Data Analytics at the Exascale for Free Electron Lasers: Overview to BES

ExaFEL Team
April 23rd 2019

EXASCALE COMPUTING PROJECT

National Nuclear Security Administration

U.S. DEPARTMENT OF ENERGY | Office of Science

Outline

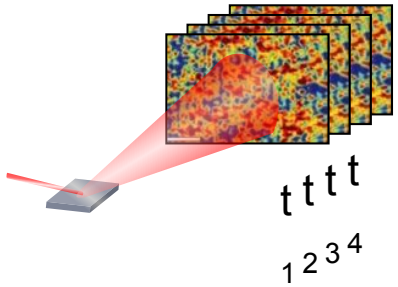
- Brief introduction to **FEL science case**
  - Project plan: big picture of how the project is organized
  - Data flow: how the data move from the beamlines to HPC and back
  - ExaFEL science cases: nanocrystallography and single particle imaging
- **KPPs**: quantify what ExaFEL needs to achieve to be successful
  - LCLS data analysis framework: features and scalability
  - Evolution of the analysis framework (Legion integration)
- **Progress and next steps**

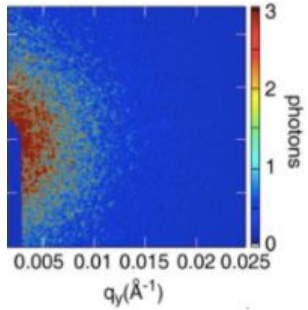
Brief Introduction to ExaFEL Science case

LCLS-II: many workflows, massive throughputs

Coherent Scattering

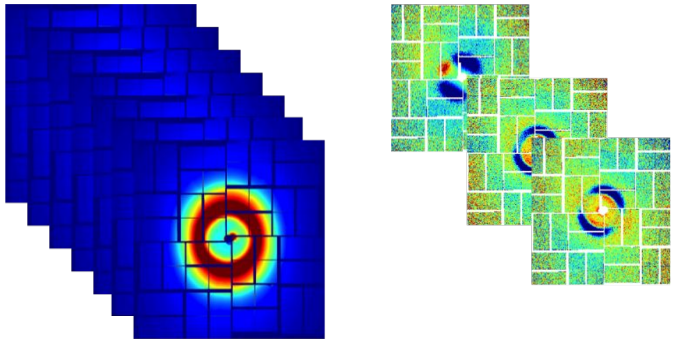
XPCS XSVS





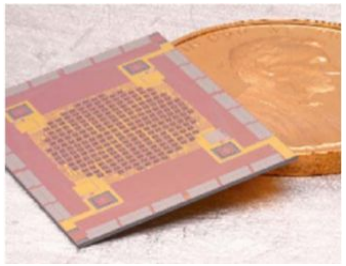
2023: 20 GB/s, 4 TF (reduction), 34 TF (analysis)
2026: 80 GB/s, 34 TF (reduction), 270 TF (analysis)

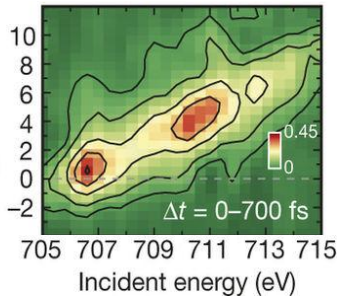
Liquid scattering



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

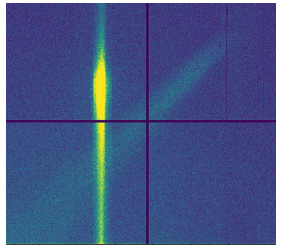
Resonant Inelastic Scattering

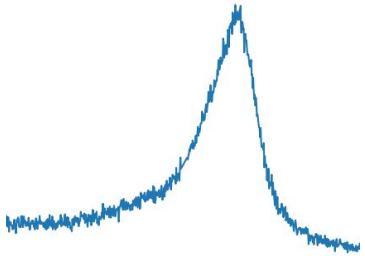




2023: 20 GB/s, 4 TF (reduction), 1 TF (analysis)
2026: 200 GB/s, 40 TF (reduction), 2 TF (analysis)

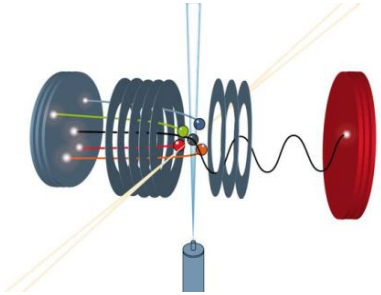
X-ray Emission Spectroscopy (XES)

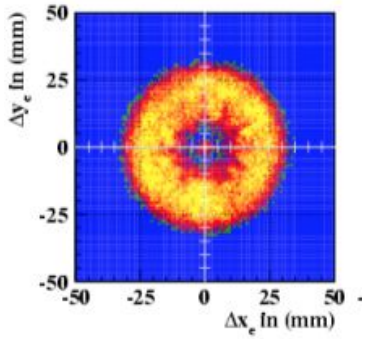




2023: .1 GB/s, < 1 TF (reduction), .2 TF (analysis)
2026: 10 GB/s, <1 TF (reduction), 20 TF (analysis)

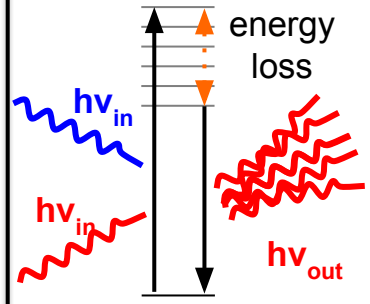
Coincidence Spectroscopy

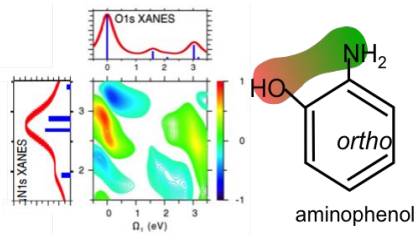




2021: 200 GB/s, <1TF (reduction), <1TF (analysis)

Nonlinear Spectroscopy





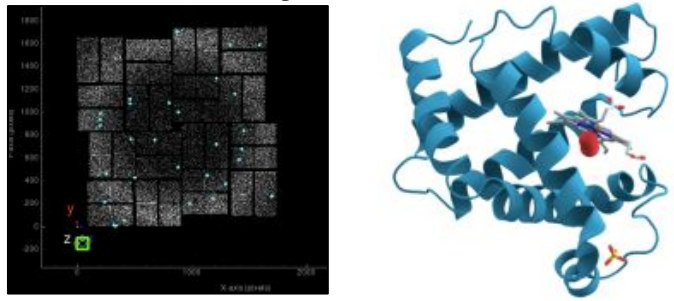
2023: 20 GB/s, 3 TF (reduction), <1 TF (analysis)
2026: 80 GB/s, 16 TF (reduction), <1 TF (analysis)

EXASCALE

Ability to reduce data and flexibility to handle multiple workflows will become critical

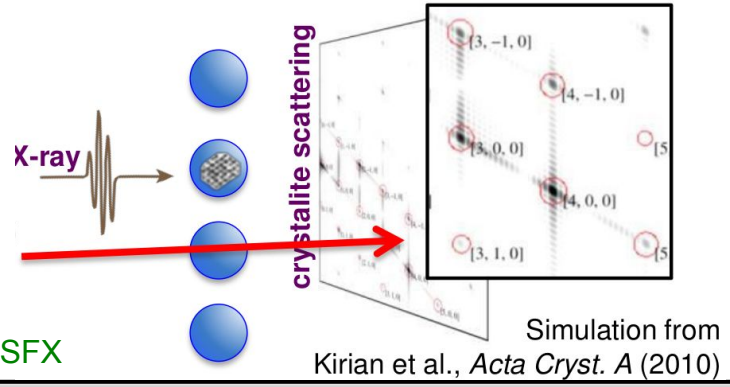
LCLS-II and LCLS-II-HE: Inverse Problems and Compute Intensive Workflows

Macromolecules Structure and Dynamics



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

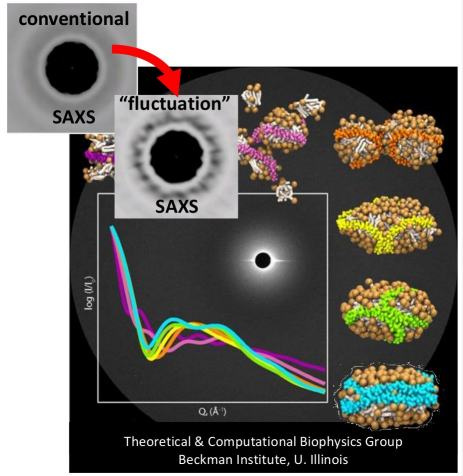
Understanding & Controlling Materials Nucleation Pathways



~SFX

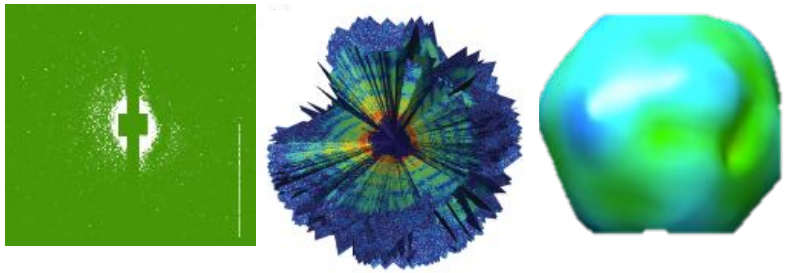
Simulation from Kirian et al., *Acta Cryst. A* (2010)

Understanding & Controlling Nano-materials Self-Assembly



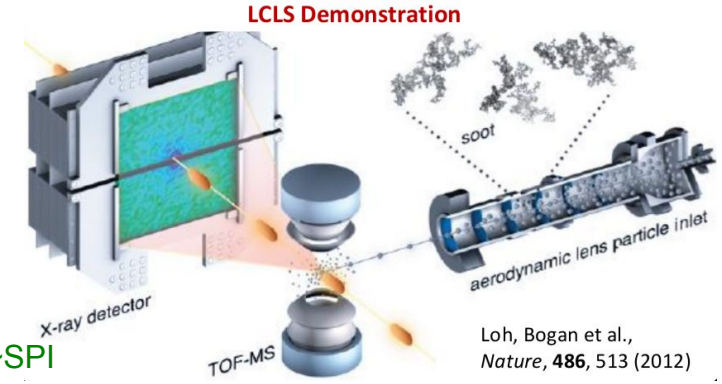
~SPI

Viruses Imaging



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

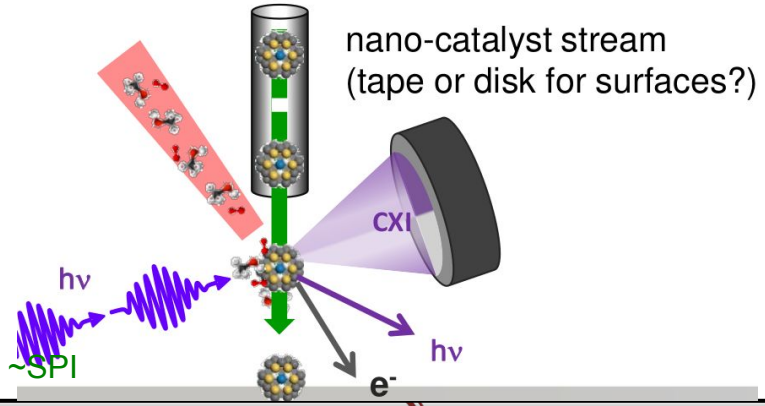
Chemistry and Morphology of Combustion Aerosols (soot)



~SPI

Loh, Bogan et al., *Nature*, **486**, 513 (2012)

Coupled electronic & nuclear dynamics in heterogeneous (nano) catalysis



~SPI

Ability to surge highest demand experiments to HEC will be critical

Computing Requirements for Data Analysis: a Day in the Life of a User Perspective

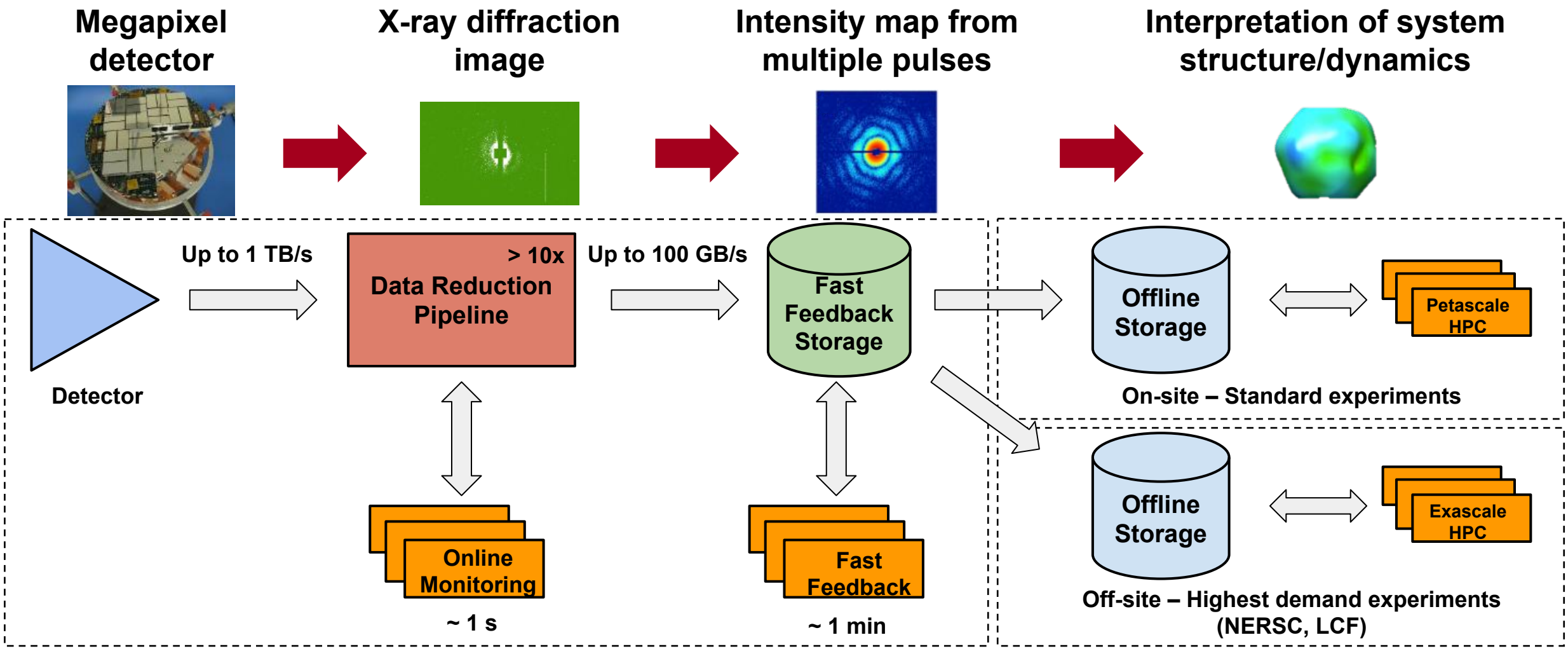
- During **data taking**:
  - Must be able to get real time (~1 s) **feedback** about the **quality of data taking**, e.g.
    - Are we getting all the required detector contributions for each event?
    - Is the hit rate for the pulse-sample interaction high enough?
  - Must be able to get **feedback** 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?

ExaFEL focus

- 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 in preparation for the next shift
- During **4 months** after the experiment: must be able analyze the raw and intermediate data on **fast access storage** in preparation for publication
- **After 4 months**: if needed, must be able to **restore** the archived data to test new ideas, new code or new parameters

EXASCALE COMPUTING PROJECT

LCLS-II data flow can be represented as a 4-stage process: detection, data reduction, online analysis, full interpretation



ExaFEL focuses on providing full interpretation capabilities in quasi real time using HEC resources

The Challenging Characteristics of LCLS Computing

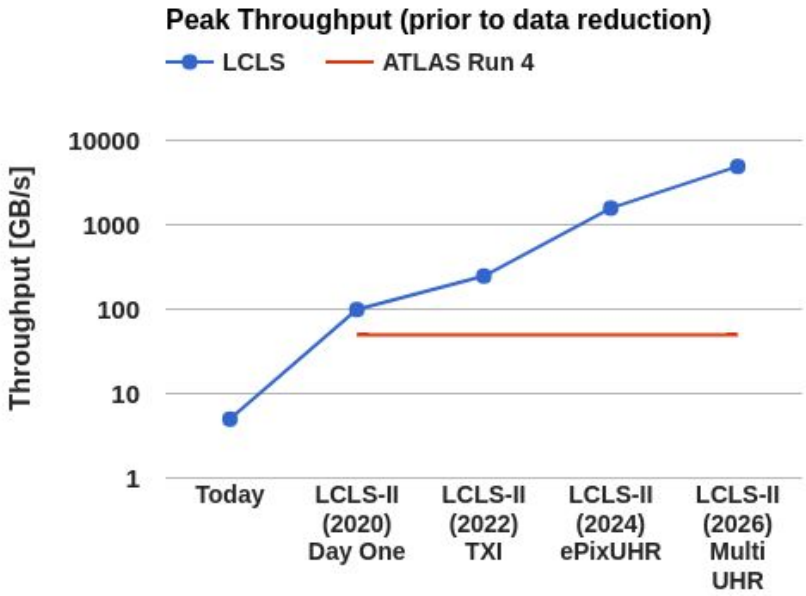
1. **Fast feedback** is essential (seconds / minute timescale) to reduce the time to complete the experiment, improve data quality, and increase the success rate
2. **24/7 availability**
3. **Short burst** jobs, needing very short startup time
4. **Storage** represents significant fraction of the overall system
5. **Throughput** between storage and processing is critical
6. Speed and flexibility of the **development cycle** is critical
   Wide variety of experiments, with rapid turnaround, and the need to modify data analysis during experiments

Example data rate for LCLS-II (early science)

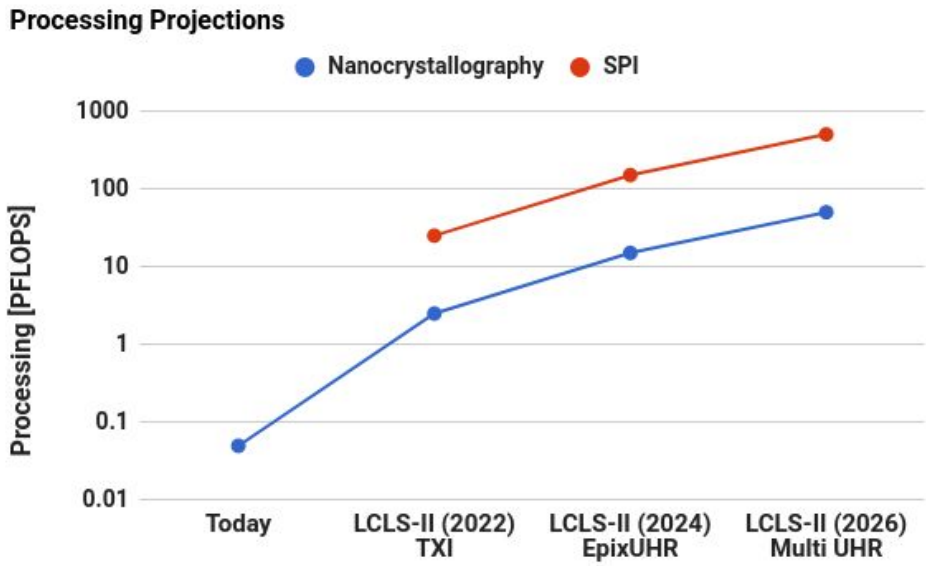
- 1 x 4 Mpixel detector @ 5 kHz = **40 GB/s**
- 100K fast digitizers @ 100kHz = **20 GB/s**
- Distributed diagnostics 1-10 GB/s range

Example LCLS-II and HE (mature facility)

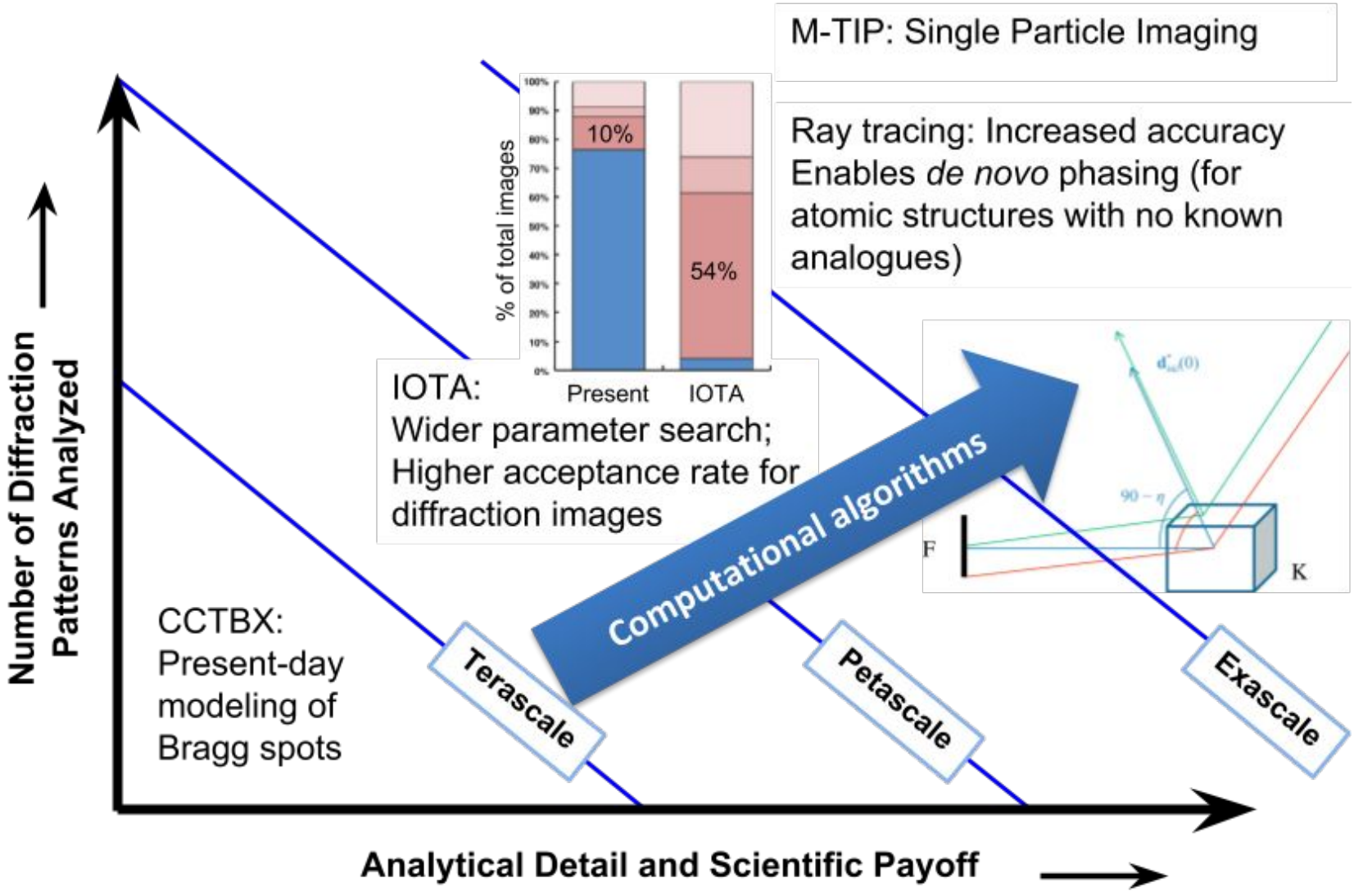
- 2 planes x 8 Mpixel ePixUHR @ 50 kHz = **1.6 TB/s**



Sophisticated algorithms under development within ExaFEL (e.g., M-TIP for single particle imaging) will require exascale machines



From Terascale to Exascale: what we'll be able to do that we cannot do today



Exascale vastly expands the experimental repertoire and computational toolkit

EXASCALE COMPUTING PROJECT

ExaFEL Project Plan

1. **Algorithmic improvements** for high data throughput experiments

2. Port LCLS **data analysis framework** to supercomputer architecture, allow scaling from hundreds of cores (today) to millions of cores

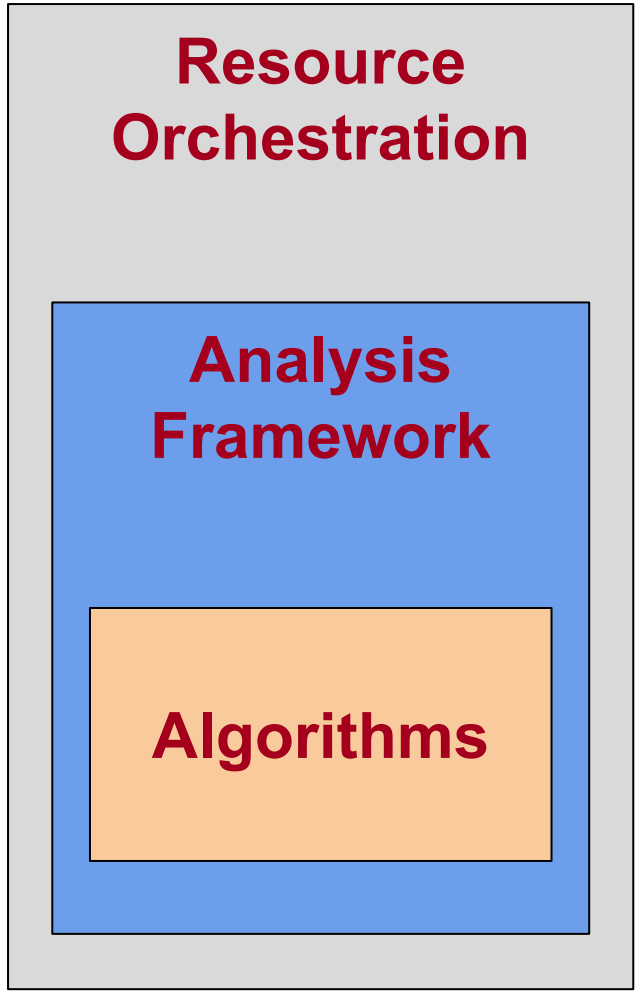
3. Design and develop the **orchestration** of all the resources required to:

   Stream the data on-the-fly from LCLS beamlines to NERSC over ESnet

   Execute the analysis on the supercomputer

   Visualize the results of the analysis back to the experimenters in quasi real time

These developments will benefit all LCLS experiments, not just the algorithms selected for ExaFEL

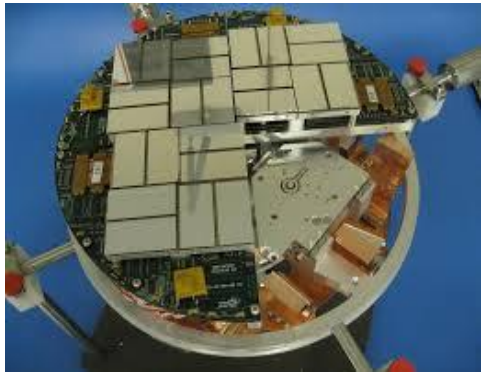


EXASCALE COMPUTING PROJECT

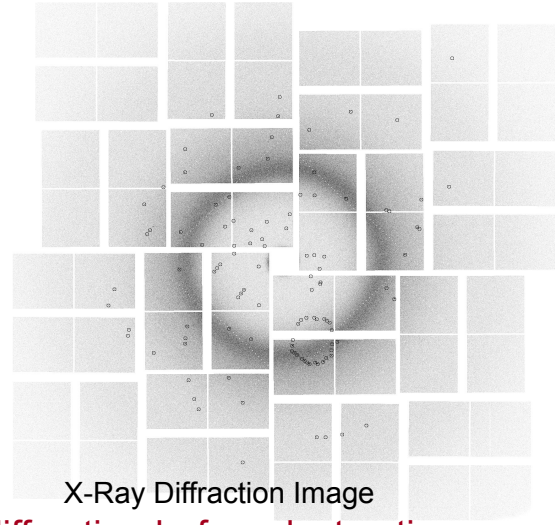
Example of computing intensive algorithms for ExaFEL: Scaling the nanocrystallography pipeline

Avoidance of radiation damage and emphasis on **physiological conditions** requires a transition to fast (fs) X-ray light sources & large datasets

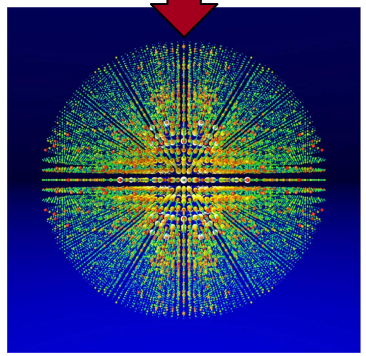
- Main steps in the algorithm are
  (1) identifying the Bragg diffraction spots
  (2) deducing the geometry of the lattice repeat,
  (3) refining the model again
  (4) summing the X-ray signal in each spot for further analysis



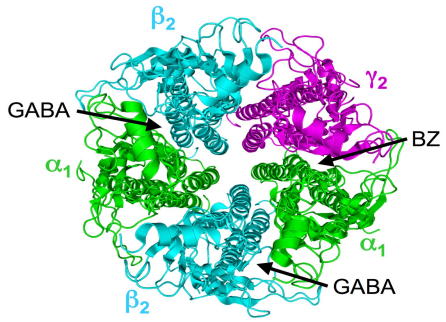
Megapixel detector



X-Ray Diffraction Image
"diffraction-before-destruction"



Intensity map (multiple pulses)



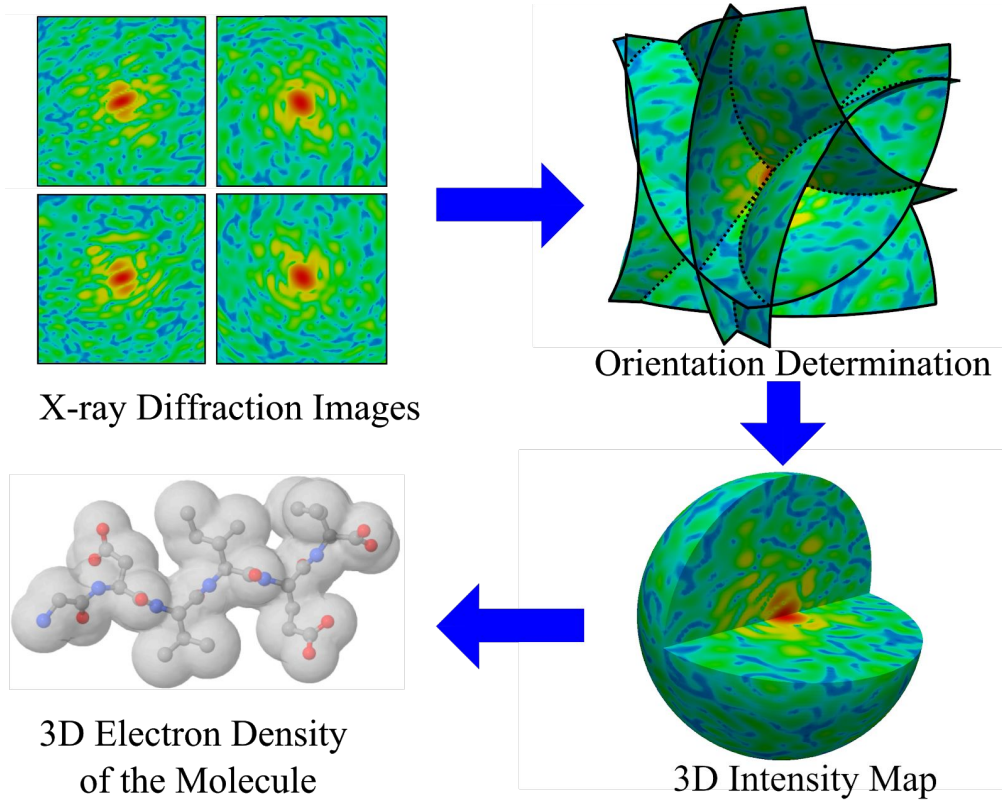
Electron density (3D) of the macromolecule

EXASCALE COMPUTING PROJECT

Example of computing intensive algorithms for ExaFEL: M-TIP - a new algorithm for single particle imaging

M-TIP (**Multi-Tiered Iterative Phasing**) is an algorithmic framework that **simultaneously determines conformational states, orientations, intensity, and phase** from single particle diffraction images

- The aim is to reconstruct a 3D structure of a single particle
  - We can NOT measure: a) the orientations of the individual particles and b) phases of the diffraction patterns
  - MTIP is an iterative algorithm that deduces these two sets of unknowns given some constraints



[1] Donatelli JJ, Sethian JA, and Zwart PH (2017) Reconstruction from limited single-particle diffraction data via simultaneous determination of state, orientation, intensity and phase. PNAS 114(28): 7222-7227.

Work done in collaboration with CAMERA

EXASCALE COMPUTING PROJECT

KPPs: quantify what ExaFEL needs to achieve to be successful

ExaFEL Application Key Performance Parameters: Definitions & Requirements

Must be able to keep up with data taking rates: fast feedback *(seconds / minute timescale)* ***is essential*** *to reduce the time to complete the experiment, improve data quality, and increase the success rate*

ExaFEL Key Performance Parameter: Number of events analysed per second

Pre ExaFEL capability (LCLS-I): 10 Hz

- LCLS-I operates at 120 Hz, hit rate is ~10% ⇨ 10 events/s for reconstruction

Target capability (LCLS-II and LCLS-II-HE): 5 kHz

- LCLS-II high rate detectors are expected to operate at 50 kHz by 2024-2026, hit rate ~10% ⇨ 5000 events/s for reconstruction (after DRP)
- DRP = Data Reduction Pipeline (for SFX and SPI: vetoes events which are not hits)

EXASCALE COMPUTING PROJECT

ExaFEL KPPs: Plans & Achievements

		FY17	FY18	FY19	FY20	FY21	FY22	FY23
Algorithm	**Expected Ratio to SFX**	**Cori PII (30 PF)**	**Cori PII (30 PF)**	**Cori PII (30 PF) Summit (200PF)**	**Summit (200PF)**	**NERSC-9 (>60PF) Summit (200PF)**	**A21 (1EF)**	**A21 (1EF) Frontier (1 EF)**
SFX	x1	**135 Hz (6%)**	**3 kHz (52%)**					
SFX with IOTA	x2-x5		**119 Hz (36%)**		**5 kHz (OED)**			
SFX diffuse scattering	x2-x5				5 kHz (SD)			
SFX with x-ray tracing	x10-x20			100 Hz (OED)	1 kHz (OED)			
SFX with x-ray tracing & diffuse	x10-x20					1 kHz (OED)	**5 kHz (OED)**	
SPI with M-TIP	x10-x20			100 Hz (SD)	1 kHz (SD)	1 kHz (NSD)	1 kHz (OED)	**5 kHz (OED)**

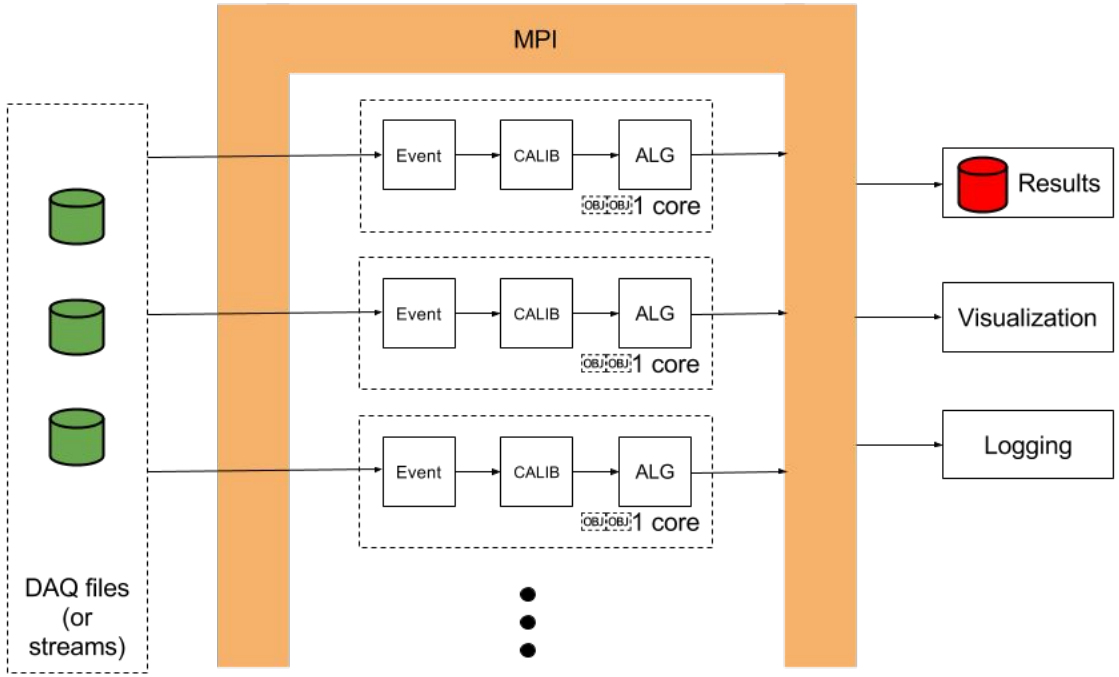
OED = Observational Experimental Data
SD = Simulated Data
NSD = Noisy Simulated Data

EXASCALE COMPUTING PROJECT

LCLS Data Analysis Framework

Main features LCLS data analysis framework :

1. **Rapid development** with simple photon-science-standard Python programming language
2. **Complexity is hidden**: parallelization, common algorithms, detector corrections, parallelization, file formats
3. Allows for **real-time analysis** in an identical fashion as offline analysis



The framework, which is the same for all LCLS experiments, handles data **marshalling, parallelization over events and calibration**

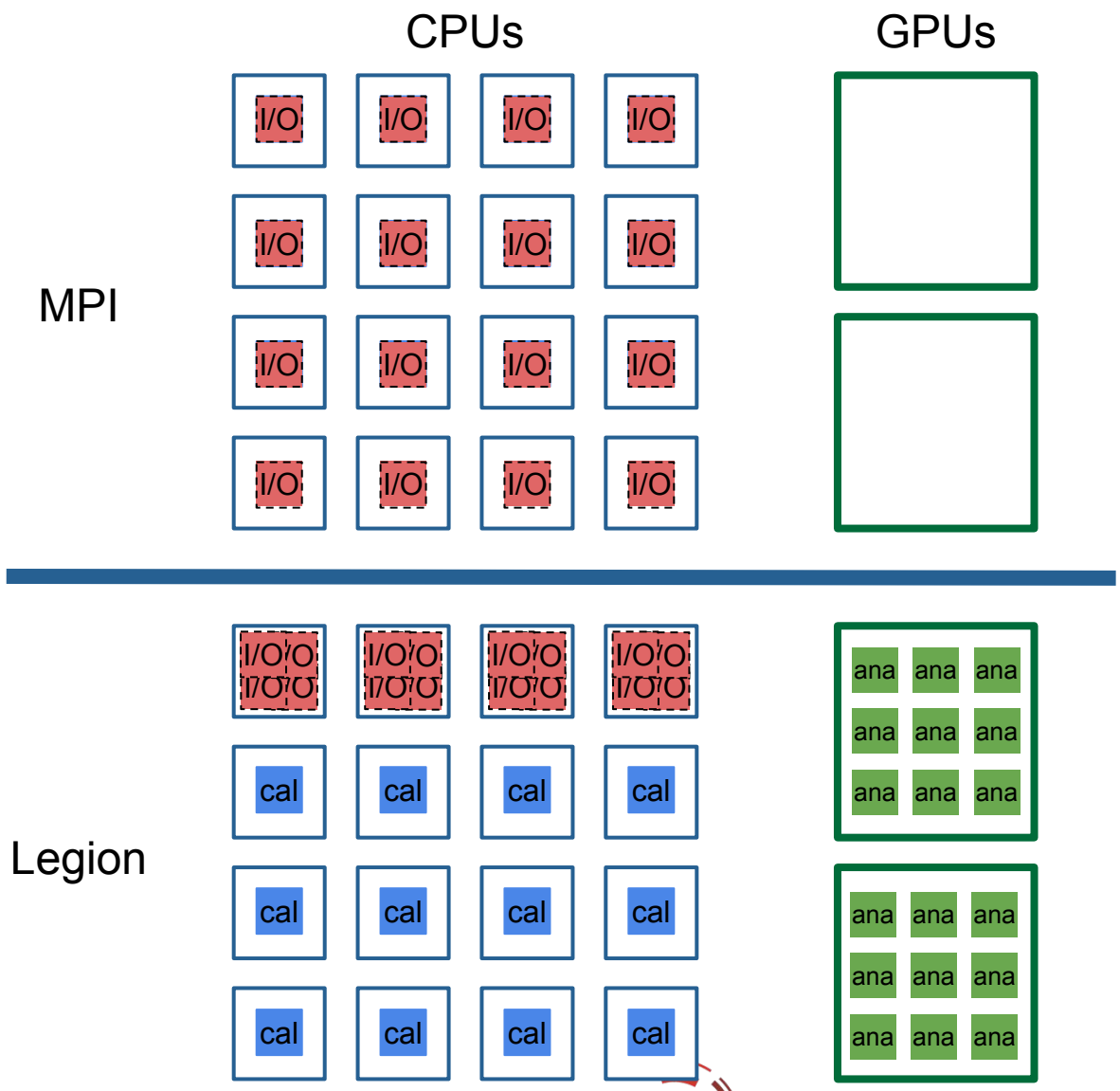
Scalability is key: the more time we spend in the ALG box for each event ⇒ the more cores we need to run in parallel to keep up with data taking rates

EXASCALE COMPUTING PROJECT

Evolution of the LCLS analysis framework: integration with Legion

- **Maximizes throughput** of data analysis via **flexible assignment** of resources
- **Overlaps compute**, I/O, communication
- Provides **performance portability** to future architectures such as Summit

Work done in collaboration with Stanford University and CS group at SLAC



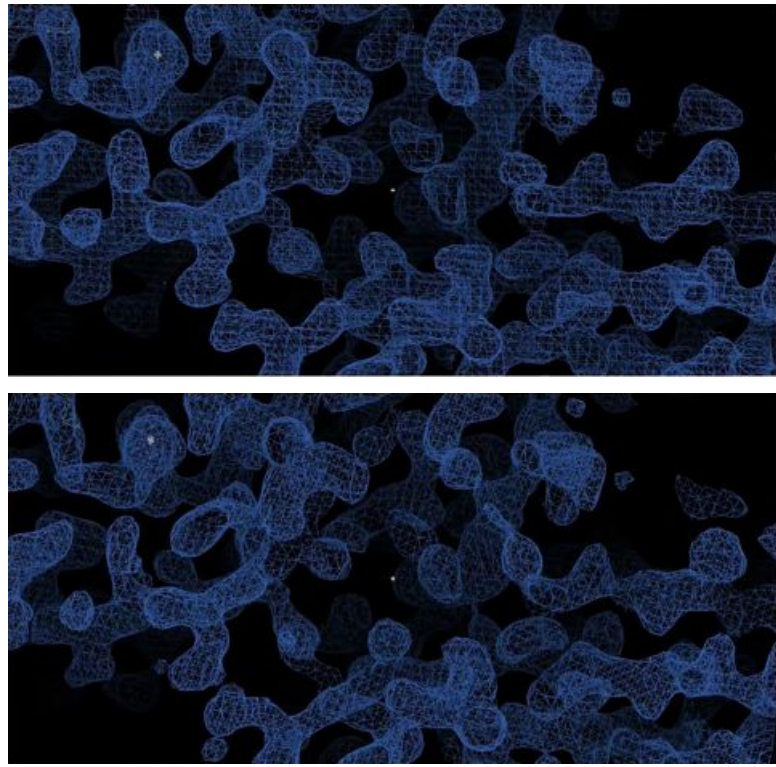
EXASCALE COMPUTING PROJECT

Last year progress and future work

Last Year Progress

ExaFEL has completed several milestones spanning different areas, ranging from resource orchestration to data movement to algorithmic improvements:

- Integration of the LCLS analysis framework with the **Legion** framework for improved scalability and portability
- Integration, improvement and evaluation of the SZ **lossy compression** algorithm for data reduction
- Ability to selectively request an **uncongested data path** over **ESnet**
- Evaluation and selection of different **data transfer** technologies
- Introduction of a new **data format optimized for data streaming** and data reduction (xtc2)
- **Increase of the usable data set** for nanocrystallography experiments by introducing the Integration Optimization Triage and Analysis algorithm
- Ability to scale the **merging step** in nanocrystallography
- Porting of the nanocrystallography code to the **Summit** supercomputer at ORNL



Simulated lossy compression shows Se-SAD can tolerate absolute error bound of 10 ADUs without any problems

EXASCALE COMPUTING PROJECT

Future Work

- FY20:
  - Optimize orchestration (streaming, memory usage for events, startup times)
  - Realistic simulation of SPI data accounting for beam features, sample injector and detector, for M-TIP to ingest
  - Accelerate x-ray tracing on GPUs and deploy to Summit and NERSC-9
  - Accelerate M-TIP on GPUs and deploy to Summit and NERSC-9 (against simulated data)
- FY21:
  - Expand data path to last mile at SLAC and NERSC (current path limited to ESnet segment)
  - Merge SFX with diffuse scattering
  - Run M-TIP against realistic simulated data
- FY22:
  - Automate workflow
  - Run M-TIP against actual experimental data
  - Scale SFX with x-ray tracing and diffuse scattering to target rate on A21
- FY23:
  - Scale M-TIP to target rate on A21 and Frontier

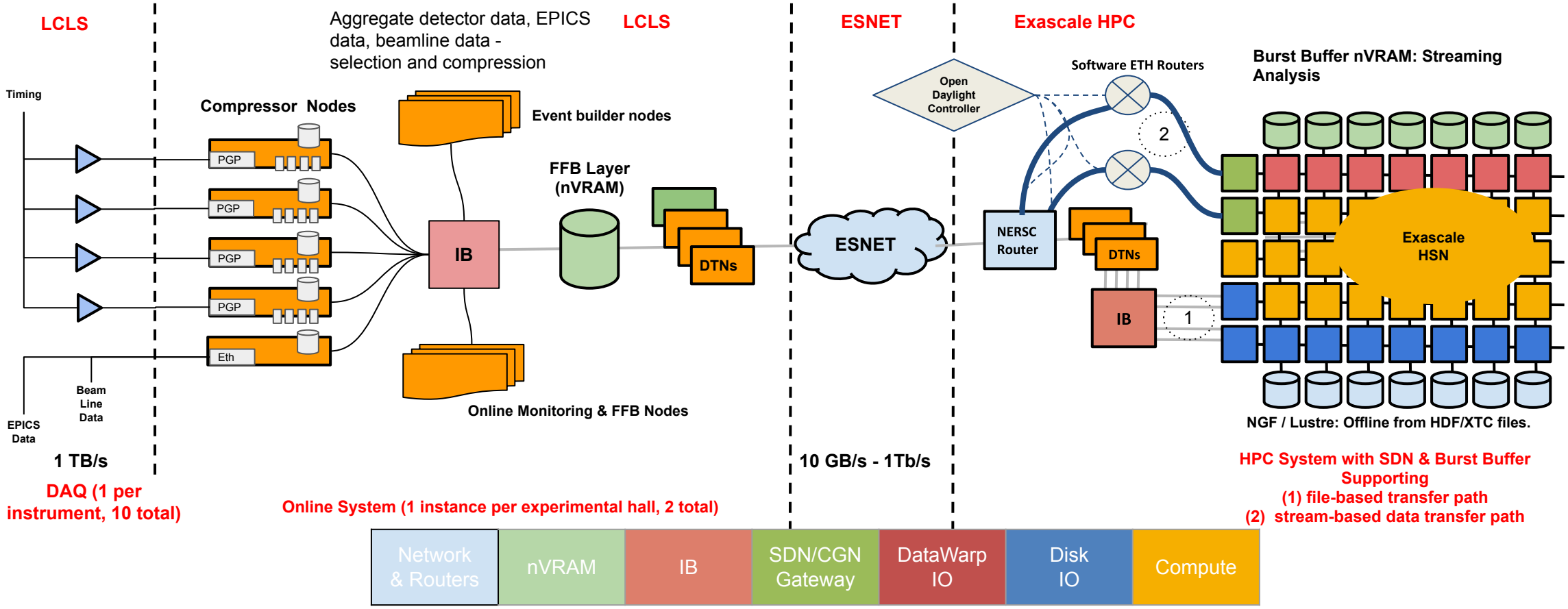
Conclusions

- There's been significant progress in first two years and the next years look very exciting - some of the development in our ability to **solve inverse problems at scale** could be revolutionary for FEL science
- It took some time to hire the right expertise, but, as of **March 2019, ExaFEL is fully staffed**
- While the ExaFEL algorithmic development is focused on **inverse problems** (SFX/diffuse/SPI - as they represent the most computing intensive techniques), **all LCLS high throughput experiments will benefit from this development (e.g. XPCS for material science studies)**

EXASCALE COMPUTING PROJECT

Backup Slides

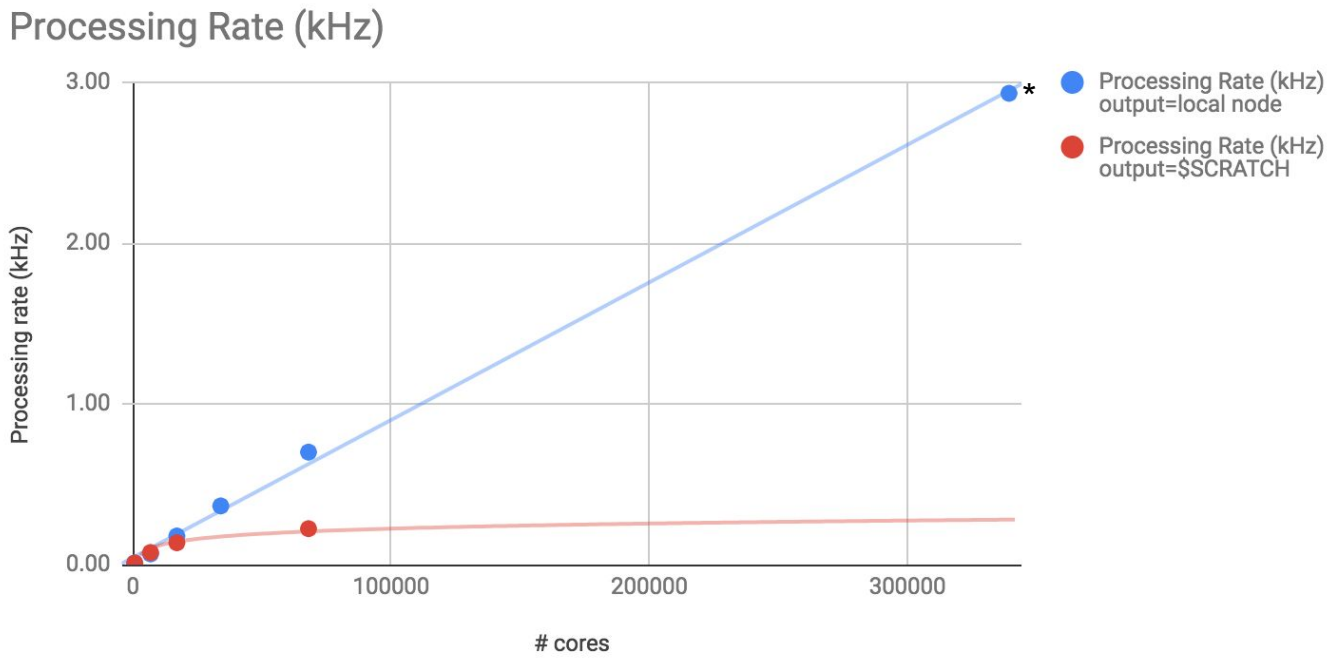
ExaFEL Data Flow



EXASCALE COMPUTING PROJECT

Psana tasking optimization on Cori PII

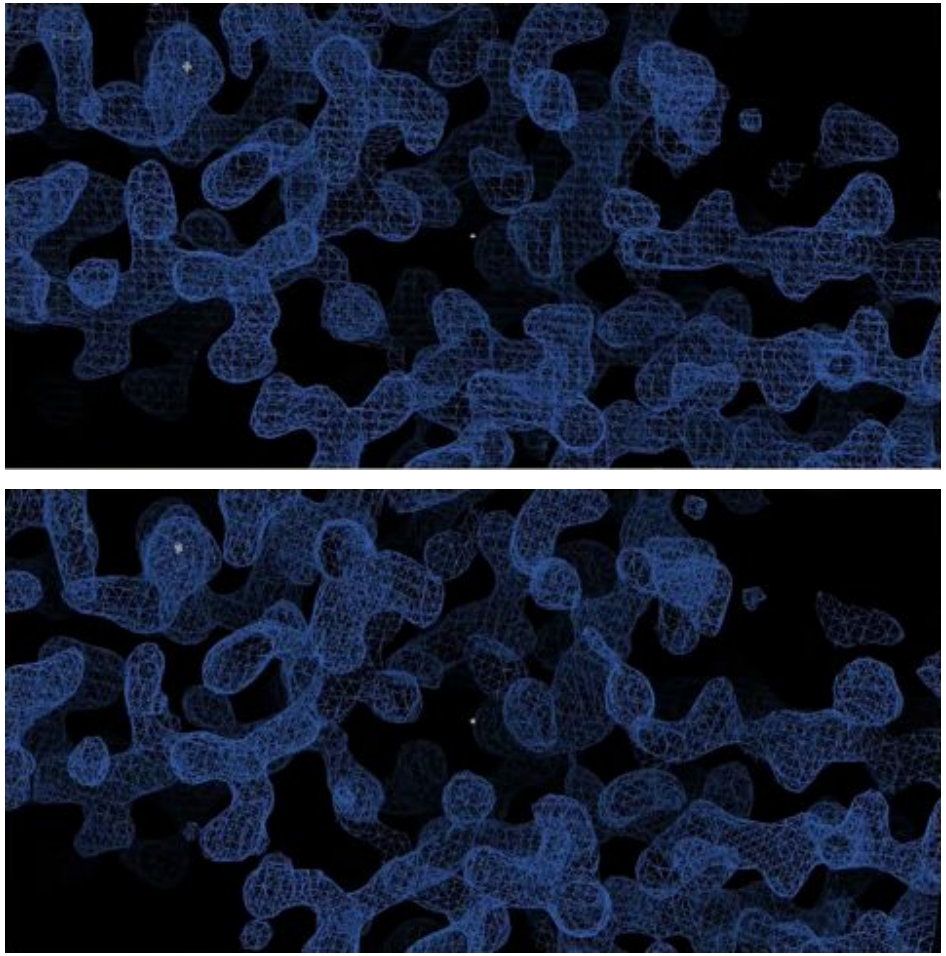
Processing rate = no. of events / wall time



- Parallelization algorithm in Psana2 was improved to accommodate higher rate of data streaming
- We observed linear scaling of cctbx upto 52% of Cori-II (340,000 cores!)
- CCTBX output to Lustre filesystem saturated around 500 nodes (red points). We need to develop a more efficient way to output results.

EXASCALE COMPUTING PROJECT

Lossy Compression (in collaboration with EZ ST team)



Simulated lossy compression shows Se-SAD can tolerate absolute error bound of 10 ADUs without any problems

Lossy compression with absolute error bound of 30 ADUs with SZ v2.0

	Test 1	Test 2
Raw / Calib	Raw	Calib
Datatype	Float32	Int16
Compression Speed (MB/s)	180	100
Decompression Speed (MB/s)	230	180
Compression Ratio	**9.0**	**3.7**

- Integer compression developed for ExaFEL in SZ
- Algorithm can be further optimized
- Plans to develop SZ on FPGAs

EXASCALE COMPUTING PROJECT

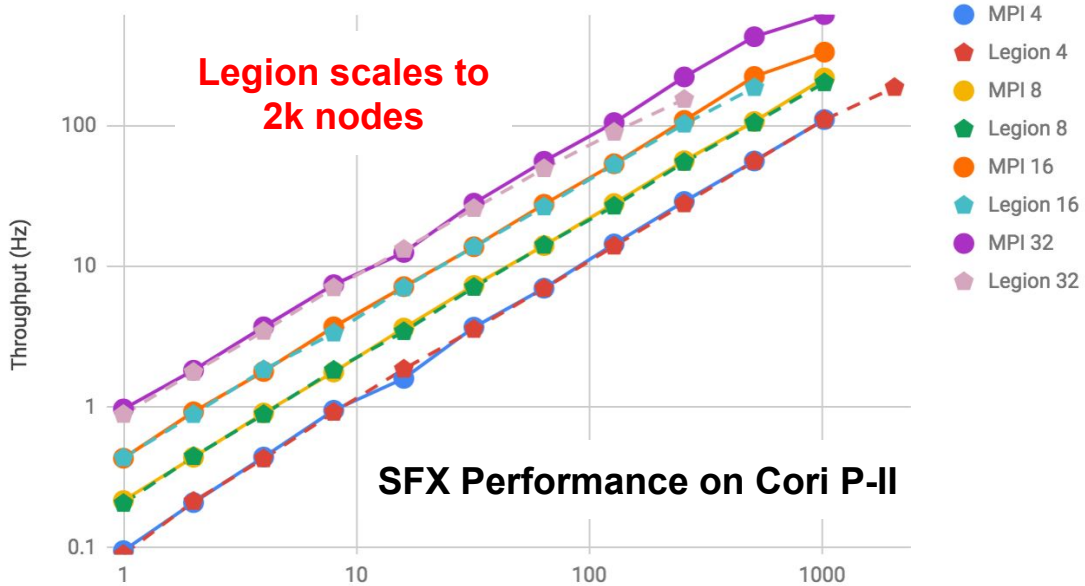
Psana-tasking: Progress and Next Steps

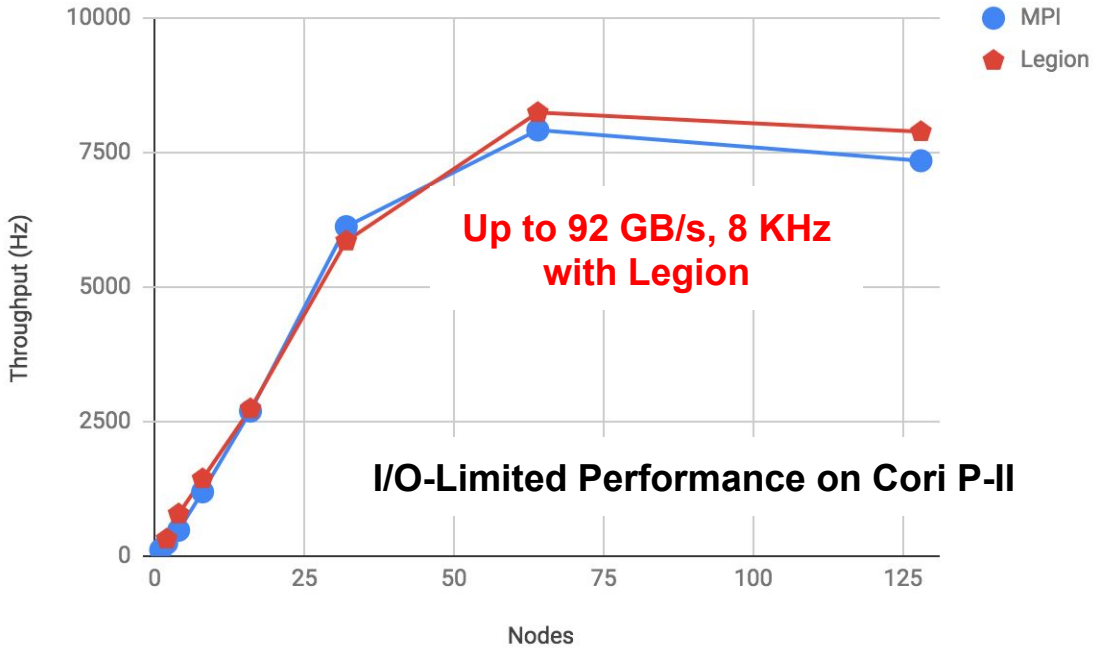
Progress:

- Port to Cori PII and Summitdev
- Contributions to Legion (STPM10):
  - Expanded Python support
  - First implementation of **lifeline load balancing**
- Ported SFX demo to psana-tasking
- Scaled SFX demo to **2K nodes on Cori P-II**
- Achieved **8 KHz** data rate in I/O limited case
- Use of **GASNet-EX** (STPM17) enabled scaling to 32 cores per node
- Support for **GPU tasks** in psana-tasking

Next steps:

- Scale to **full machine** on Cori and/or Summit
- More integration using GASNet-EX to further **improve memory usage and scalability**
- Complete Legion support for **multiple Python interpreters per runtime**





Nanocrystallography: Progress and Next Steps

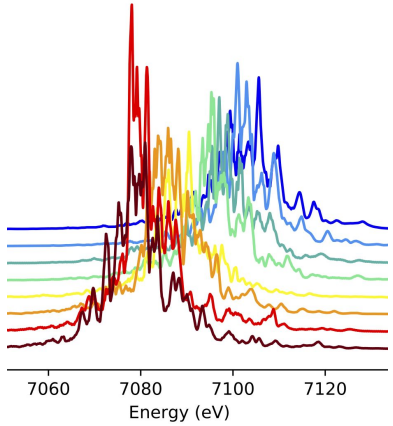
Progress:

- Port to Cori PII and Summitdev. Completed profiling on KNL (together with CODAR team) .
- Found optimal algorithm for iterative non-linear least squares parameter optimization (with **Strumpack** team)
- **Data merging**: MPI-based parallelism to distribute the workload (completed Oct 2018)
- **IOTA indexing:** higher success rate for processing diffraction patterns (completed Oct 2018)

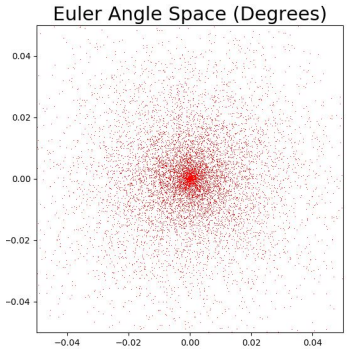
Next steps:

- Bragg spot integration
  - use more detailed physical models to achieve (1%) accuracy, enabling new science (time resolution, metalloenzyme spectroscopy, conformational dynamics of proteins)
- **Approach (pixel by pixel "ray tracing")**: physical parameters → simulate diffraction adjust parameters → simulation fits data
  - SIMTBX (SIMulation ToolBoX) implemented in 2017.
    GPU & OpenAcc ports created in 2018.
    Will incorporate into data processing Year 3.

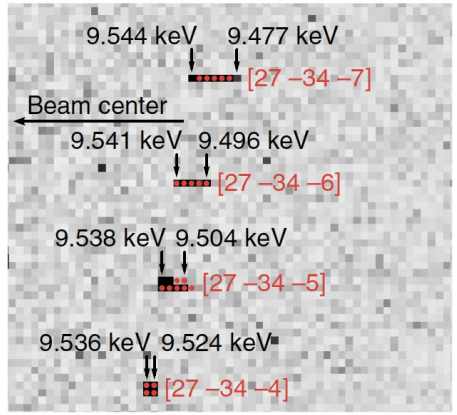
Physical Modeling



Unique X-ray spectrum for each shot



Each crystal is an ensemble of mosaic domains



Result: Each Bragg spot has a unique shape and size

This becomes Exascale:
1.5 core-hours per simulation, single process
5000 diffraction images/second at LCLS-II

EXASCALE COMPUTING PROJECT

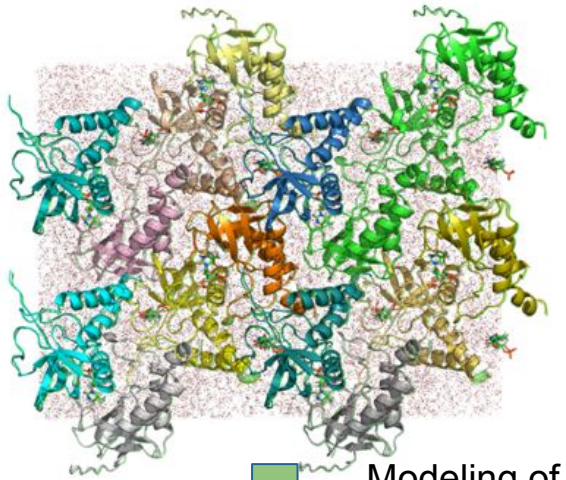
Diffuse scattering: Progress and Next Steps

Progress:

- Implemented parallel diffuse scattering data processing pipeline in C using MPI and OpenMP
  - Code publicly released at https://github.com/mewall/lunus
- Processed a SFX dataset collected at LCLS
  - 317 Rayonix LCLS diffraction images
- Achieved 100 Hz frame rate on 250 nodes of Cori KNL
  - 2,000 Pilatus 6M synchrotron diffraction images

Next steps:

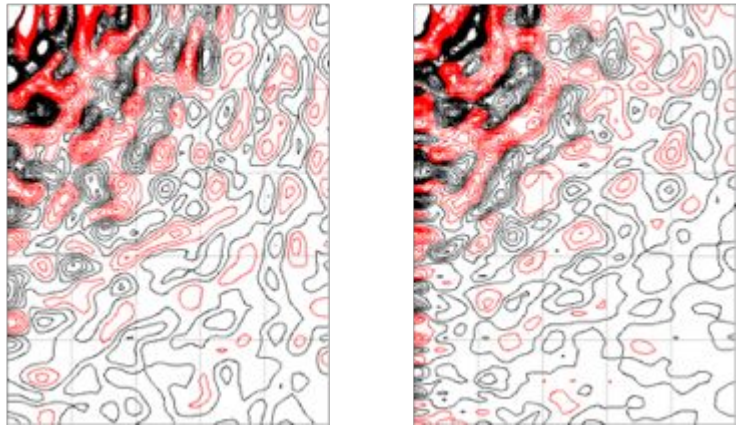
- Improve on-node performance
  - Mode filter using GPU target with improved algorithm
  - Threaded orienting of individual diffraction images and accumulation of intensity values in 3D
- Adapt Lunus for multi-panel detector (e.g. CS-PAD)
- Further scale to process 100,000 images at 5 kHz
- Python wrappers for Lunus for integration with CCTBX and psana ⇨ **e2e diffuse scattering pipeline**



Modeling of diffuse data in terms of protein dynamics

Wall, IUCrJ 5 (2018) 172

MD Model (unit cell) | Data



Fourier Transform of Diffuse Intensity

EXASCALE COMPUTING PROJECT

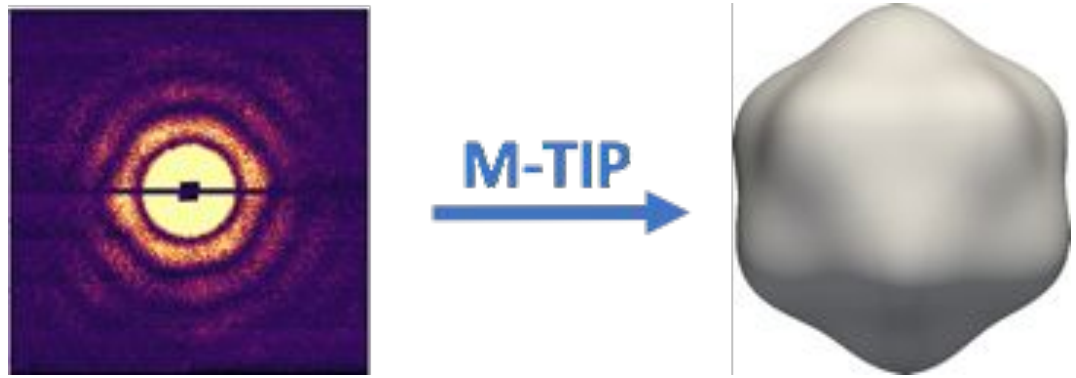
Single Particle Imaging: Progress and Next Steps

Progress:

- **Successful 3D reconstruction** of RDV and PR772 viruses from experimental LCLS SPI data using M-TIP
- **Designed new Cartesian** to Non-uniform framework to replace the current polar framework in M-TIP
  - Based on efficient inversion of a non-uniform fast Fourier transform (NUFFT) via the LSQR algorithm
  - **Improves scalability** - New approach can be fully parallelized over all images, whereas the old polar approach was done one image at a time
  - **Reduces complexity** from $O(DN^{4/3})$ to $O(D + N)$ (D = # data points, N = # grid points)

Next steps:

- Combine LSQR approach to NUFFT inversion with Cartesian version of M-TIP
- Develop resolution-adaptive local orientation matching to further increase scalability
- Integrate M-TIP with the PSANA framework ⇨ **e2e single particle imaging pipeline**



12nm reconstruction of an RDV virus from experimental LCLS data under icosahedral symmetry constraints

EXASCALE COMPUTING PROJECT

Resource Orchestration: Progress and Next Steps in the SDN Data Path over ESnet

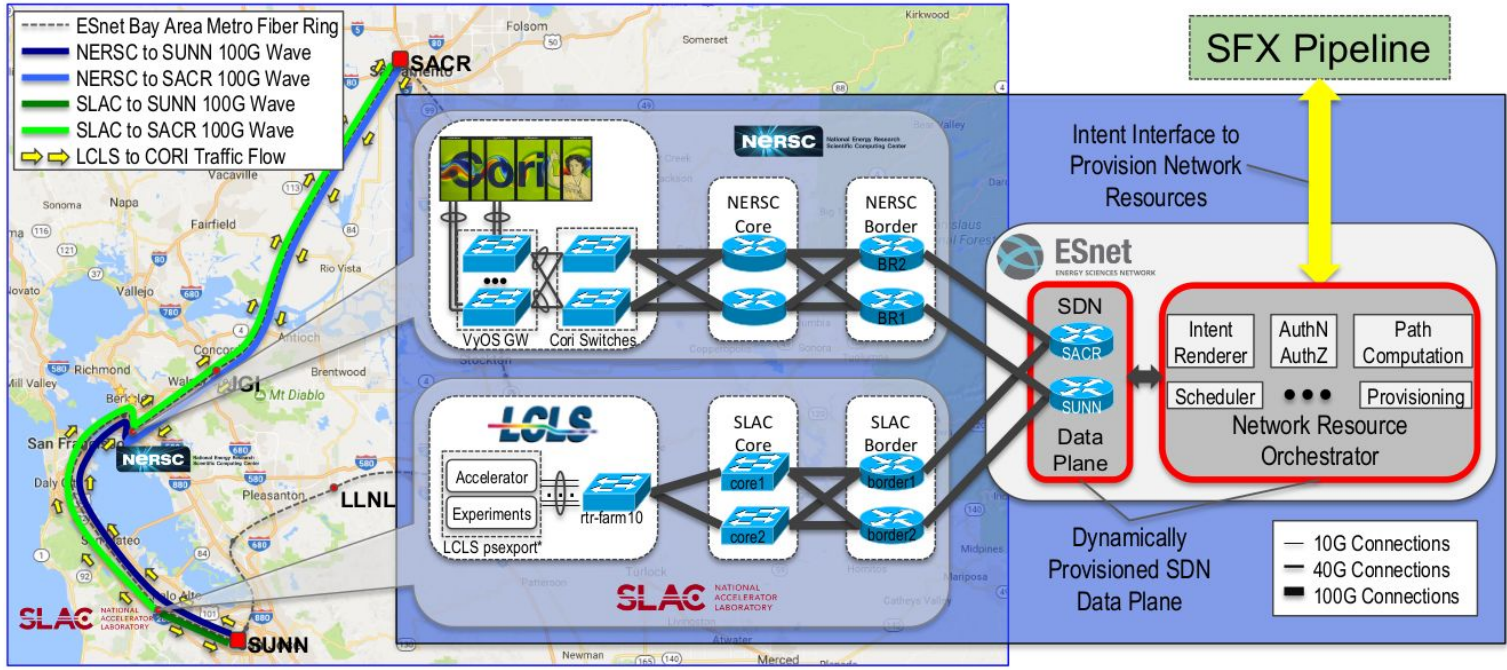
- Network Resource Orchestrator exposes an intent interface to accommodate network bandwidth scheduling requests, allowing science application workflows to interact with the network to support coordination of network, compute, and instrument resources
- SDN (software defined networking) data plane enables the dynamic reconfiguration of the network to provide bandwidth guarantees to support predictable and repeatable data transfer times

Progress:

- Network provisioning intent API is complete and tested, along with prototype client code to exercise API

Next steps:

- Integrating network provisioning client into workflow



EXASCALE COMPUTING PROJECT

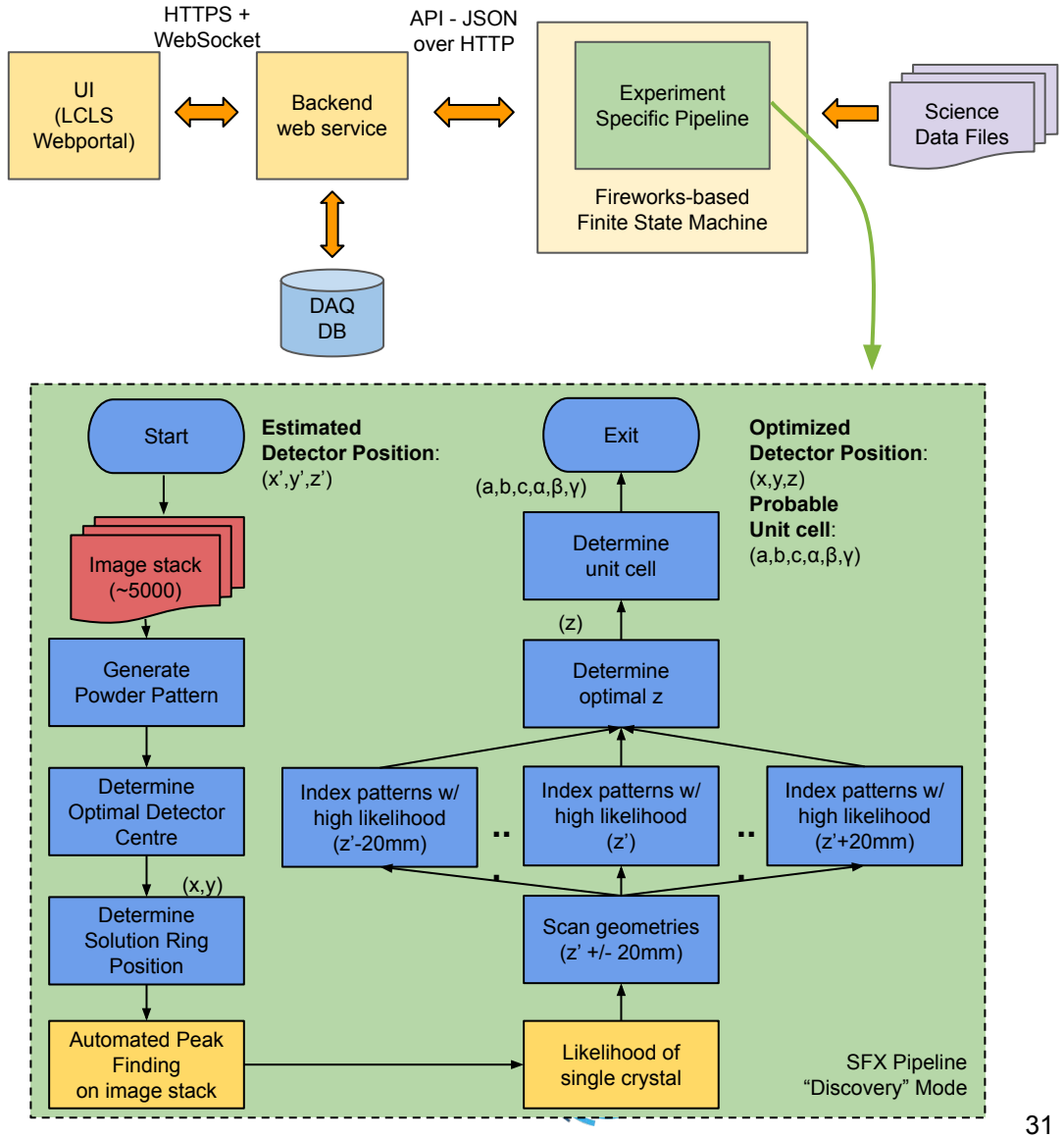
Resource Orchestration: Progress and Next Steps in Workflow Automation

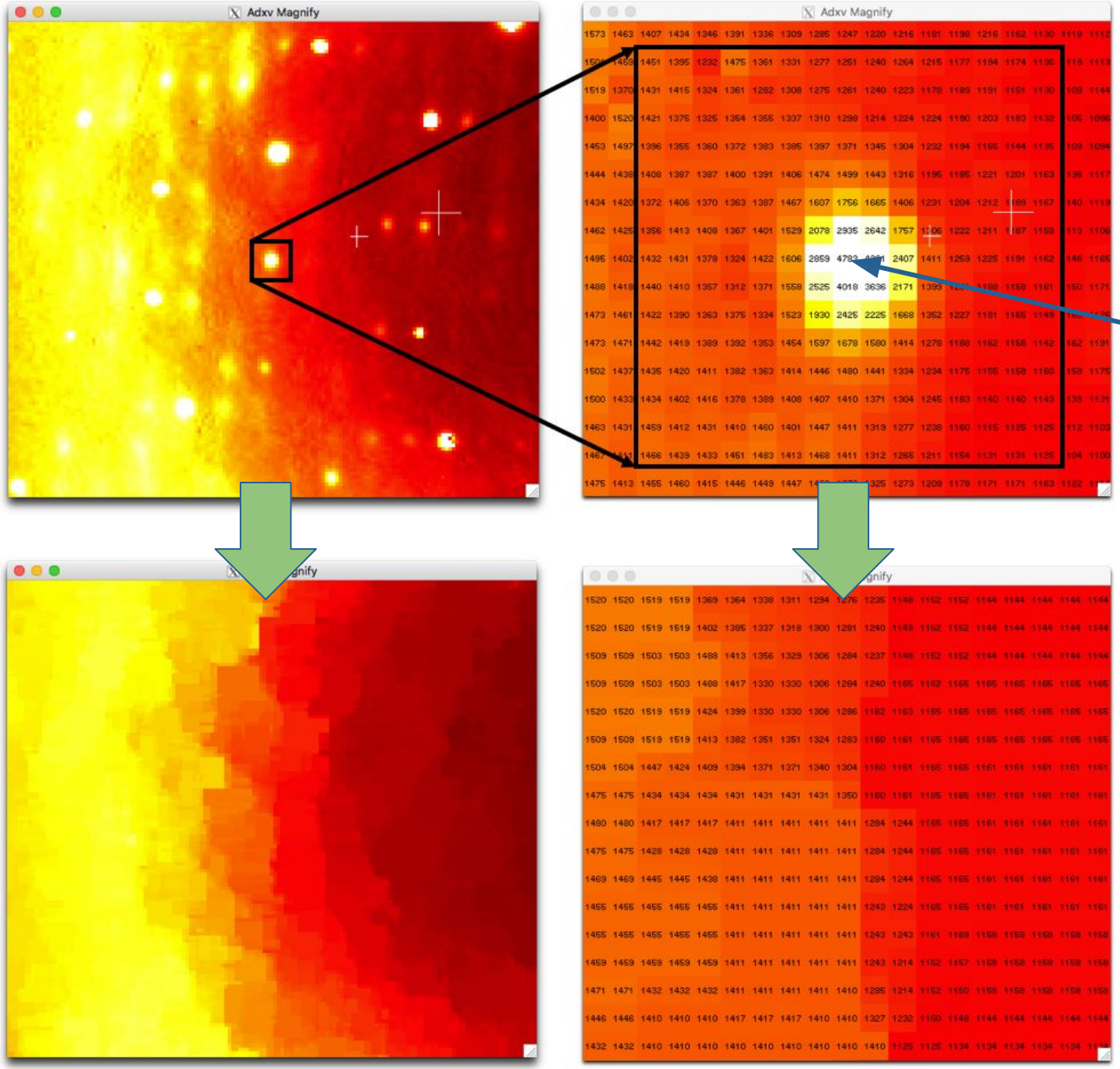
Deployed mechanism for **automating the execution of the analysis**

- Analysis execution **synchronized with data taking** (through the DAQ database)
- Ability for the experimenters to **monitor and control** the workflow through the web portal (aka electronic logbook)

Next steps:

- Make the workflow **more robust and better documented**





Mode Filter

- A method to remove sharp peaks from diffraction images
- Draw a box around each pixel in the image
  - A typical box width is 15-20 pixels
- Replace the central pixel value with the most common value in the black box (the **mode**)
- Use resulting images, e.g., for computing scale factors in merge of diffuse scattering data
- Threaded implementation by calculating the mode in parallel for different pixels
  - Working on GPU implementation

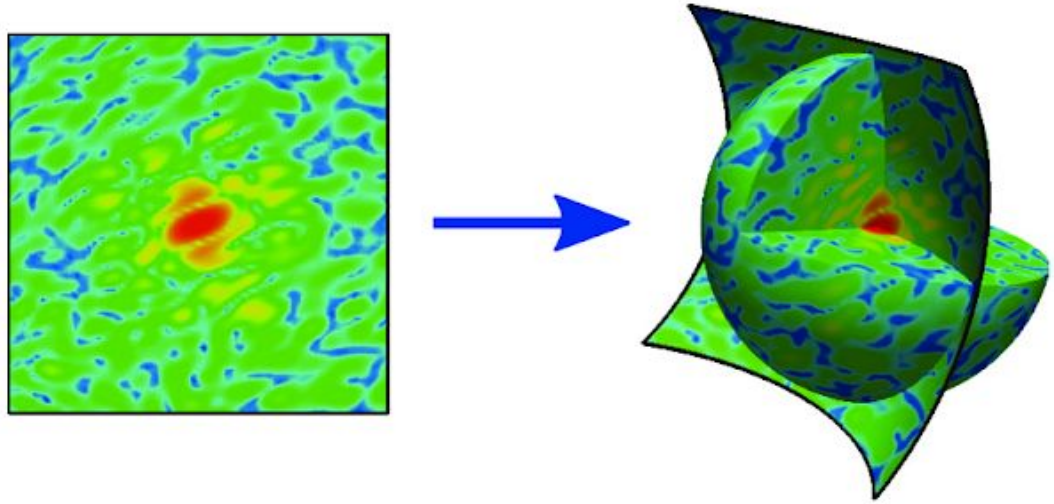
EXASCALE COMPUTING PROJECT

Single-Particle Imaging Reconstruction Problem

Single-Particle Diffraction Images:
Image $J^{(k)}$ samples I along a spherical slice rotated according to the image orientation R_k:

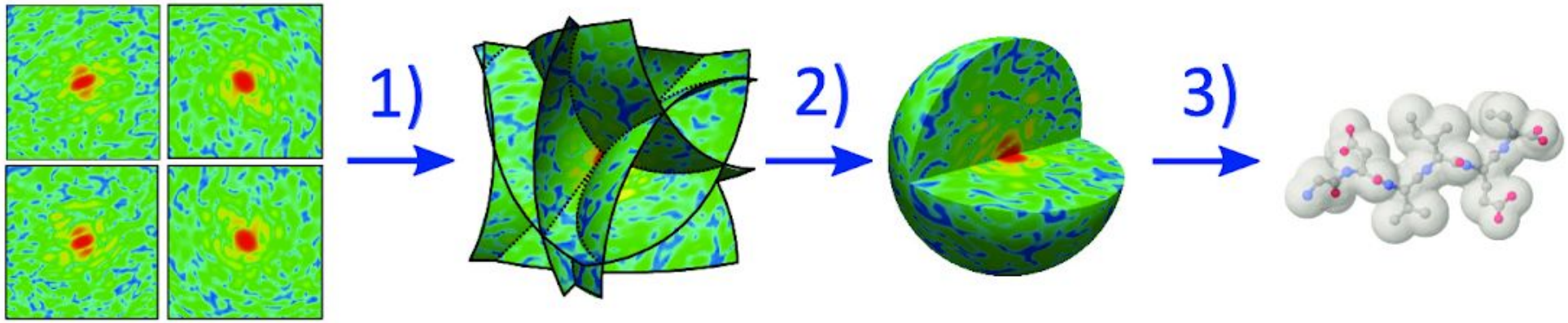
$$J^{(k)}(q,\phi) = I^{(R_k)}(q,\theta(q),\phi),$$

where $\theta(q) = \arccos(q\lambda/2)$.



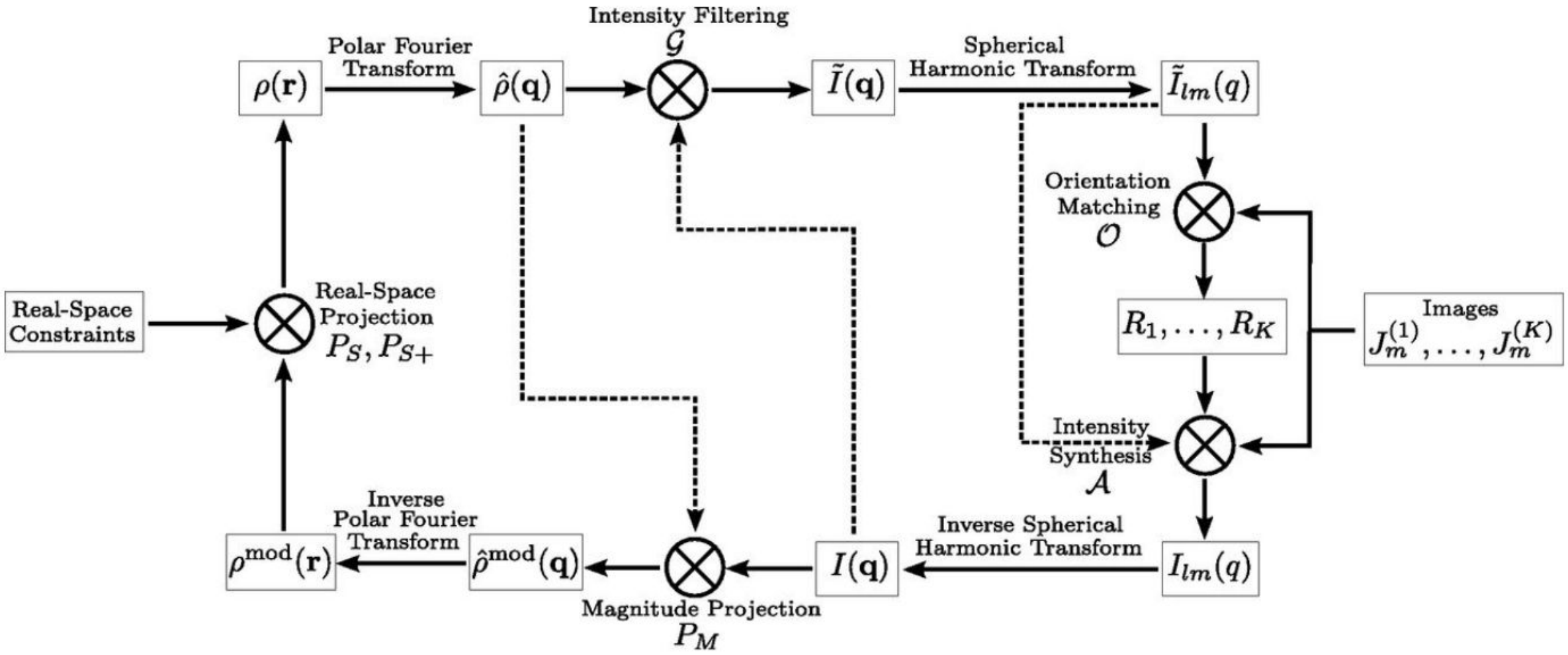
Challenges:

1) Orientation Problem: Determine the orientation R_k of each image $J^{(k)}$.
2) Intensity Reconstruction: Extract the 3D intensity function I from the set of images.
3) Classical Phase Problem: Reconstruct the electron density ρ from the intensity function I.



EXASCALE COMPUTING PROJECT

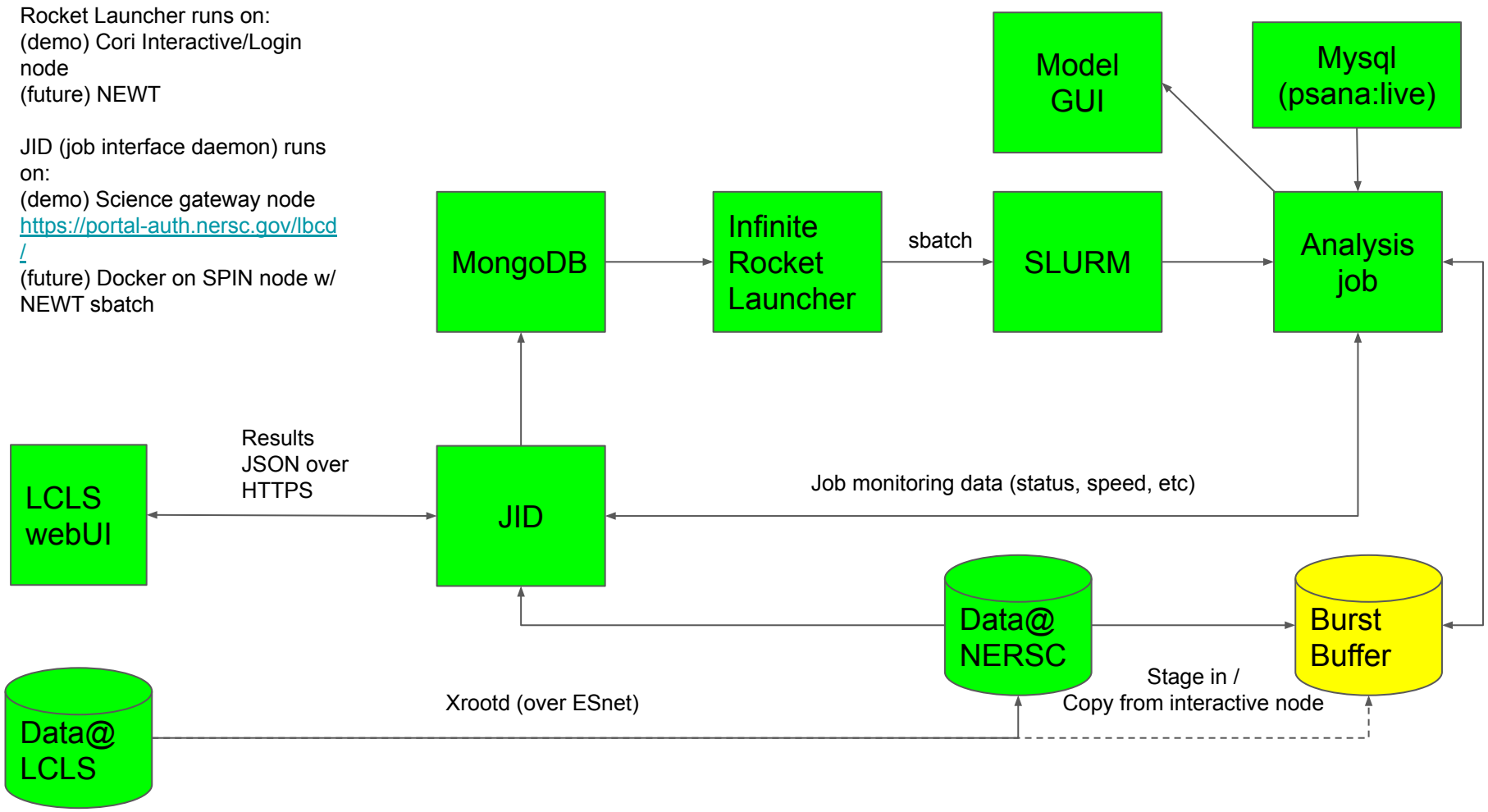
MTIP framework outline



ExaFEL Workflow for the demo

Deployed mechanism for **automating the execution of the analysis**

- Analysis execution **synchronized with data taking** (through the DAQ database)
- Ability for the experimenters to **monitor and control** the workflow through the web portal (aka electronic logbook)



EXASCALE COMPUTING PROJECT