Massive-scale Data Analytics at the Linac Coherent Light Source

CHEP 2023, Norfolk

Jana Thayer/LCLS Data Systems Division Director

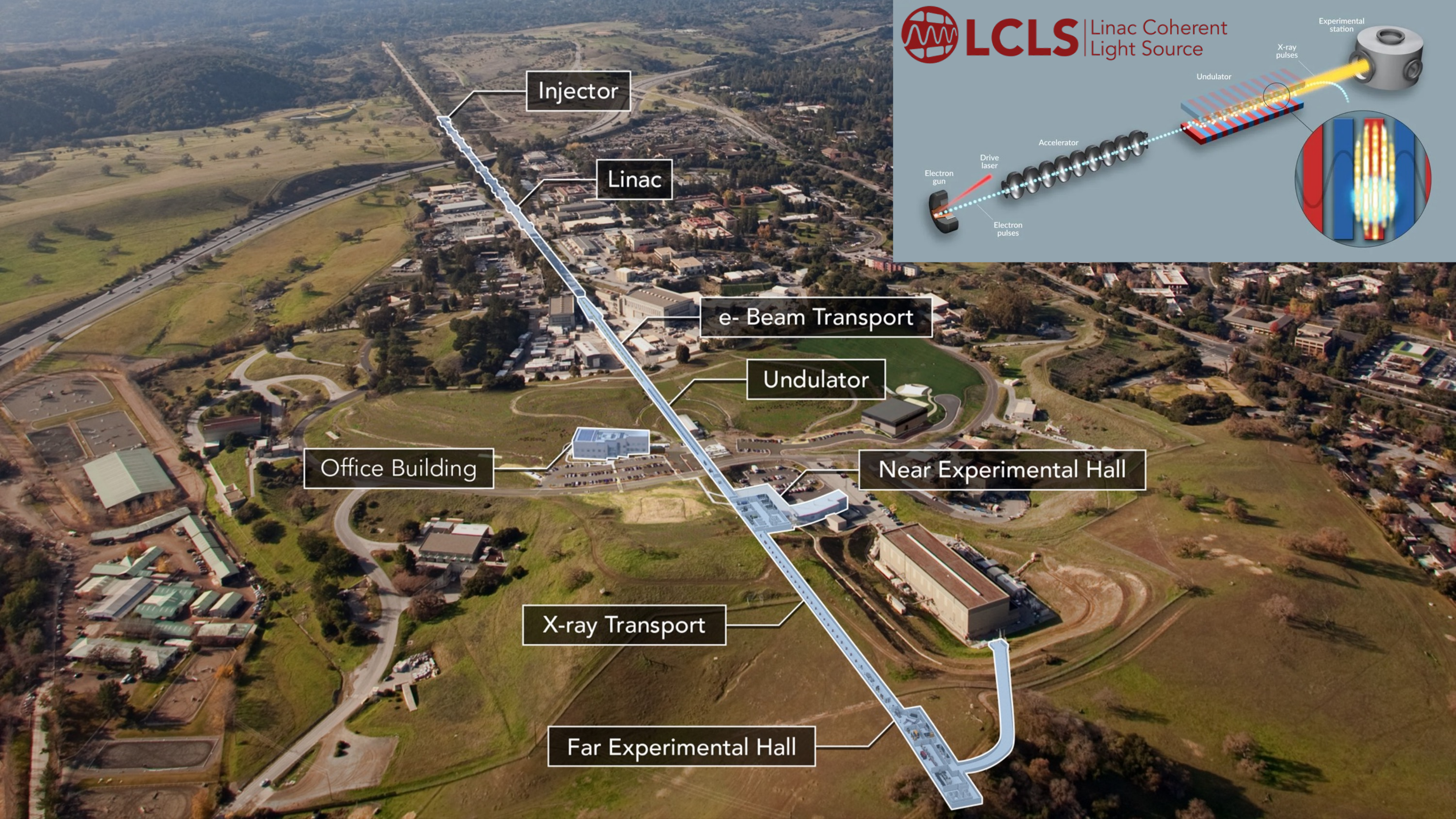
May 11, 2023

Many thanks to the people doing all the honest work:
Ric Claus, Dan Damiani, Chris Ford, Mikhail Dubrovin, Victor Elmir, Wilko Kroeger, Xiang Li, Valerio Mariani, Silke Nelson, Ariana Peck, Frederic Poitevin, Chris O'Grady, Julieth Otero, Omar Quijano, Murali Shankar, Monarin Uervirojnangkoorn, Matt Weaver, Seshu Yamajala, Chuck Yoon, Cong Wang, Zhantao Chen

Outline

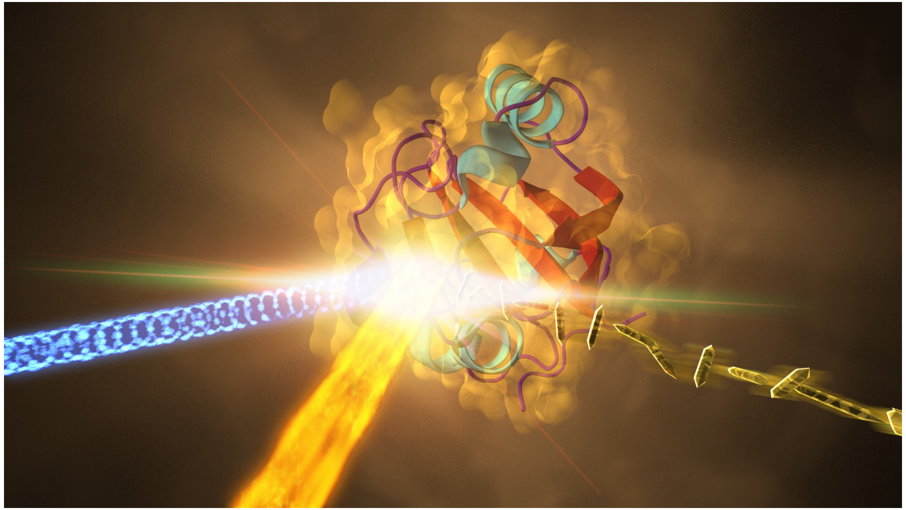
New infrastructure and analysis methods that leverage massive data quantities will maximize the science output from LCLS

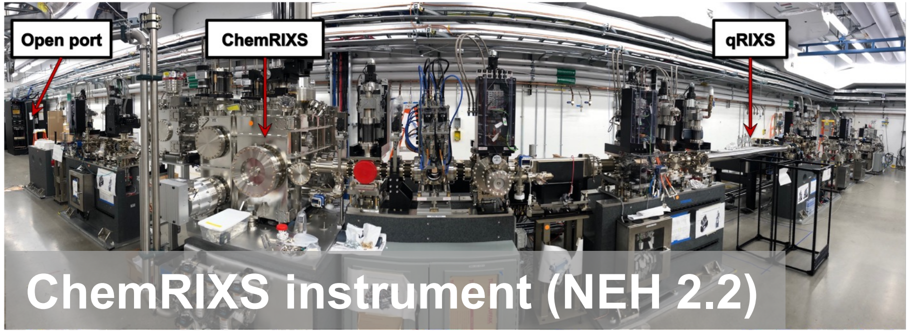
- About LCLS
- Challenges for LCLS Data Analytics
- Patterns for workflows
  - On-the-fly data reduction: Data Reduction Pipeline
  - Heterogeneous pipelines: Building intelligent, adaptable detector systems
  - Quasi-real-time workflows with results within minutes
    - Compute-intensive streaming analysis of Tb/s data streams
    - Adaptable real-time ML workflows
- The future: Integrated Research Infrastructure and HPDF
- Conclusions

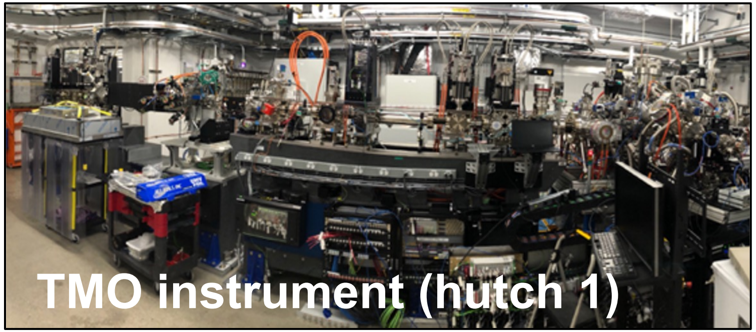

LCLS
Linac Coherent Light Source
Experimental station
X-ray pulses
Undulator
Accelerator
Drive laser
Electron gun
Electron pulses
Injector
Linac
e- Beam Transport
Undulator
Office Building
Near Experimental Hall
X-ray Transport
Far Experimental Hall

LCLS-II: SLAC's largest data and computing challenge

- Ultrafast x-ray pulses from LCLS are used like flashes from a high speed strobe light, producing stop action movies of atoms and molecules.
- Both data processing and scientific interpretation demand intensive computational resources.
- Characterized by time-sensitive and data integration-intensive workflows



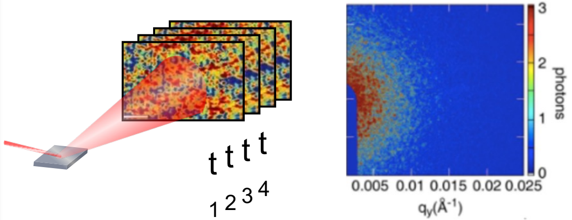




Many compute-intensive workflows, one scalable data system

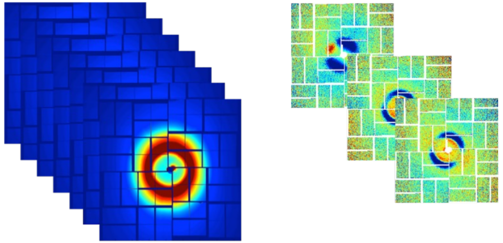
Coherent Scattering

XPCS XSVS



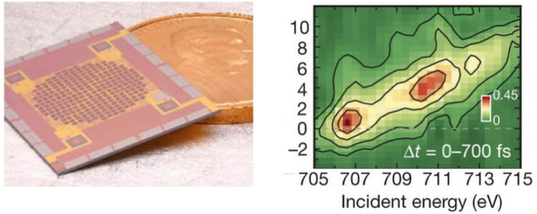
2023: 20 GB/s, 4 TF (reduction), 34 TF (analysis)
2026: 80 GB/s, 34 TF (reduction), 270 TF (analysis)

Liquid scattering



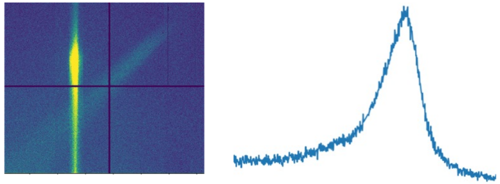
2023: 40 GB/s, < 1 TF (reduction), 20 TF (analysis)
2026: 320 GB/s, <1 TF (reduction), 50 TF (analysis)

Resonant Inelastic Scattering



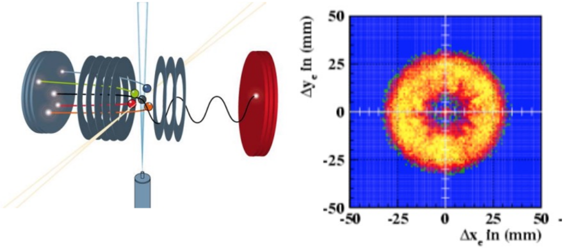
2023: 20 GB/s, 4 TF (reduction), 1 TF (analysis)
2026: 200 GB/s, 40 TF (reduction), 2 TF (analysis)

X-ray Emission Spectroscopy (XES)



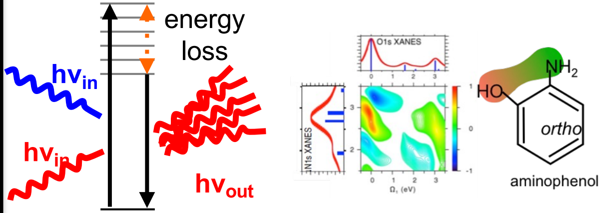
2023: .1 GB/s, < 1 TF (reduction), .2 TF (analysis)
2026: 10 GB/s, <1 TF (reduction), 20 TF (analysis)

Coincidence Spectroscopy



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

Nonlinear Spectroscopy



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

SLAC

Data Life Cycle and the Compute Needs of the Typical User

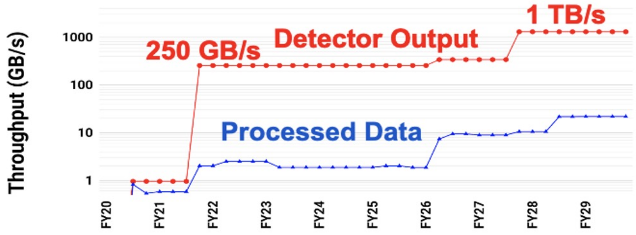
A Day in the Life of a User

- Typical experiments last 5 days of 12 hour shifts.
- During **data collection**:
  - Must be able to get real-time (~1 sec) **feedback** and the **quality of data-taking**, e.g.
    - Are we getting all the required detector contributions for each event?
    - Is the hit rate for the pulse-sample interaction high enough?
  - Must be able to get **feedback** about the **quality of the acquired data** with a latency lower (~1 min) than the typical lifetime of a measurement (~10 min) in order to optimize the experimental setup for the next measurement, e.g.
    - Are we collecting enough statistics? Is the S/N ratio as expected?
    - Is the resolution of the reconstructed electron density what we expected?
- During **off shifts**: must be able to run **multiple passes** (> 10) of the full analysis on the data acquired during the previous shift to optimize analysis parameters and code in preparation for the next shift
- During the **4 months** after the experiment, must be able to analyze the raw and intermediate data on **fast access storage** in preparation for publication
- **After 4 months**, must be able to **restore** archived data to test new ideas, new code, new parameters

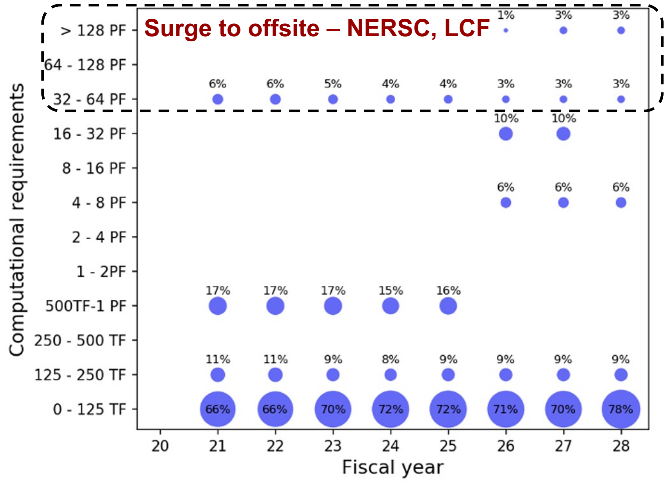
LCLS Data Challenges

- **LCLS-II Upgrade:** greater data *velocity* and *volume*
  - Data Rates: 120 Hz to 1 MHz (10000x)
  - Raw Data Volumes: 2 GB/s to 200 GB/s (100x)
  - Recorded Data Volumes: 2 GB/s to 20 GB/s (10x)
  - Computational Requirements: 80% ~1 PF, 20% ~1 ExaFLOP
- **Fast Feedback:** real-time analysis (seconds/minutes) is essential to the users' ability to make informed decisions during an LCLS experiment.
- **Variability:**
  - Wide variety of experiments with turnaround ~days
  - Large dynamic range: device readout 0.01 Hz - 1 MHz
  - Data Complexity: Variable length data (raw, compressed)
  - Access patterns to data vary by experiment and detector
  - Analysis is a mix of tried-and-true & innovative techniques
- **Time to Science:** Development cycle must be fast & flexible
- **No user left behind:** alleviate the pressure on users to gather resources to mount a significant computing effort.

LCLS Data Throughput



LCLS Computational Requirements

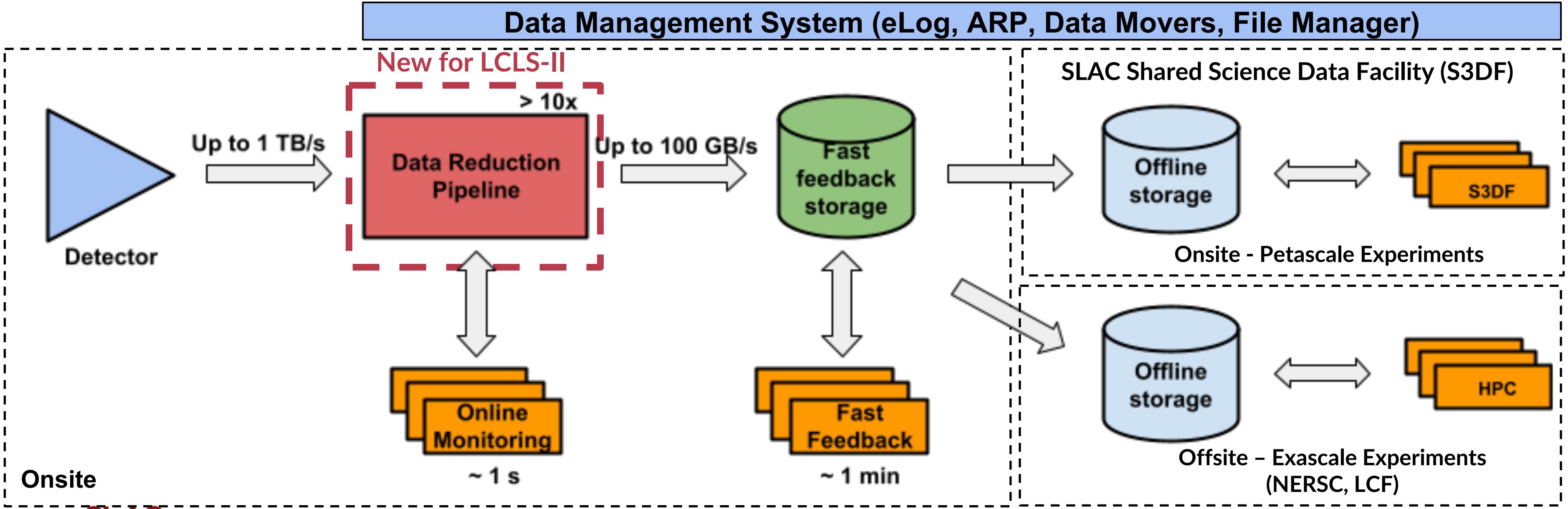


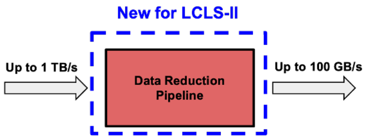
SLAC

LCLS Data System, a scalable, adaptable system

LCLS facility provides stewardship of core hardware/software infrastructure for data acquisition, data reduction, online monitoring, data storage and management, and offline analysis processing and framework.

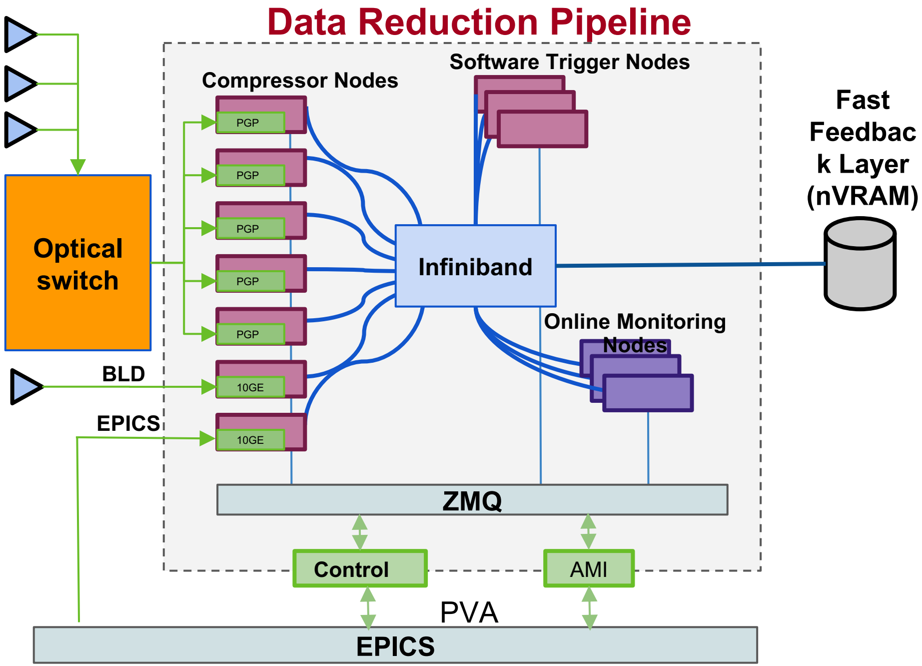
Users provide the last mile: develop their own analysis on top of this stack.





LCLS Data Reduction Pipeline (DRP)

All experiments reduce data during processing → Data Reduction Pipeline does it in real time



Data reduction toolbox of parameterized algorithms runs on DRP compute layer: **compression, feature extraction, trigger/veto, multi-event reduction**

Software Trigger Nodes perform online event-build collecting data from multiple detectors from the same event. Two decisions per event, per shot: **Store or not? Send data to online monitoring?**

100 Hz stream of “leaked”, unreduced events accompanies reduced data.

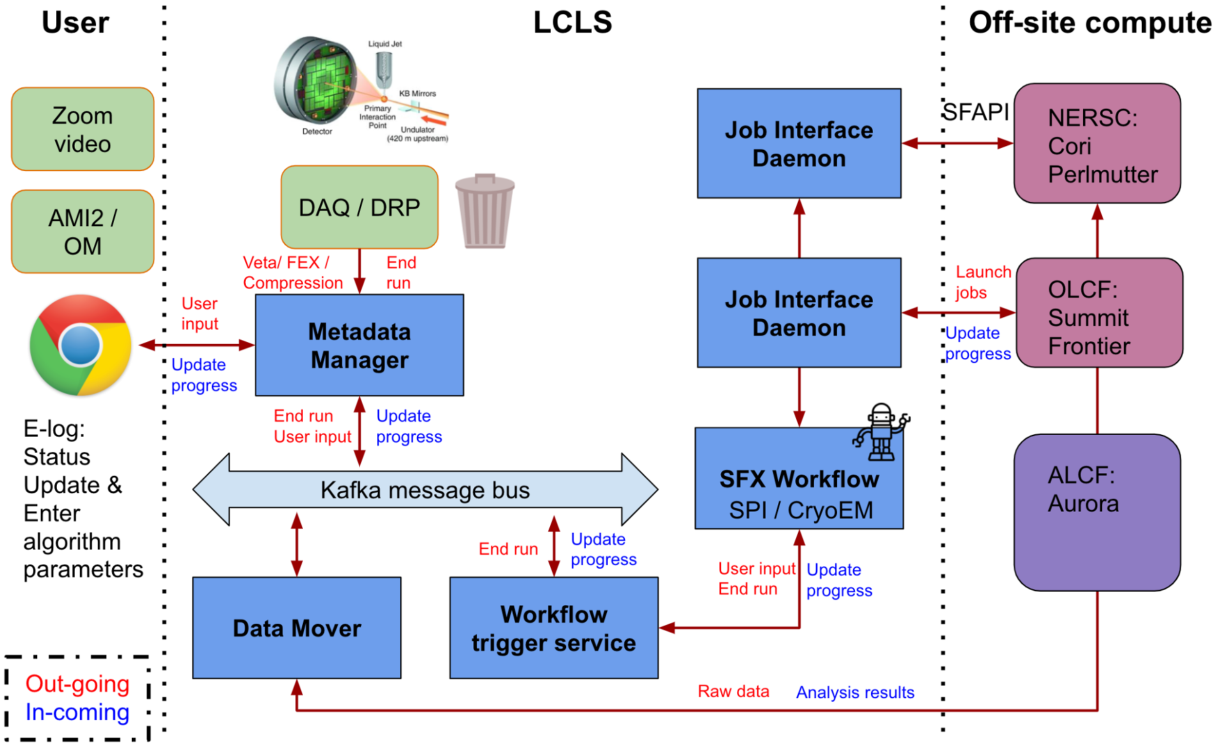
CPU and FPGA available in DRP .

Status: Successfully acquiring 200 GB/s at 1 MHz with data reduction

SLAC

Automated Data Movement and Run Processing

Automated data/metadata capture and movement; configurable and automated workflows



https://doi.org/10.1107/S1600577520004257

SparkPix-RT: Intelligent Detector Systems for 1 MHz readout

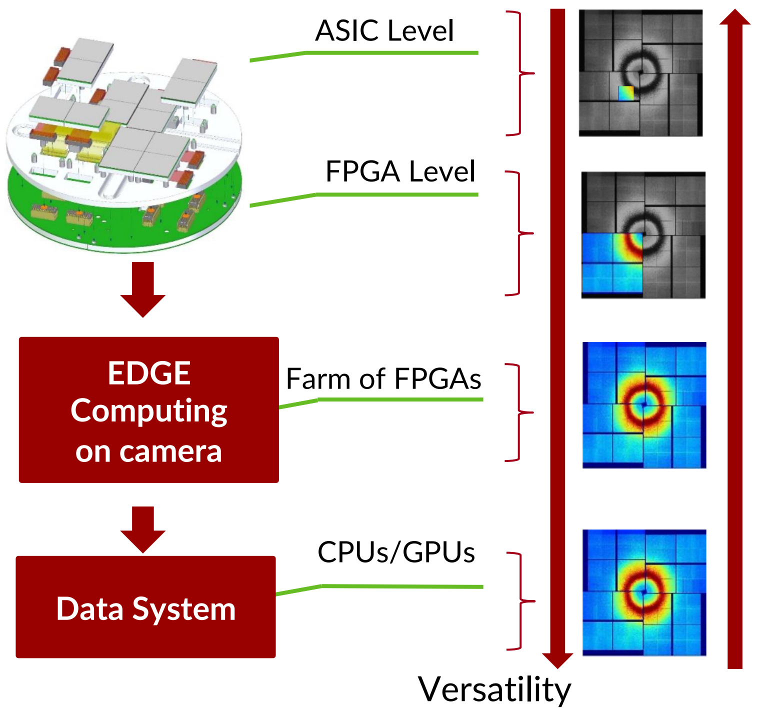
Triggering/Compression within ASIC reduces data volume, alleviates data transport bottleneck

LCLS needs streaming feature extraction on high rate data from large imaging detectors to enable smart, autonomous experiments

- Fast readout of the desired data
- Alleviate downstream network, storage, and computing bottlenecks

Enabling data reduction/feature extraction at the ASIC/FPGA level is challenging because ASICs/FPGAs may only see a fraction of the image (potentially uncorrected) while the offline analysis sees fully calibrated, complete images.

- Develop intelligent auto-correction techniques
- Provision for buffering and deadtime
- Develop triggering capability
- Determine what kind of information extraction is feasible
- Provide for data validation and model training



SLAC

Single Particle Imaging High Rate Analysis

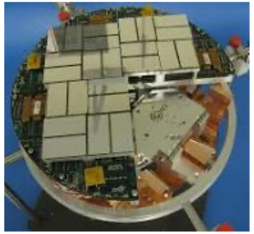
Determine the 3D molecular structure of nanoparticles (time-resolved, at room temperature)

Three key steps for high rate analysis workflow:

1. DRP: Fast Data Reduction using Veto. Hit or Miss? (> 10x reduction)
2. Accurate classification: Identify single hits
3. Reconstruction: Orientation recovery: phase retrieval

- Individual particles are injected into the focused LCLS pulses
- Scattering patterns are collected pulse-by-pulse
- Particle concentration dictates "hit" rate

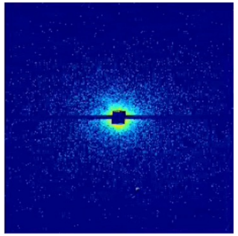
Multi megapixel detector



- 8 kHz in 2024 (4 MP)
- 40 kHz in 2027 (16 MP)

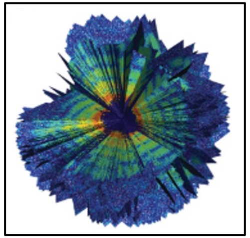
60 GB/s → 1 TB/s

Coherent scattering image

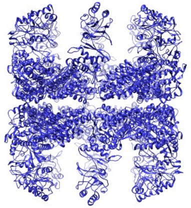


6 GB/s → 100 GB/s

Intensity map from multiple pulses



Interpretation of system structure / dynamics



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

3 TFlops
16 TFlops

Data Analysis
- Orient patterns
- 3D intensity map
- Phase retrieval

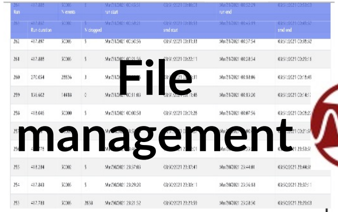
270 PFlops
1340 PFlops

https://doi.org/10.48550/arXiv.2206.11992

Quasi-real-time analysis using NERSC

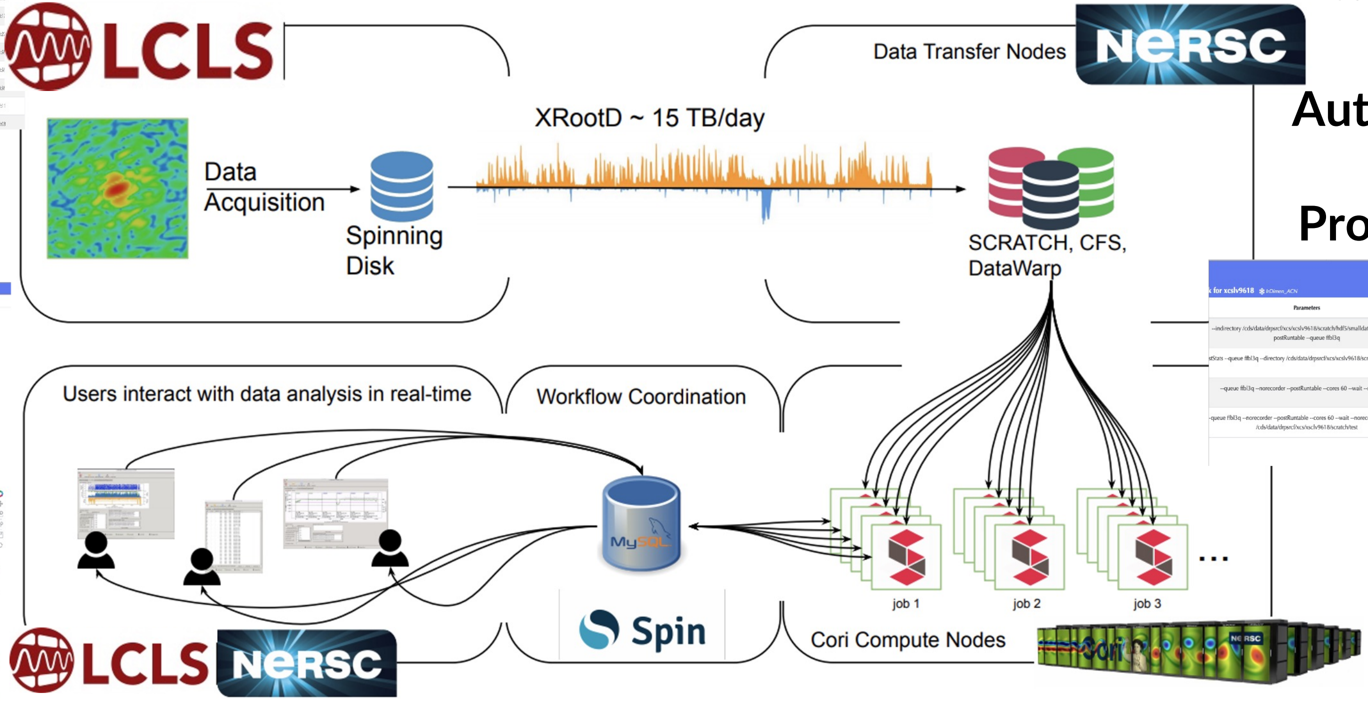
ExaFEL project streams data to NERSC for analysis results within minutes

File management

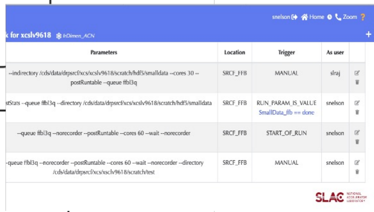


Transparent data movement

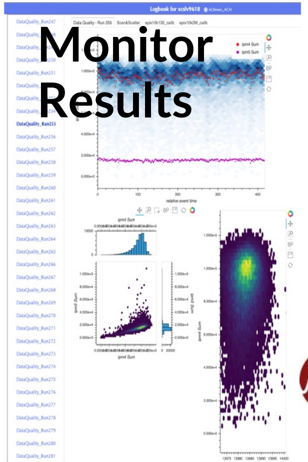
Testing data transfer to other facilities



Automated Run Processing



Monitor Results

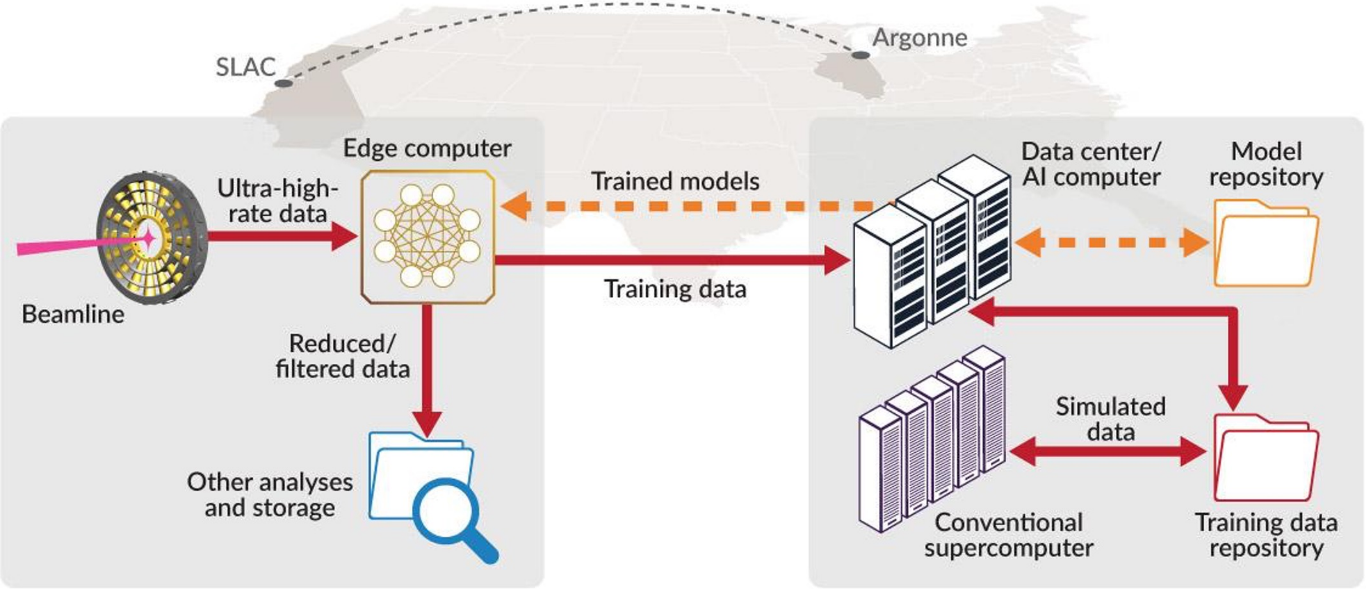


https://doi.org/10.48550/arXiv.2105.13967

Actionable Information from Sensor to Data Center (AISDC)

Providing Actionable Information by linking AI/ML at the Edge with HPC

Implement AI/ML at the Edge; develop workflows that stream data to a specialized data center at Argonne to achieve a turnaround time between initialization and model delivery to Edge host of < 10 minutes

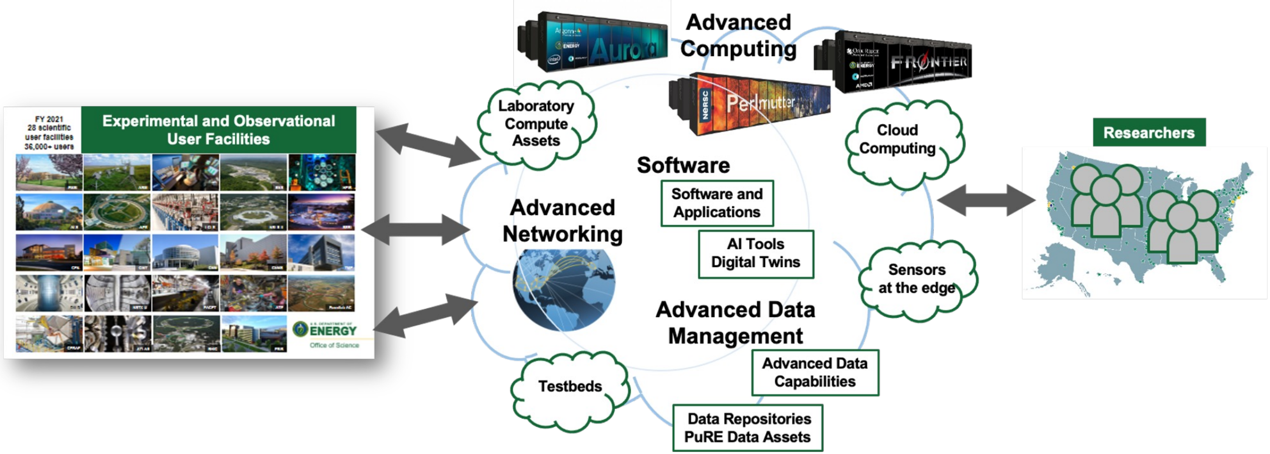


SLAC

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.

ASCR Defines a Unified Vision for Scientific Computing for Facilities

ASCR: IRI-ABA - Integrated Research Infrastructure Architecture Blueprint Activity



New modes of integrated science

Rapid data analysis and steering of experiments	Novel models for multi-facility allocation/utilization	AI-enabled insight from dynamic, vast multi-modal data	Seamless user interconnectivity via federated IDs

Vision: A DOE/SC integrated research ecosystem that transforms science via seamless interoperability.

LAB 23-3020: The ***High Performance Data Facility (HPDF)*** will serve as a foundational element enabling the DOE Integrated Research Infrastructure.

S3DF would like to participate in IRI ecosystem as both a source of data and a compute facility serving 80% of LCLS experiments

Summary

LCLS Data Systems infrastructure scales to meet high-throughput, compute-intensive demand:

- Real-time data analysis capabilities (data reduction, complex workflow orchestration)
- On-demand utilization of super-computing environments
- Pipeline spans from Detector Edge to HPC
- Strategically developing AI/ML for targeted applications
- Intelligent detectors and real-time analysis enable autonomous experiment steering

Lessons learned:

- Modularity is key for rapidly adaptable workflows
- Seamless access to computing with transparent data movement is crucial for users
- Ease-of use is important: hide the computing complexity where possible
- ML at the edge is hard and creates new compute-intensive workflows
- A common software ecosystem is greatly desired, but difficult to achieve; users will always need some experiment-specific software

Backup

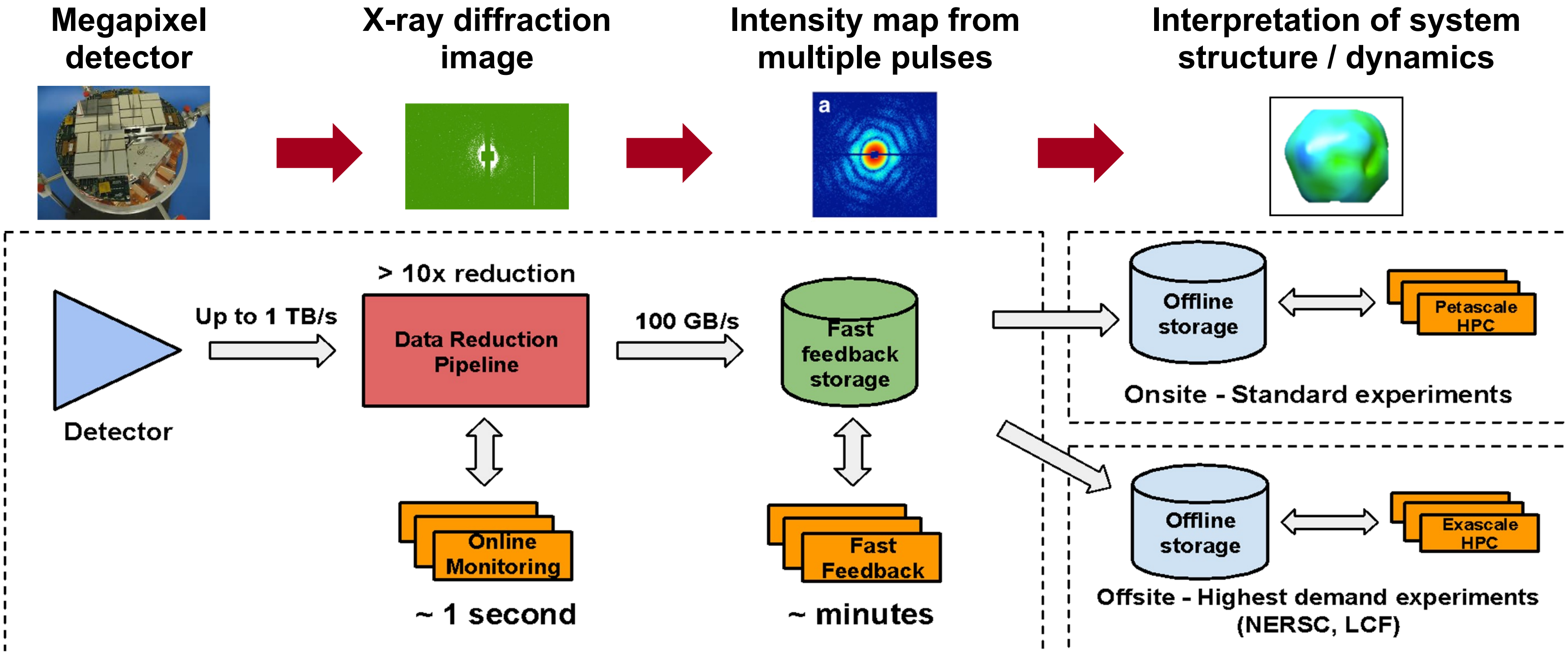
More details

LCLS and ATLAS: Similar but very different

Courtesy of Richard Mount (circa 2017)

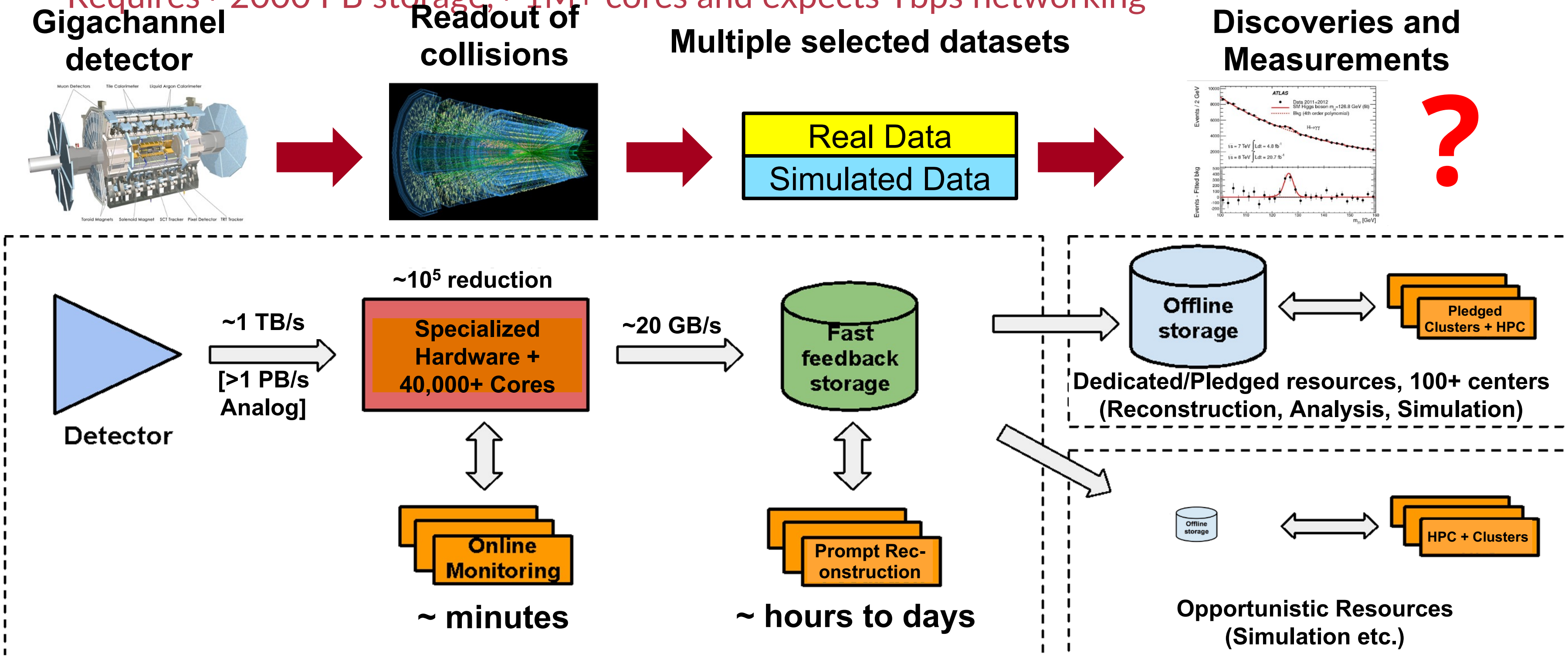
	LCLS-II 2022	LCLS-II 2029+	ATLAS Today	ATLAS 2026+
Wanted fraction of collisions	0.01 to 1.0	0.01 to 1.0	$< 10^{-6}$	$< 10^{-5}$
Typical experiment duration (same data-taking conditions)	3 days	3 days	3 years	3 years
24x7 availability of offline computing	Essential	Essential	Desirable	Desirable
Required turnround for data-quality checks	Seconds to minutes	Seconds to minutes	Hours to days	Hours to days
Raw digital data rate	200 GB/s	1000+ GB/s	160 GB/s	1,000 GB/s
Zero-and-Junk-suppressed rate	10 GB/s	100+ GB/s	1.5 GB/s	20 GB/s
Storage need dominated by	Mainly raw data		Mainly simulated and derived data	
Role of Simulation	Growing in science analysis Growing in experiment design		Vital in physics analysis Vital in experiment design	
Analysis, Simulation and Workflow Software development community	Individuals (in the past) → Organized effort		~100 organized collaborators (mainly research physicists)	

LCLS-II data challenge



Compare with HL-LHC (2029+)

Requires ~2000 PB storage, ~1M+ cores and expects Tbps networking

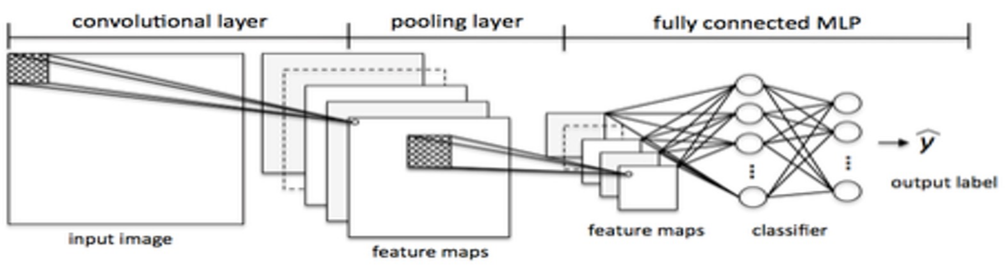


Courtesy Richard Mount, circa 2017

Progress: Actionable Information to reconstruct attosecond pulses

SLAC NeuralNet Library (SNL) framework enables deployment of AI inference in FPGAs

- Developed a structured AI inference library in High-Level Synthesis which enables high rate data processing & low latency feedback by deploying AI inference in FPGAs.
- Target AI networks with 10+ layers
- Implement CookieNet feature extraction to reconstruct time-energy distribution of an attosecond FEL pulse in real-time (at 1 MHz) in FPGA to reduce 100 GB/s →1 GB/s
- Support rapid deployment of weights and biases → networks to adapt to changing experimental conditions



Feature extraction followed by classification

Deployment will consist of PCIE based FPGAs in rack mount servers

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.