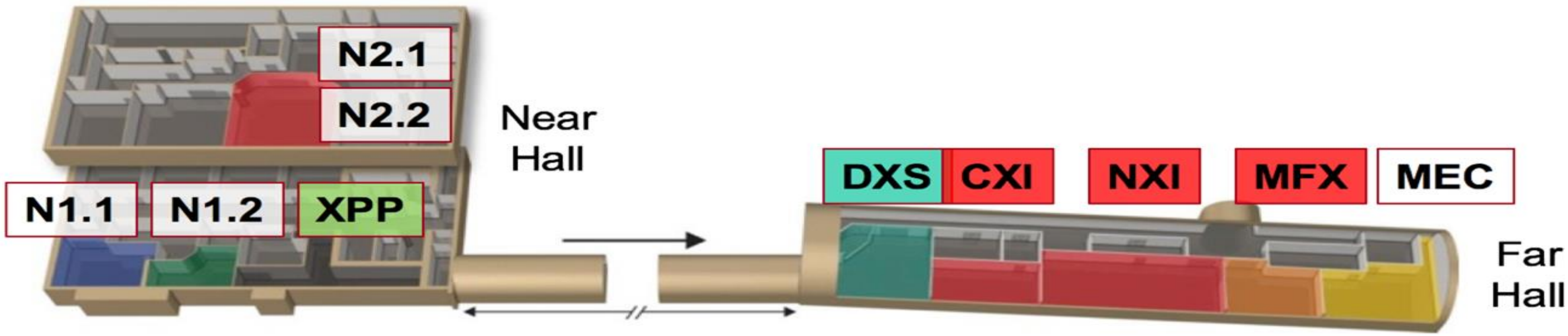
Building a Data System for LCLS-II

Jana Thayer, IFDEPS, March 13th 2018, Annecy, France

LCLS-II, a major (~ B$) upgrade to LCLS is currently underway. Online in 2020. Repetition rate will increase from 120 Hz to 1 MHz.

LCLS–II and –HE X-ray instruments, detectors, and data systems

LCLS–II and –HE require a new suite of X-ray instruments, detectors, and data systems, consistent with the leap from 120 Hz to 1 MHz



LCLS-II instrument development (underway)

Instrument	Photon Energy	Detector Needs	First Light
NEH 1.1	250-2500 eV	2D ToF Charged Particle (1 MHz) 2D ToF Multi-Particle	11/2020
NEH 2 (LJE)	250-1600 eV	2D High Spatial Resolution (5 µm) TES - 1000 pixel (≤1 eV, ≥10 kHz)	11/2020
NEH 2 (RIXS)	250-1600 eV	2D High Spatial Resolution (5 µm) 2D Imaging (≥ 2 kHz)	1/2022
NEH 1.2	400-6000 eV		1/2023

Early LCLS-II Facility Detectors and Readout Rates

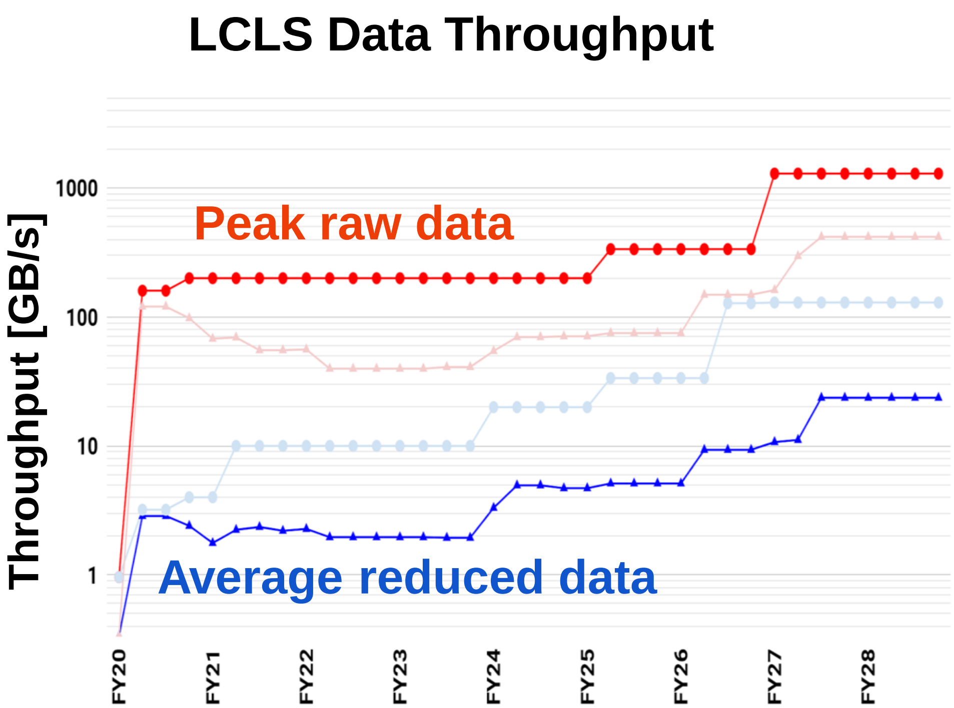
Application	Detector	Detector Size	Detector Rate	Data Rate (GB/s)	Year
Spectroscopy	TES	1000 pixels, 1-2 MHz sampling	100 kHz	20 - 40*	2021
	RIXS-ccd	4096 pixels	1 kHz	< 1	2020
Scattering/Imaging tender/hard	epix10k	various sizes	120 Hz	< 1	2018
	epixHR	4 MPixel	5 kHz	40	2023
	Jungfrau	4 MPixel	2 kHz	16	2018
	Hard X-ray Detector	1 MPixel	120 Hz	< 1	2022
	Very High Frame Detector	4 MPixel	40 kHz	320	2027
Scattering/Imaging soft	epixM/vfccd/FLORA	4 MPixel	10 kHz	80	2021
	Very high frame detector	4 MPixel	40 kHz	320	2023
Particle Detector	Digitizer: 20 channels, 5 GHz sampling	20 ch x 50,000 points	100 kHz	200	2020
	MCP Delay-line		100 kHz	< 1 GB/s	2020
	Tixel/Particle Detector	0.5 - 1 MPixel	>1 kHz	< 1 GB/s	2023

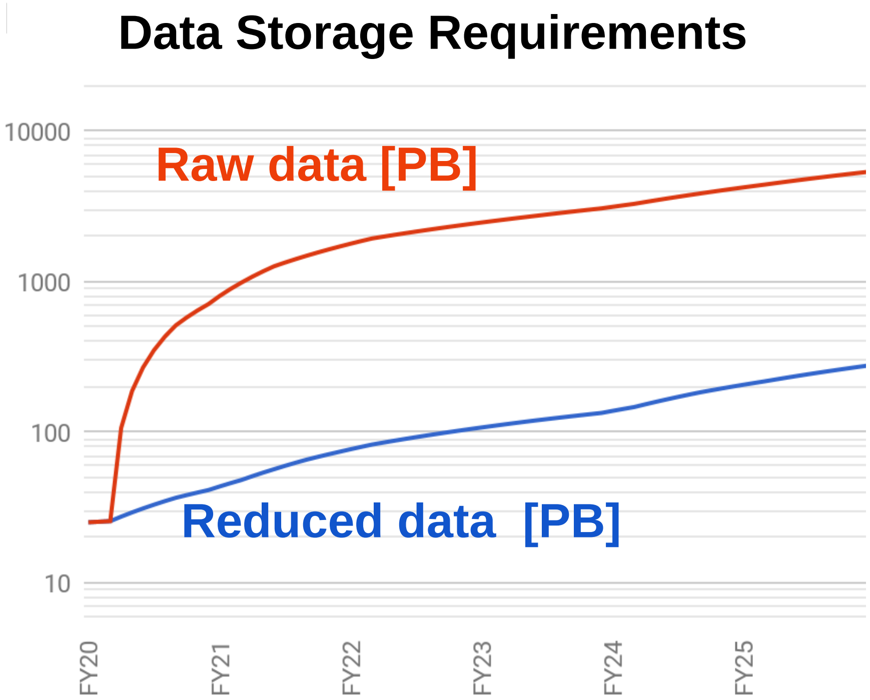
LCLS computing has some challenging characteristics

1. **Fast feedback** is essential (seconds / minute timescale) to reduce the time to complete the experiment, improve data quality, and increase the success rate
2. **24/7 availability**
3. **Short burst** jobs, needing very short startup time
   Very disruptive for computers that typically host simulations that run for days
4. **Storage** represents significant fraction of the overall system, both in cost and complexity
5. **Throughput** between storage and processing is critical
   Currently most LCLS jobs are I/O limited
6. Speed and flexibility of the **development cycle** is critical
   Wide variety of experiments, with rapid turnaround, and the need to tune data analysis during experiments (20+ unique workflows identified)

LCLS is uniquely challenging due to the data throughput, the variety of experiments and the need for fast feedback

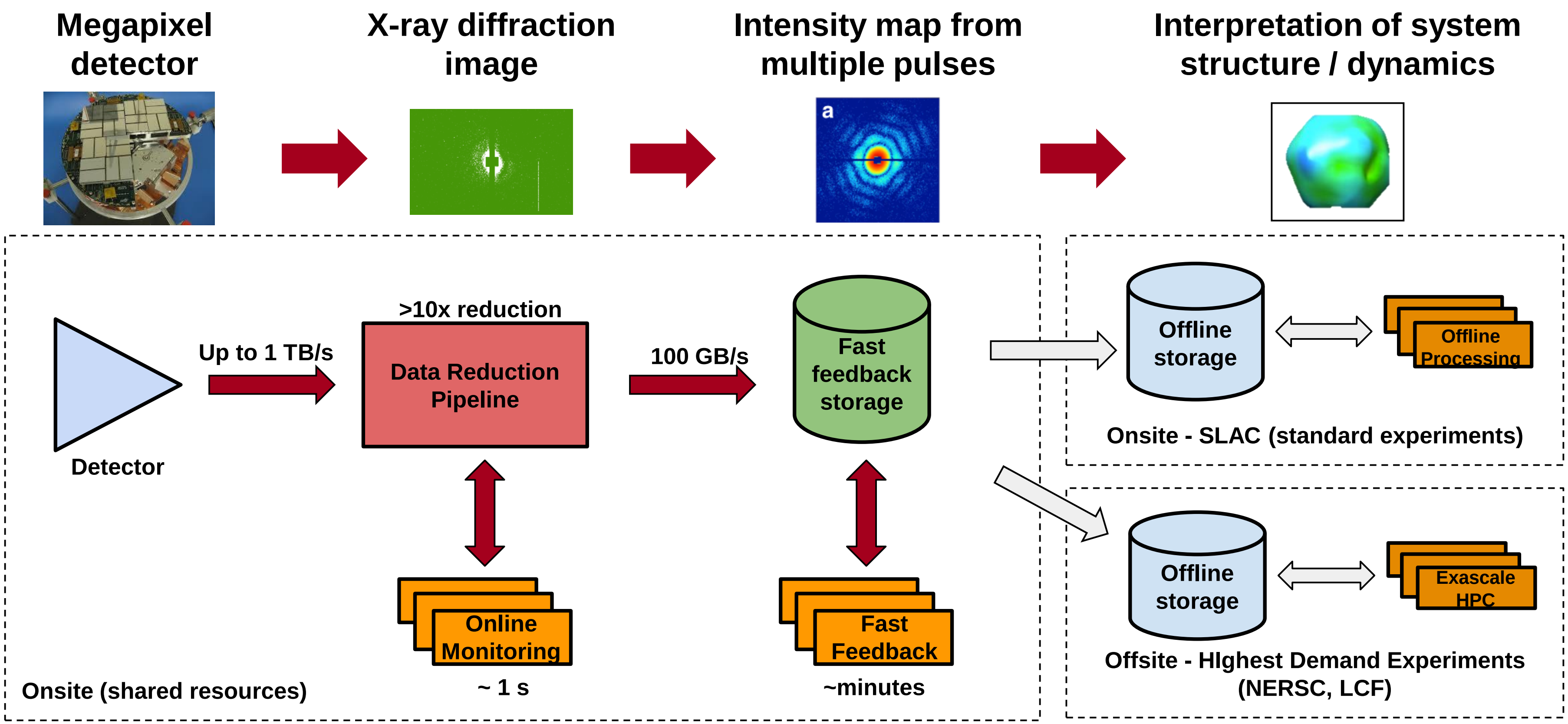
LCLS-II Throughput and Storage Challenges





Throughput/storage requirements are extremely challenging: <u>data reduction needed</u>

LCLS-II Data System

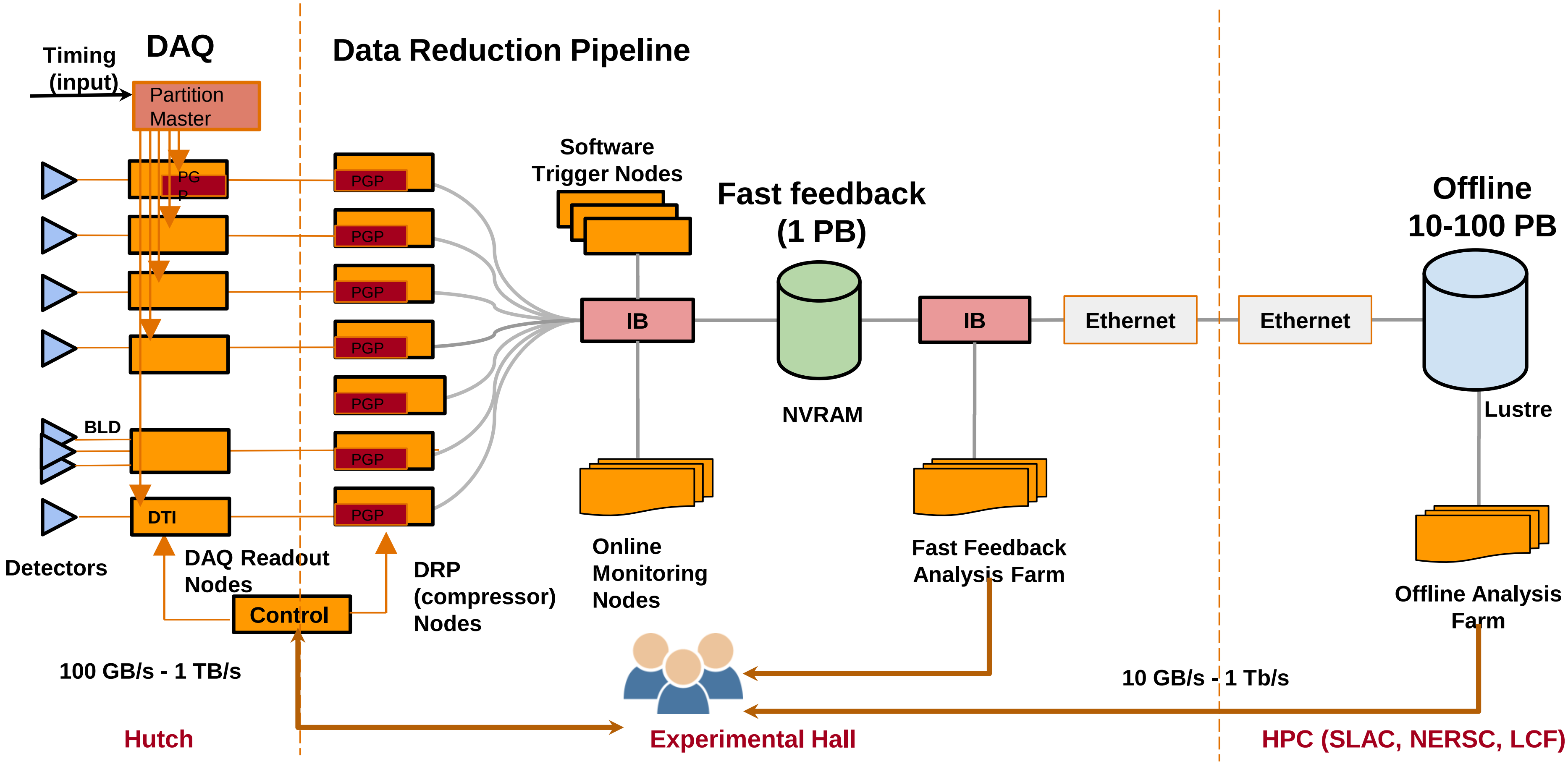


Data reduction mitigates storage, networking, and processing requirements

LCLS strategies for feature extraction

- **In general, feature extraction should be done as close to the sensor as possible.** Moving bits is costly by any metric:
  - Power
  - Networking
  - Storage
  - Computing
  - Cost
- Raw data → Information
  - Detectors produce raw data, but how much information is actually encoded?
- Optimize data system to best serve the needs of LCLS as a whole
  - Flexibility
  - Accommodate variety of detector types, compression types - *even those that have not been invented yet.*
  - Allow detectors to be assembled in any combination at the beamline.
  - For LCLS, each pulse is its own adventure – timestamp data

LCLS-II Data System



LCLS-II DAQ advances

- **Controlled Deadtime:** LCLS-II Timing system enables controlled deadtime
  - Delivers frames of fast control data over fiber optics @ 929 kHz
  - Identifies PulseID, beam present, timing markers, control words sequenced by experiment request.
  - Timing Master capable of appending commands to frame data
    - Trigger decisions (exposure and readout control)
    - Commands: configuration control and event handling
  - Distribution will fan out command data and fan-in feedback information
  - **Sensors can now participate in controlled deadtime**
- **Hardware and software event vetoes**:
  - **L1Trigger** (hardware) can feed back signal from fast detector to throttle readout in a slow detector (for participating detectors)
  - **L3Trigger** is a software trigger decision to keep/toss all data associated with a PulseID
    - Full rate event build limited to a software trigger decision
      - Each DRP node reduces input to a trigger primitive, e.g., number of photons on a detector segment, and passes to the software trigger nodes for compilation
      - The software trigger nodes make a *monitor decision* (forward event to online analysis farm) and a *record decision* (record event in FFB data cache).

Different Levels of Data Processing

Data Reduction Pipeline (DRP):

- Purpose: **Feature extraction** of science information and rejection of data or events that do not meet scientific criteria; **reduce data in a way that does not affect the science result;** reduction of data volume *before data reaches disk*
- Multi-threaded C++ running on ~40 nodes written by core LCLS team
- Small number of algorithms (~20) supports most experiments
- *Real time data reduction must keep up with input data rate*

Online:

- Purpose: real time analysis of acquired data within <1s of readout
- Reads statistical subsample of data from memory
- Builds selectable subsets of data which flow through the Data Reduction Pipeline
- Users analyze data on-the-fly; used to direct experiment operations

Fast FeedBack (FFB):

- Purpose: near real-time feedback to allow experiments to make operational decisions
- Runs on disk-based data (reserved for running experiment) with latency of ~1 minute with parallelized-python code (top level written by users) for quick development

Offline:

- Purpose: obtain final physics results
- Mostly parallelized-python code (top level written by users)

Feature Extraction/Data Reduction Algorithms

SLAC

Diverse science at LCLS-II requires many specialized data reduction algorithms.

- Use the **SIMPLEST, MOST ROBUST** algorithm to get the job done.
- **Prescale data**: save the raw data for a selectable fraction of the events for validation offline.

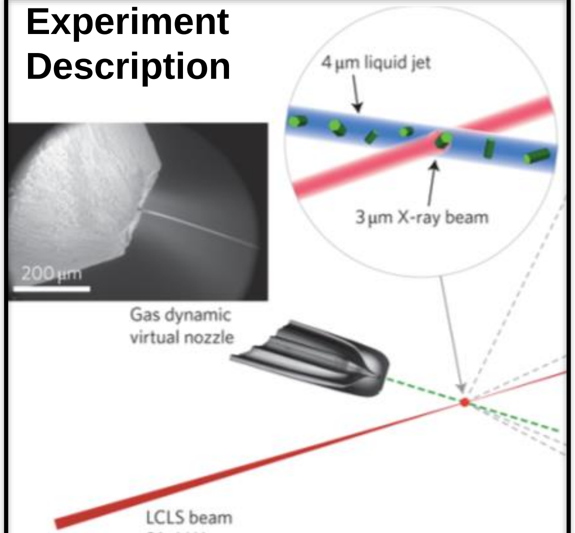
Determined all proposed LCLS-II experiment types through 2028.

Identified **~10 data reduction categories:**

- Triggering
- Accumulating
  - Includes angular integration averaging
- Binning
- Lossless compression
- Lossy compression (SZ algorithm from ANL)
- ROI
- Zero-suppression (software and firmware)
  - Includes peak finding
- Timetool calculation (firmware)

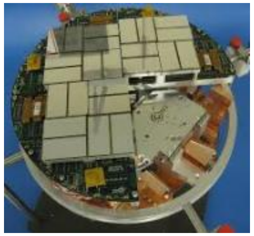
LCLS-II Data System Architecture: Single Particle Imaging Example

Experiment Description



- Individual particles are injected into the focused LCLS pulses
- Scattering patterns are collected on a pulse-by-pulse basis
- 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

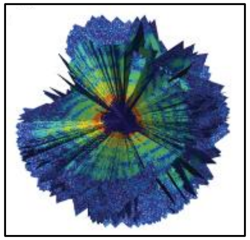
Coherent scattering image

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

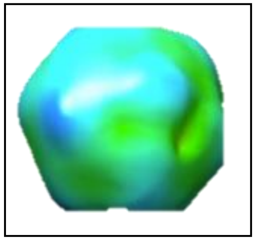
3 TFlops
16 TFlops

6 GB/s
100 GB/s

Intensity map from multiple pulses



Interpretation of system structure / dynamics



Data Analysis
- Orient patterns
- Average
- 3D intensity map
- Reconstruction

270 PFlops
1340 PFlops

Data reduction mitigates storage, networking, and processing requirements

LCLS-II and ATLAS: Similarities and differences

	LCLS-II 2022	LCLS-II 2026+	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	300+ GB/s	160 GB/s	1,000 GB/s
Zero-and-Junk-suppressed rate	10 GB/s	30+ 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)	

Credit: Richard Mount

LCLS-II data volume similar to ATLAS

Summary

LCLS-II, LCLS-II HE, and detector upgrades create demanding data throughput and processing rates, demanding a coordinated effort to upgrade the LCLS Data Systems and SLAC computing infrastructure

Data reduction as close to detector as possible.

		Phase I	**Phase II**	**Phase III**
Parameter	LCLS-I **Present**	LCLS-II comm. **2020**	LCLS-II ops **2024**	LCLS-II HE **2028**
Ave throughput	1-2.5 GB/s	2.5-25 GB/s	5-200 GB/s	1296 GB/s
Peak throughput	5 GB/s	200 GB/s	200 GB/s	1.3 TB/s
Data cache storage	50 TB/hall	1 PB	3 PB	10 PB
Peak Processing (offline)	50 TFlops	1 PFlops	5 PFlops	>130 PFlops
Disk storage	6 PB	16 PB	36 PB	>100 PB

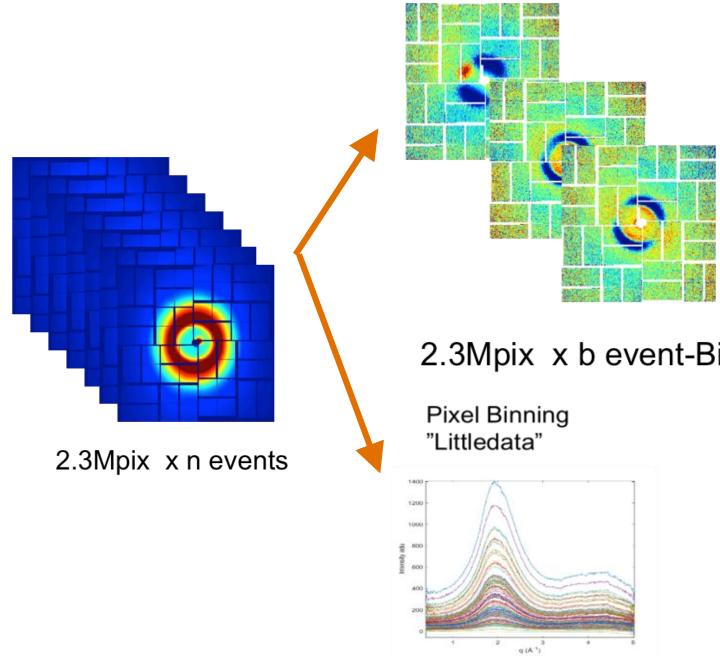
Acknowledgements:
Thank you to the people doing all the honest work: Matt Weaver, Christopher O'Grady, Ric Claus, Dan Damiani, Mikhail Dubrovin, Mona Uervirojnangkoorn, Silke Nelson, Clemens Weninger, Chuck Yoon, Sioan Zohar

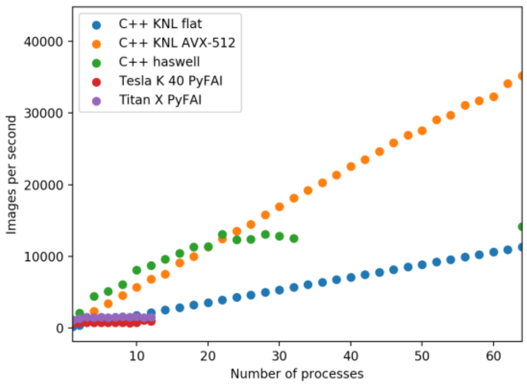
Common Data Reduction Algorithms

Common DRP Example: Binning & Angular Integration

Liquid scattering and Chemistry in solution

- Sorting images by pump-probe delay time and averaging them into time bins (very memory and I/O intensive)
- Angular integration to obtain scattering signal (memory bound)
- Benchmarking of angular integration on different hardware architectures

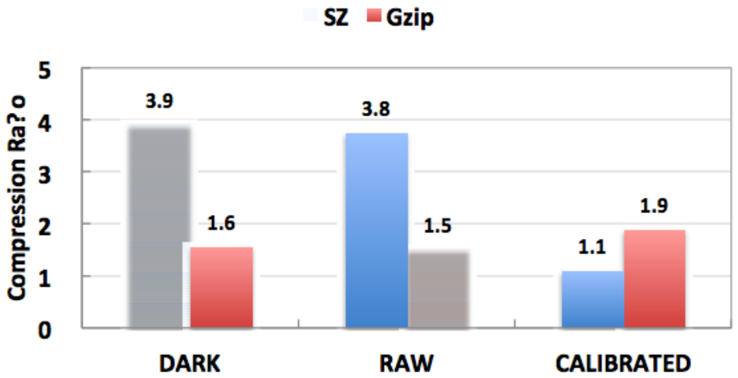




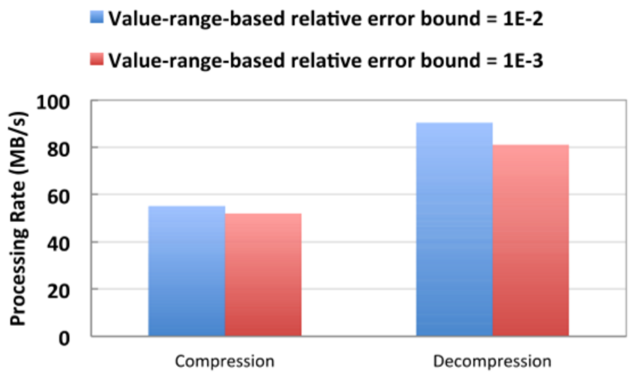
Example DRP algorithm from our data reduction toolkit: angular integration

Common DRP Example: Compression

- Some physics experiments (XPCS, FXS) have "dense" data where every shot is a hit: use compression.
- Measured 2x reduction for lossless-image compression. Measured 100ms CPU time for zlib (gzip) compression of 2 MPixel image
- Also examining ANL "SZ" lossy-compression with user-definable precision (relative or absolute errors). Validated on a crystallography dataset introducing 20 ADU error.
- SZ 50MB/s/core: 1600 cores for 10 kHz 4 MPixel detector: daunting … a work in progress

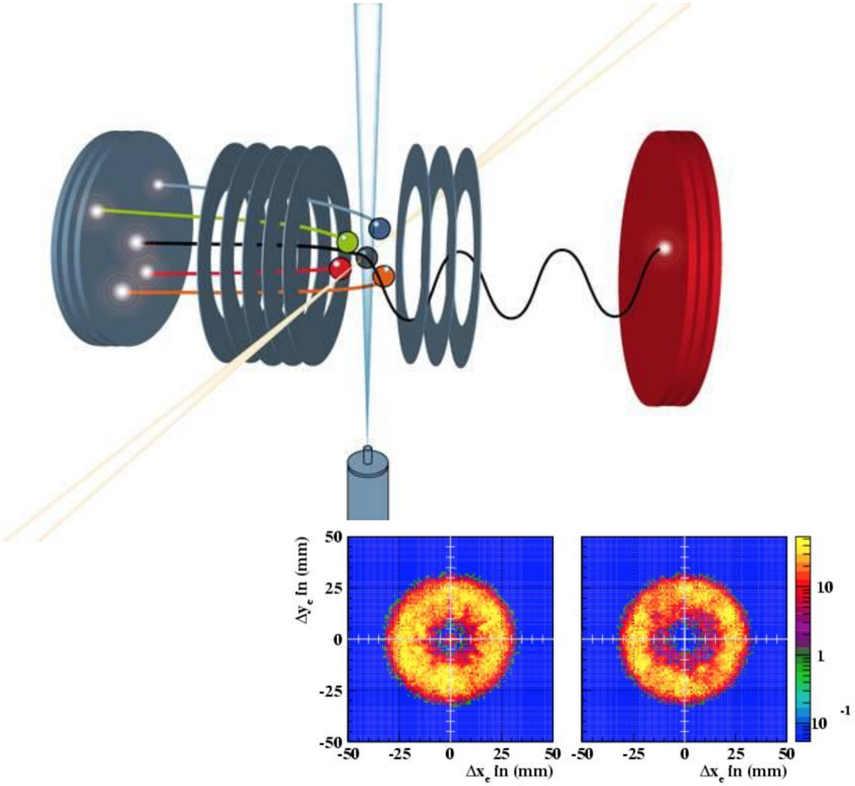


(b) Point-wise rela?ve error bound = 10^{-2}



Example of DRP algorithm from our toolkit: lossless and lossy compression

Common FFB/Offline Example: Reconstruction of particle momenta



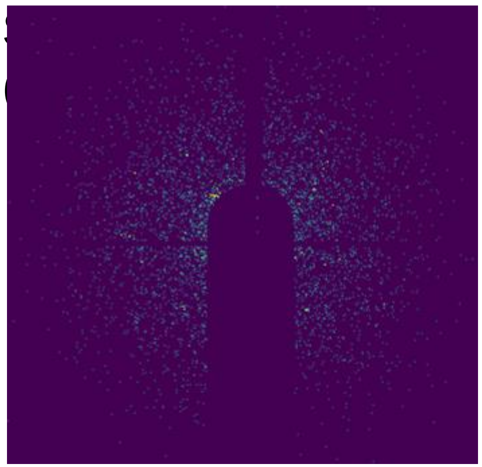
Coincidence spectroscopy of electrons and ions

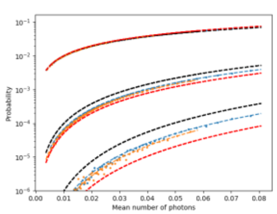
- Digitizers measure arrival time of particles (1.25 GHz sampling)
- Zero suppression done in hardware for data reduction
- Algorithm reconstructs particle information from timing information
- initially > 20 TFlops required

Example of an algorithm that runs on Fast Feedback Layer

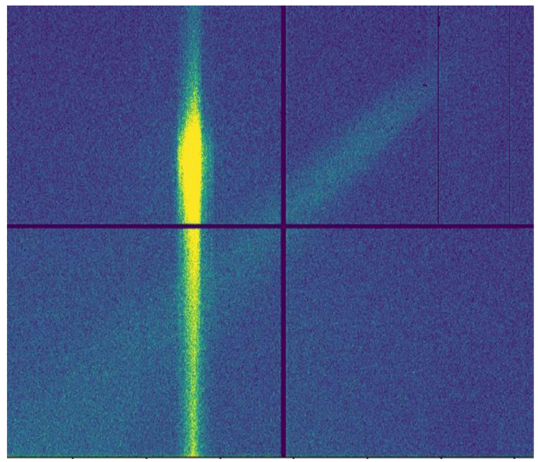
Common FFB/Offline Example: Photon finding

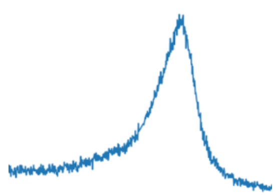
X-ray Photon Correlation





X-ray Emission Spectroscopy





Reconstructing photon hits on image detector is important algorithm for many experiments

- 2 Threshold droplet algorithm
- 50 ms processing time for 1 MPix Camera (including detector corrections)
- 70 TFlops for 0.5 MPix @ 40kHz

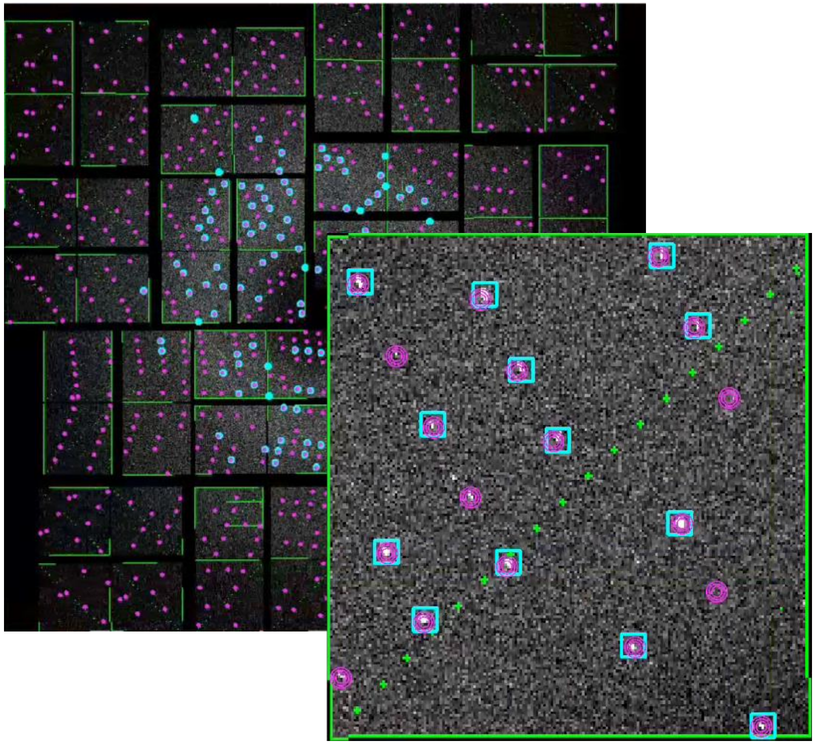
Example of an algorithm that runs on Fast Feedback Layer

Common FFB/Offline Example: Indexing

Serial Femtosecond Crystallography

Determine atomic structure of Biomolecules and proteins

- One of the most computationally expensive analyses
- Significant experience: have run this with MPI on 30,000 cores @NERSC
- Finding Bragg spots in image, 30 TFlops for 4MPix & 40 kHz
- Indexing: find orientation of crystal, 2 PFLops for 4MPix & 40 kHz
- Critical need for near real-time FFB Indexing to verify crystal quality



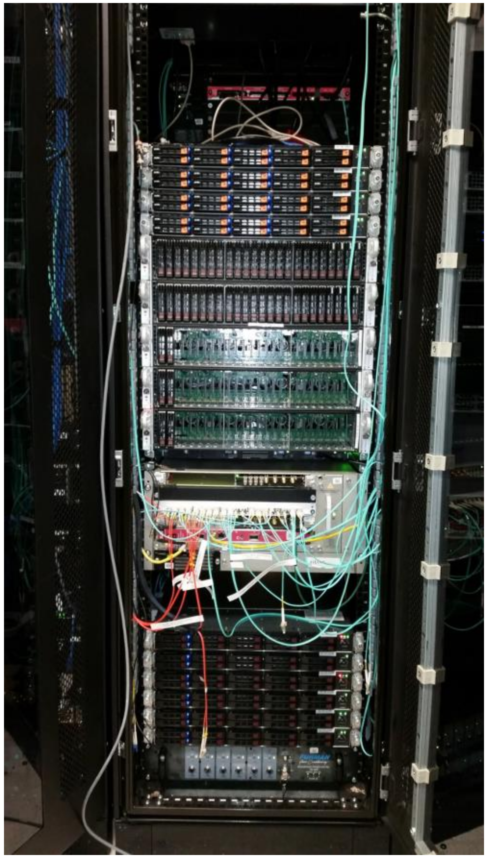
Computationally intensive example of algorithm that runs on FFB or Offline

Data Reduction in specific workflows

- XPCS (X-ray Photon Correlation Spectroscopy)
  - Every event is a hit, photons are (often) dense
  - Either lossless compression, SZ compression or only saving speckles
  - Need to enumerate various cases more carefully (hard/soft x-ray, detector distance, bragg-spot/diffuse…)
- FXS (Fluctuation X-Ray Scattering: high concentration limit of SPI)
  - With good detector corrections and beam-center knowledge, believe we can compute angular correlations and sum resulting images
  - Working with CAMERA on this
- TES (Transition Edge Sensor) Detector
  - cross talk correction is computationally intensive (firmware)
  - event time-overlap complicates separation of data into events
- SFX (Serial Femtosecond Crystallography) in the unlikely multi-hit case

LCLS-II Prototype

Prototype for LCLS II Data Reduction Pipeline



Experiment Timing

FFB

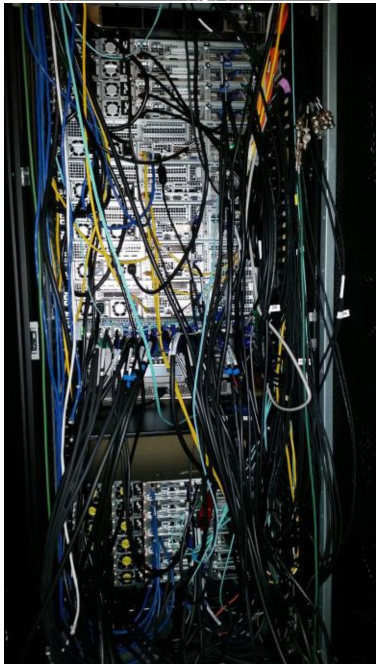
Sensors

DRP A

IB Switch

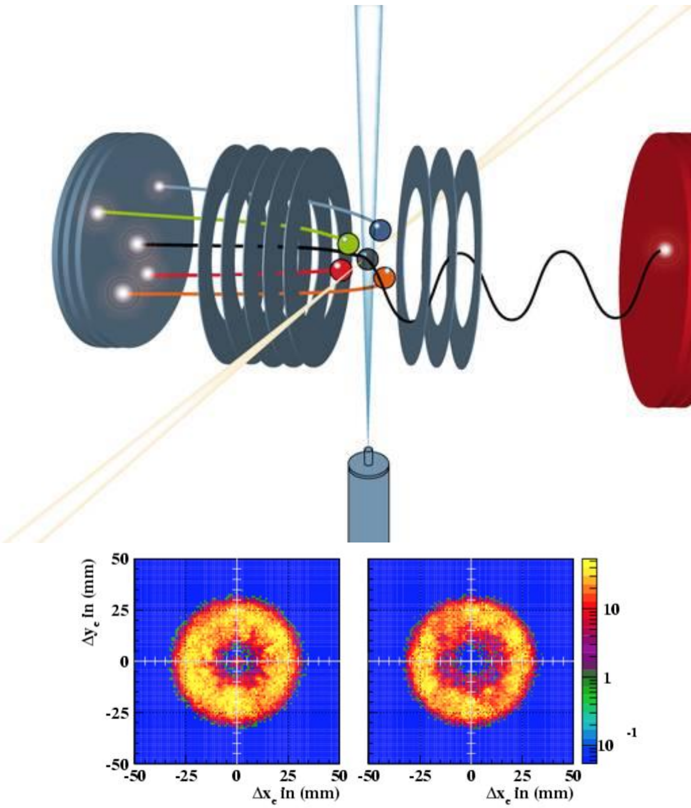
Accelerator Timing

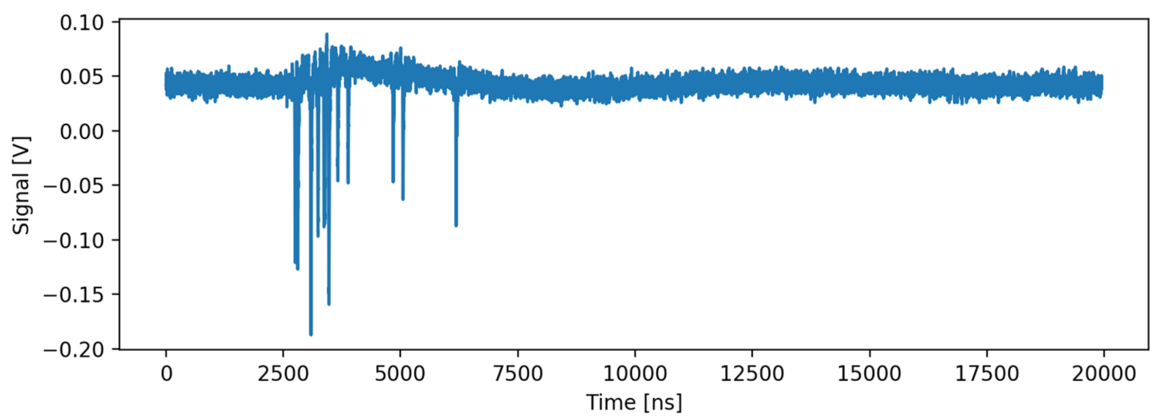
DRP B



Data Reduction: High Speed Digitizer

- High speed digitizer with 1.25 GHz sampling rate
- Up to 20 digitizer channels
- Zero suppression done in hardware for data reduction





Data Reduction: High Speed Digitizer

Results

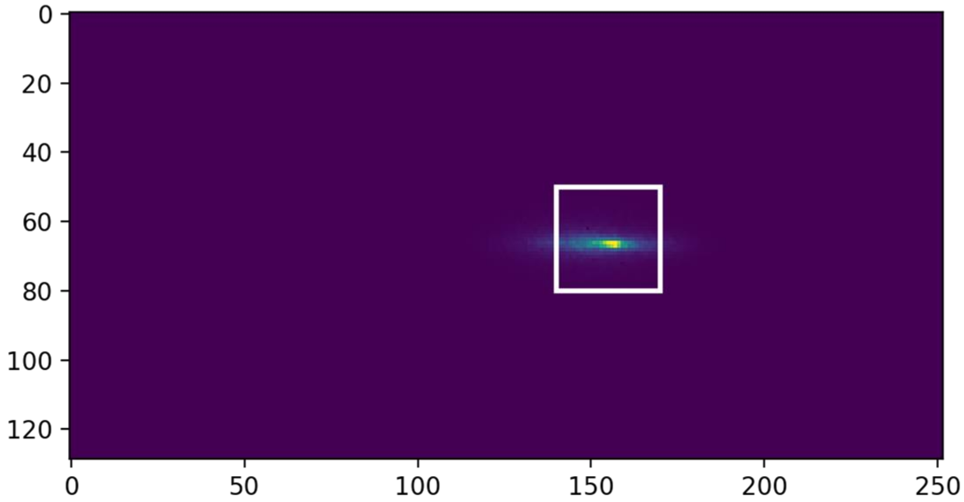
- 2 channels @70kHz (530MB/s) writing to xtc file
- Deadtime functionality has been demonstrated and can be attributed to a source

Lessons learned

- I/O limited (single writer only around 500-700 MB/s to Lustre) (see Data Management talk)
- Current PGP driver is rate limited by interrupts

-> PGP driver for new PGP card generation will address issue

Data Reduction: Area detector ROI (hardware emulated)



- Region of interest for data reduction
- 30 * 30 ROI

Results

- 1 channel @10kHz writing ROI to HDF5 / xtc file

Lessons learned

- Throughput limited by current PGP card to 2GB/s per node (PCI 2.0 × 4)