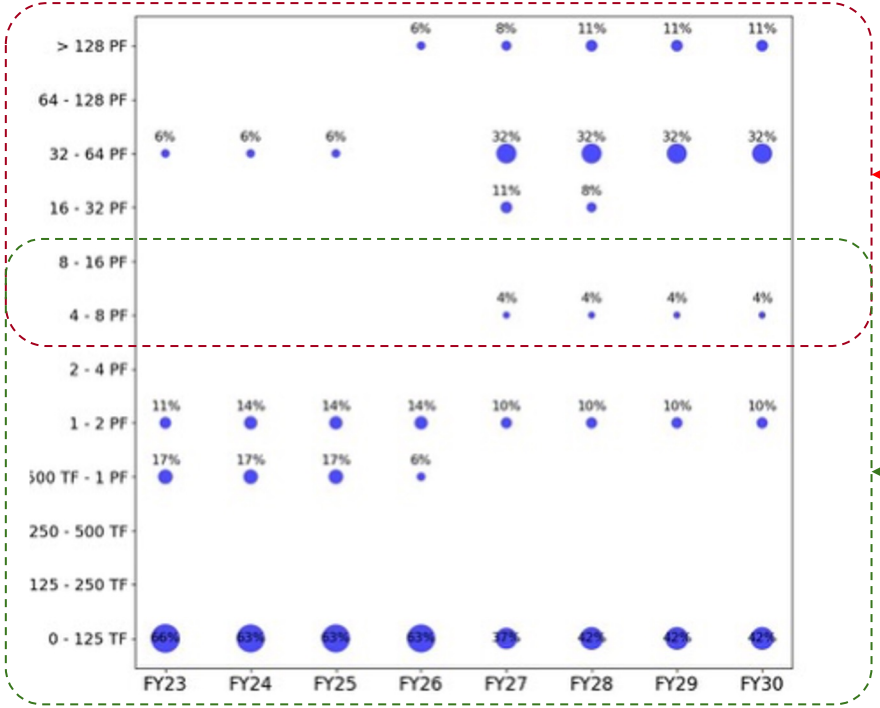
Recorded Data Rates:
2 GB/s to 20 GB/s (10x)

Recorded Data Volumes:
~1 PB per 12 hour shift
5 - 10 PB per 5 day experiment
50 - 100 PB per year aggregated for the facility, and growing.



Challenge: Why we need High End Computing

Strategy: maintaining critical capabilities at SLAC to cover majority experiments and for fast feedback while surging highest demand experiment to NERSC/LCF



Surge to offsite (NERSC & LCF)

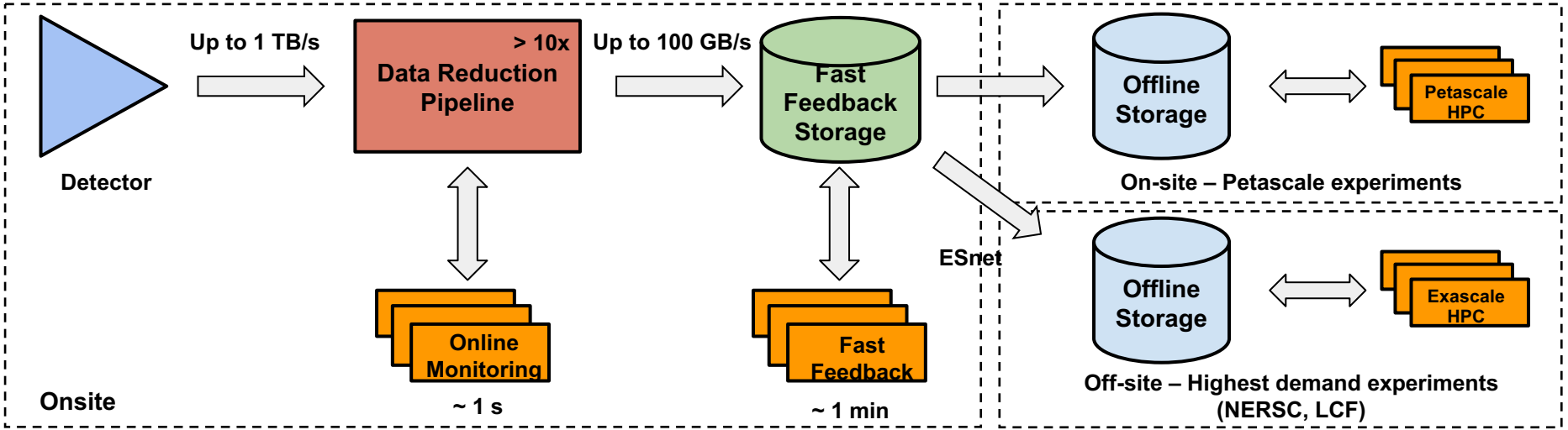
Local resources

Computational Requirements:
80% ~1 PF, 20% ~1 ExaFLOP

SLAC

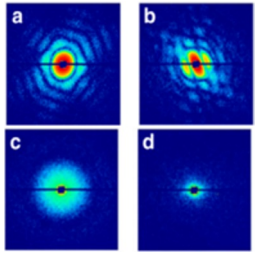
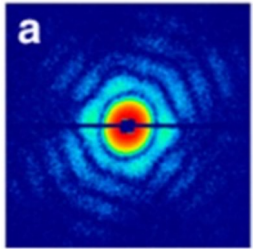
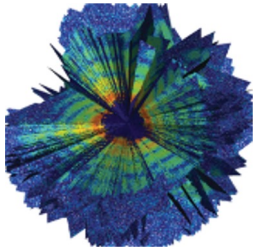
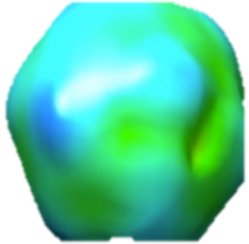
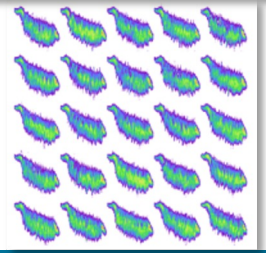
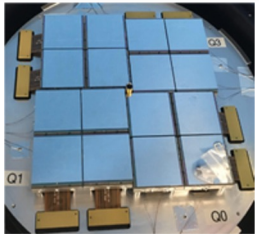
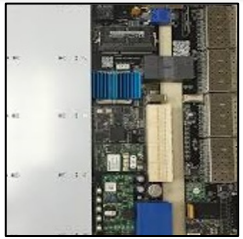
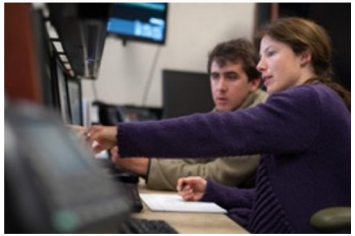
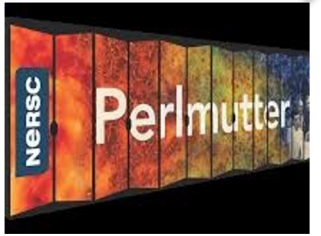
LCLS Data System, a scalable, adaptable system

Mix of automatic, on-demand, and user driven data flows - combination of onsite and offsite resources



Example of Massive Throughput Workflow: Coherent Imaging

One workflow must encompass several areas and disciplines: integrated approach required

Accelerator Diagnostics	X-ray Images	Instant Data Reduction	Assess Data Quality from Integrated Data	Interpretation of Structural Dynamics
				
				
Machine Learning Optimization	**High Frame Rate Detector**	**Edge Computing**	**Local Fast-feedback Computing**	**High End Computing**

Data Reduction at the Edge

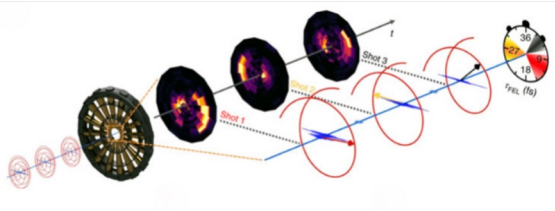
Produce actionable information with low latency for fast feedback and experiment steering

Fast ML at the Edge: Data Reduction for attosecond streaking

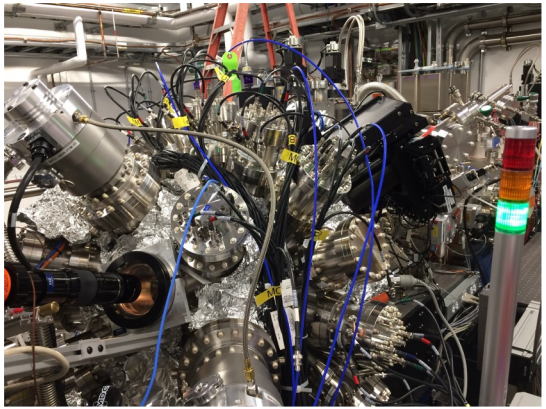
MRCO reconstructs attosecond pulses using ML at the Edge

Gain insight into attosecond electron dynamics:

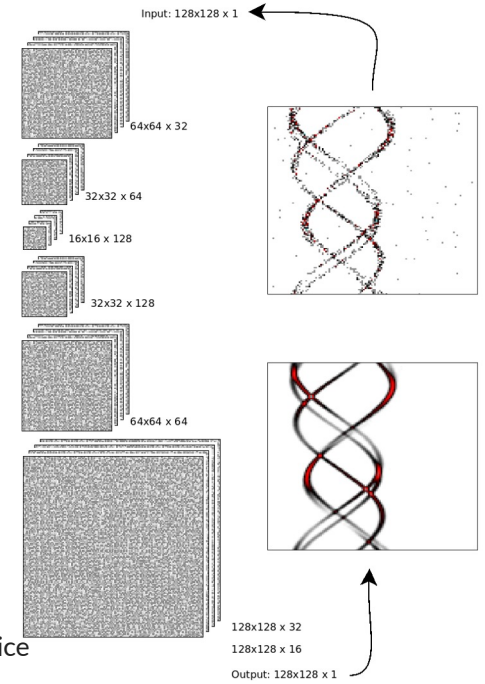
- MRCO/Cookiebox: Angle-resolved Electron Spectroscopy determines photoelectron angular distributions during photochemical processes



- Deploy AI inference in FPGAs: developed an AI inference library in High-Level Synthesis using SLAC Neural Net Library; enables high rate data processing & low latency feedback
- Implemented CookieNet feature extraction to reconstruct time-energy distribution of an attosecond FEL pulse in real-time to reduce 100 GB/s →~1 GB/s
- Implemented in FPGA used in LCLS Data Reduction Pipeline
- Demonstrated training and inference on Graphcore and SambaNova



MRCO/Cookiebox



SLAC 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 and FWP-35896.

ML in FPGA: SLAC Neural Network Library (SNL) Framework

Goal: Provide a set of libraries to synthesize AI inference networks into FPGAs

SNL implementation is targeting scientific instruments (frame rate of 100 kHz to 1 MHz) which must continuously adapt to new data and changing environments.

- Targeted at networks of a medium size, 10 - 20 layers, 100,000s of trainable parameters,
- Dynamic reloading of weights and biases to avoid re-synthesis.
  - Cannot re-synthesize for new training set; cannot risk FPGA implementation failing due to increase in resource usage , timing failure, or change to internal interconnect structure.
- High speed training is needed to support this as are real time bias and weight updates.

Features:

- Supports a Keras-like API for layer definition and configuration, modular and extensible
- Currently supported layer types: Conv2D, MaxPooling, AveragePooling, Dense, Reservoir.
- Current activators: LeakyRelu, Relu

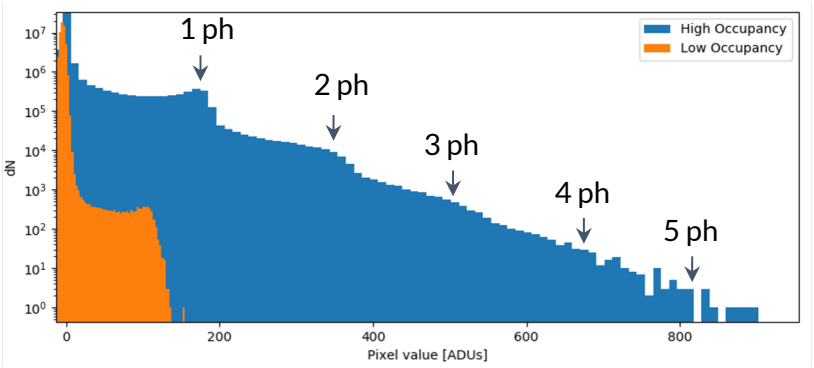
To Do: Quantization, attention layers for transformers (foundation models), global optimization suggestions

Smart Sensors: SparkPix-S and SparkPix-RT

Detectors with sparsified readout at ASIC enable leap from 100 kHz detector rates to 1 MHz

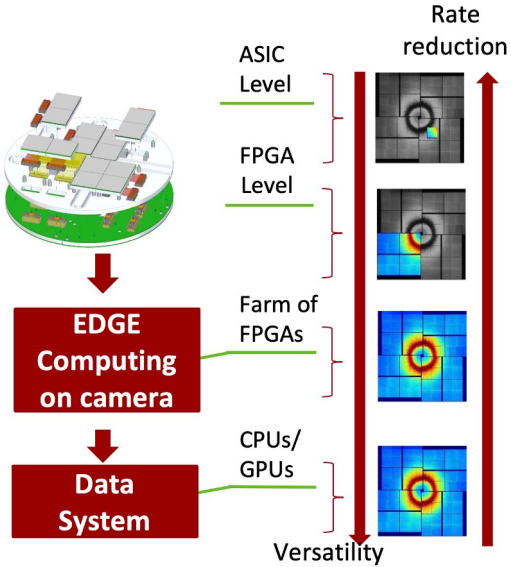
SparkPix-S: Pixel-threshold

- Information in both XPCS and XSVS experiments is **"sparse"** and confined in a limited # of pixels/frame, each pixel containing a limited # of photons
- 2D detector with fine spatial resolution, operating at the full rate of the machine, and discriminating between 0, 1, 2, 3.... photons/pixel/frame with high QE



SparkPix-RT

- Solve data transmission bottleneck by implementing compression algorithm solutions in ASIC
  - bit-level compression
  - auto-correction techniques (pedestal)
- R&D needed to deal with calibration and segmentation



SLAC

Better Science through ML

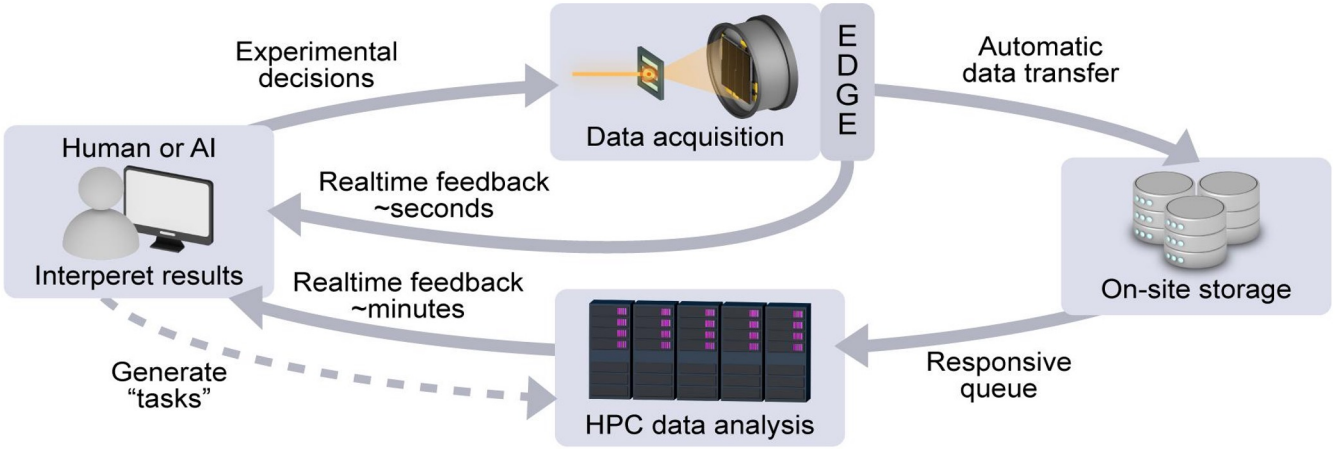
Lower the barrier to doing science through a unified approach from sensors to the data center

Figures: Greg Stewart at SLAC

Use ML to analyze data at the rate the production (1 MHz)

Analyze data at the rate of production using ML and providing access to network and compute

- Introduce AI/ML feature extraction at the edge to produce actionable information to feed experiment steering decision making mechanisms.
- AI-assisted decision making (running offline) uses analyzed information and other inputs to steer experiment.
- Embrace the use of heterogeneous pipelines (FPGA, CPU, GPU) and make them flexible, resilient, and transparent to use and configure

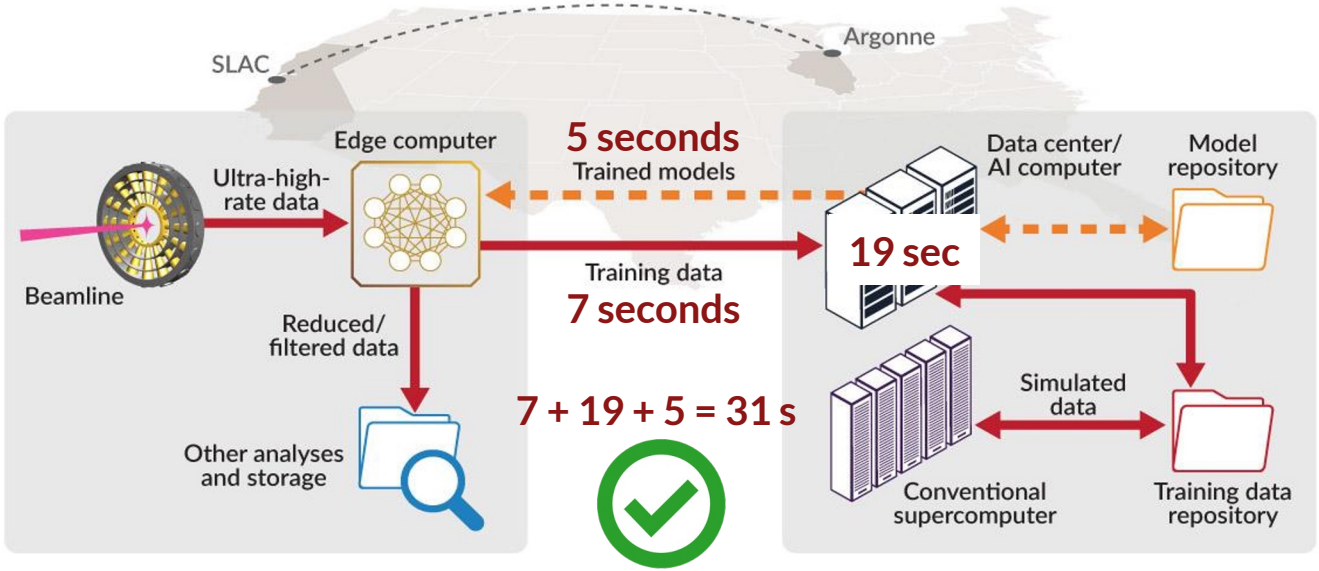


More good information, faster → better decisions → better data → experiment success!

Connect scientific instruments and HPC to create **smart instruments**

Provide actionable information by developing on-the-fly inference at the edge using ML trained remotely on streamed data - rapid (re)training workflows

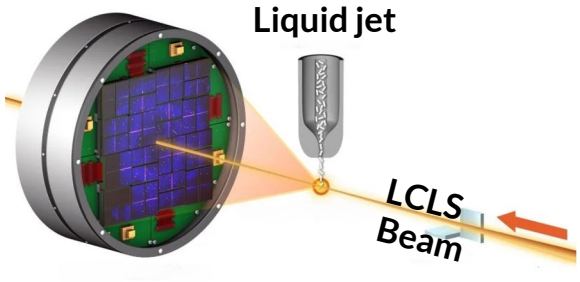
AI/ML at the Edge can introduce new, compute-intensive workflows, such as those required to re-train a model on streaming experimental data. Experiment conditions can change within 1000 seconds, so rapid re-training necessary.



This material is based on work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Award Number FWP-100643.

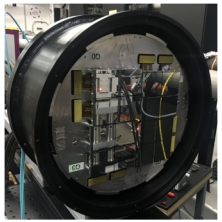
1st generation DRP: Veto for Crystallography and Single Particle Imaging

Experiment Description



- Individual nanocrystals are injected into the focused LCLS pulses
- Diffraction patterns are collected on a pulse-by-pulse basis
- One exposure per crystal
- Each image processed independently
- Crystal concentration dictates "hit" rate

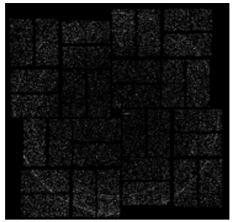
Megapixel Detector



- 4 MP@5 kHz in 2024
- 16 MP@40kHz in 2028

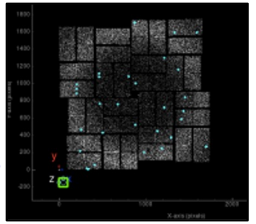
60 GB/s → 1 TB/s

X-ray diffraction image

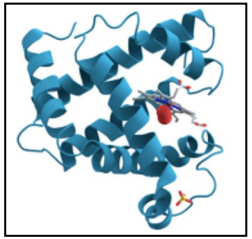


6 GB/s → 100 GB/s

Intensity map from multiple pulses



Interpretation of system structure / dynamics



Data Reduction
- Remove"no hits"
- >10x reduction

3 TFlops
16 TFlops

Data Analysis
- Bragg peak finding
- Index / orient patterns
- Average
- 3D intensity map
- Reconstruction

4 PFlops
20 PFlops

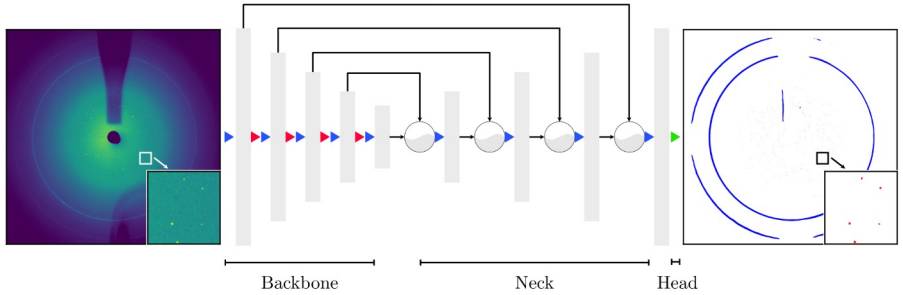
autocorrection, calibration

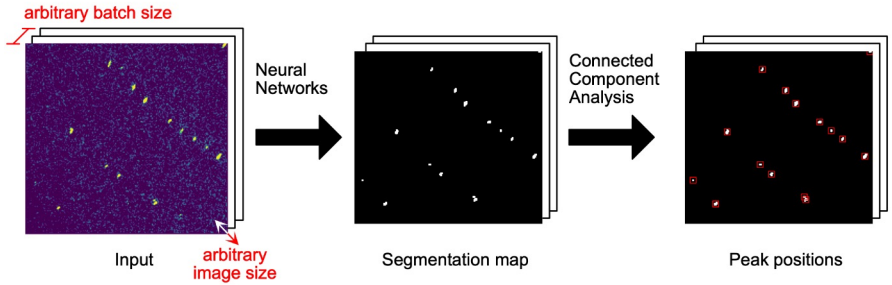
PeakNet

Indexing, averaging, 3D intensity map, reconstruction

Next generation DRP: write peaks

PeakNet: A 1 MHz AI-based Autonomous Bragg Peak Finder





PeakNet: A neural network for autonomous Bragg peak detection in real-time serial crystallography eliminates manual tuning, adapts in real-time to shot-to-shot background changes, and offers fast processing for high data rates.

Wang, C. et al., 2023 (https://doi.org/10.48550/arXiv.2303.15301)
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.

Significance and Impact

- *Use PeakNet in Data Reduction Pipeline to write peaks instead of raw images to disk.*
- PeakNet is a deep neural network for
  - Autonomous Bragg peak detection in real-time
  - Adapts in real-time to shot-to-shot background changes without manual tuning

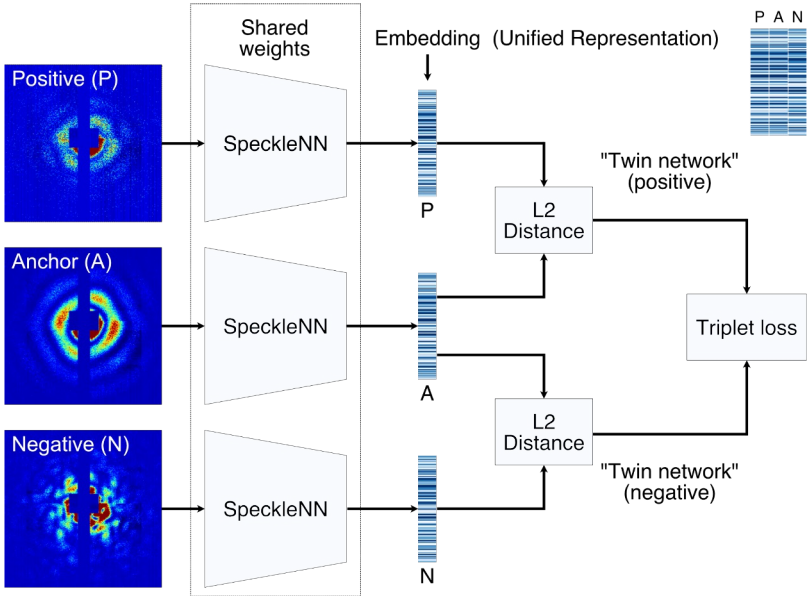
Features

- Autonomously executes pixel segmentation into 1) Bragg peaks, 2) artifact scattering, and 3) background, requiring no user parameter tuning.
- Our model, based on an attention U-Net architecture, minimizes focal loss during segmentation, accurately identifying true Bragg peaks and filtering out false peaks from artifact scattering, all without manual masking.

A modular PeakNet under development

Transitioning to a "RegNet + BiFPN + Segmentation head" architecture.
RegNet offers flexible pre-trained backbone options (e.g., ResNet, MobileNet), with BiFPN enabling multi-scale feature fusion, aiding segmentation across different scales.

SpeckleNN: AI classification of SPI images at high data rates



SpeckleNN measures speckle pattern similarities by training a model using a contrastive approach where three samples are used at a time. It learns to associate identically labeled (anchor and positive) images together and dissociate differently labeled (anchor and negative) images.

Wang, C. et al., 2023 (https://doi.org/10.48550/arXiv.2302.06895)
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.

Significance and Impact

- *Real-time data vetoing potentially reducing raw data volume and disk storage by 95%.*
- Classification of single-hit diffraction patterns for single particle imaging with limited labeled examples
- Overcomes high rate performance bottleneck: the need for speckle pattern labeling by a human for training.

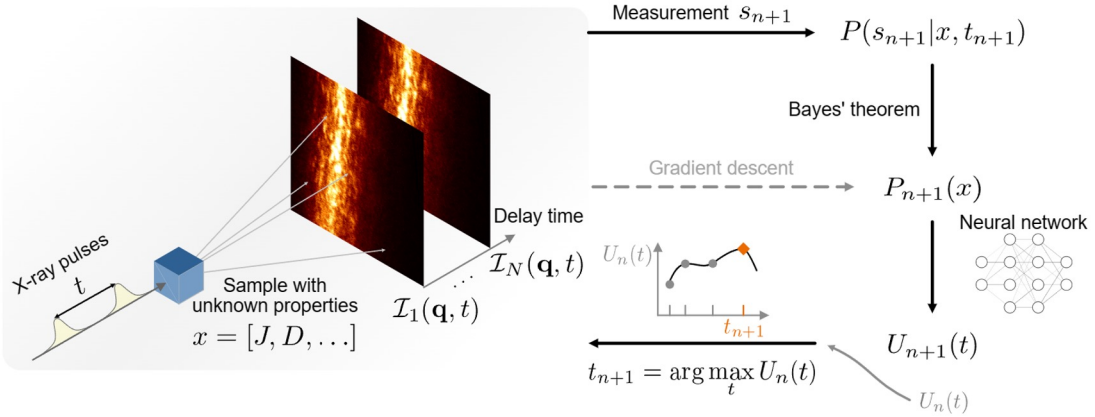
Features

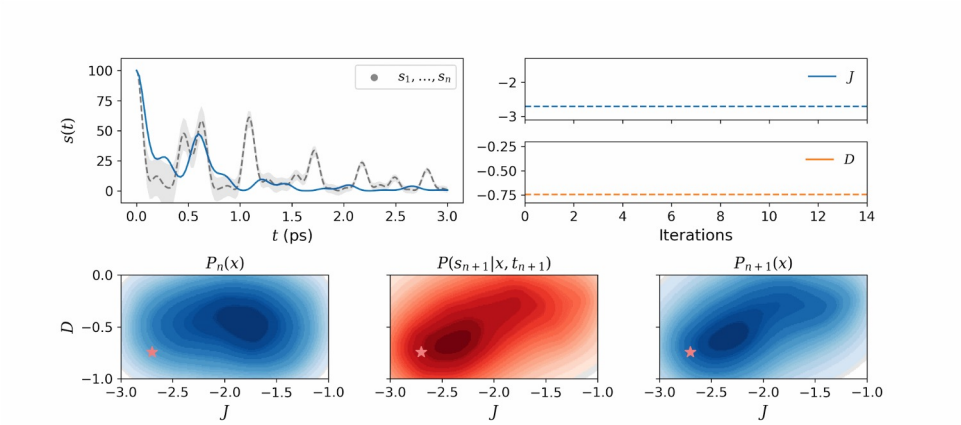
- Our model allows a flexible selection of vision backbones. A LeNet-like compact backbone (64K parameters) also delivers good performance (94% accuracy, 92% F1 score in predicting single-hit). Its small size makes it particularly amenable to deployment on FPGA devices.
- Our model maintains high performance even with only a fraction of an image available.

Experiment Steering

Actionable information produced at each layer of computing feeds decision-making algorithms that can drive experiments over seconds, minutes, or hours

Machine learning enabled real-time experiment steering





- **Help users make physics-informed decisions during their beam time.**
- A combination of neural network and Bayesian optimal design for real-time decision making and parameter estimation.
- Neural networks are used as surrogate models for rapid calculations of utility function and posterior distribution.
- Application is simulated split-and-delay measurement in LCLS: a data-driven **experiment steering framework** suggests next measurement point, time delay *t*, that maximizes information gain

Chen, Z. et al., 2023 (https://doi.org/10.48550/arXiv.2306.02015)
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 DE-SC0022216.

More good information, faster → better decisions → better data → experiment success!

Summary

Advances in computational power and analysis methods that leverage massive data quantities will maximize the science output from LCLS.

LCLS is supporting the development of a data system infrastructure capable of handling the demands of Big Science:

- Real-time data analysis capabilities (data reduction, complex workflow orchestration)
- On-demand utilization of super-computing environments
- Strategic development of AI/ML for targeted applications
- Ability to automate experiments (execution to analysis)

Many thanks to the people doing the honest work: Ric Claus, Ryan Coffee, Dan Damiani, Gabriel Dohrliac, Chris Ford, Mikhail Dubrovin, Ryan Herbst, Wilko Kroeger, Xiang Li, Stefano Marchesini, Valerio Mariani, Riccardo Melchiorri, Silke Nelson, Chris O'Grady, Amedeo Perazzo, Frederic Poitevin, Thorsten Schwander, Murali Shankar, Monarin Uervirojnangkoorn, Matt Weaver, Seshu Yamajala, Cong Wang, Zhantao Chen