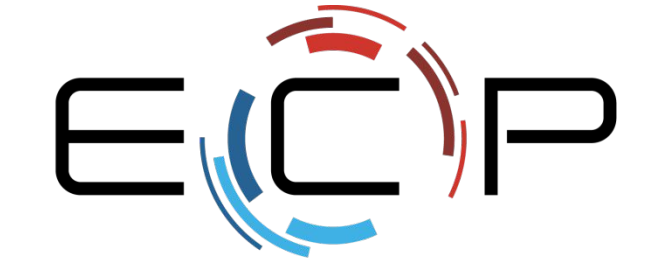
# Data Analytics at the Exascale for Free Electron Lasers

**Amedeo Perazzo**
**JLAB Computing Round Table, October 2nd 2018**




NNSA National Nuclear Security Administration
U.S. DEPARTMENT OF ENERGY | Office of Science

# Outline

- Brief introduction to **FEL science**
  - 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: psana-tasking and Legion
- **Progress and next steps**: psana, SFX, SPI, resource orchestration

# Brief Introduction to FEL Science

# Data Analytics for High Repetition Rate Free Electron Lasers

## FEL data 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 analysis

LCLS-II will increase **data throughput by three orders of magnitude** by 2025, creating an exceptional scientific computing challenge

# 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*

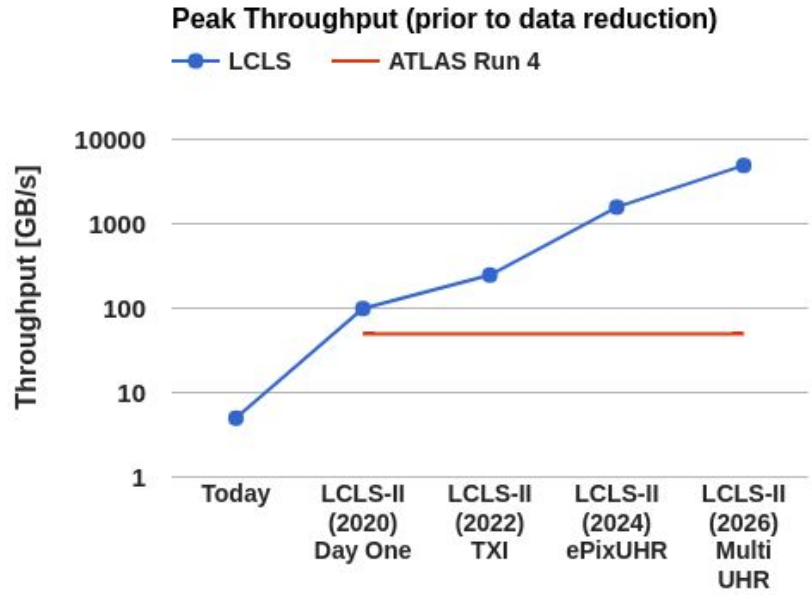
**Example data rate for LCLS-II (early science)**

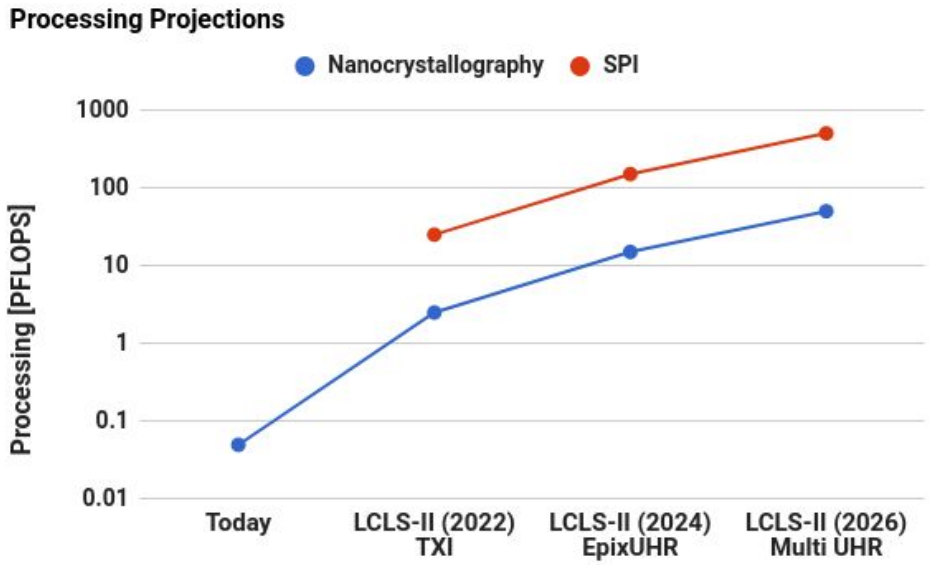
- 1 x 4 Mpixel detector @ 5 kHz = **40 GB/s**
- 100K points fast digitizers @ 100kHz = **20 GB/s**
- Distributed diagnostics 1-10 GB/s range

**Example LCLS-II and LCLS-II-HE (mature facility)**

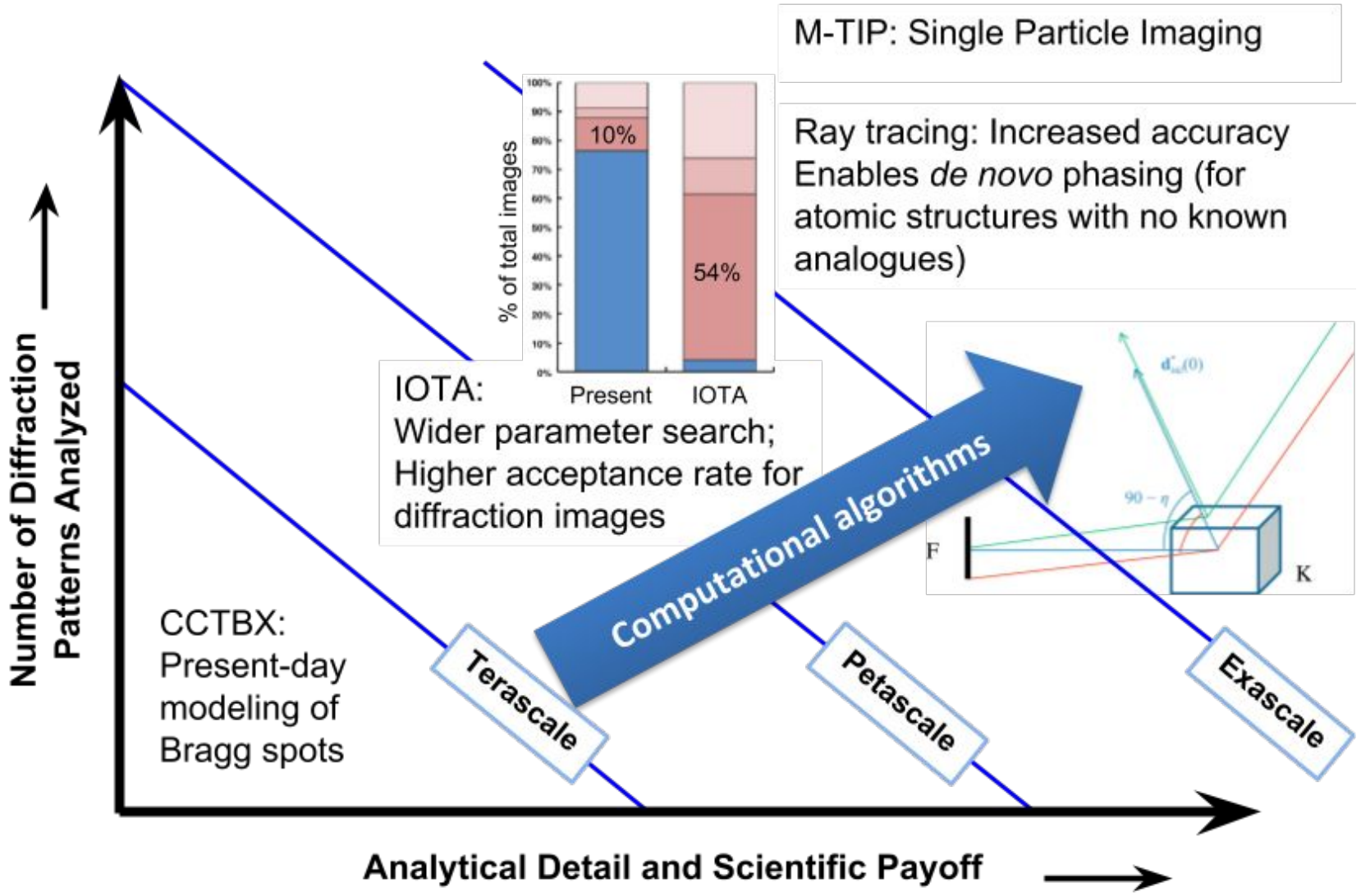
- 2 planes x 8 Mpixel ePixUHR @ 50 kHz = **1.6 TB/s**

**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

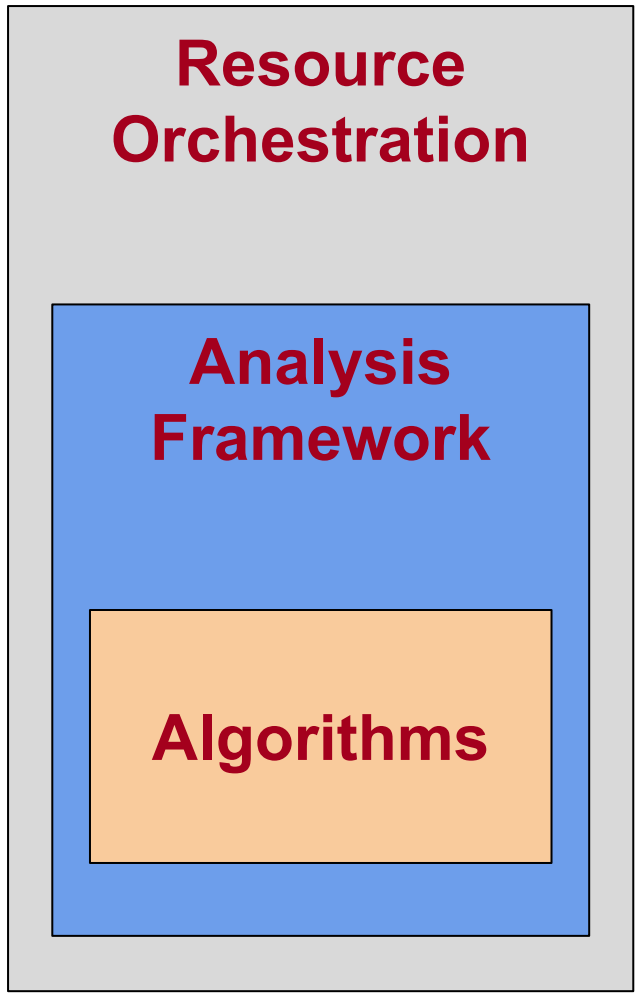



**Exascale vastly expands the experimental repertoire and computational toolkit**

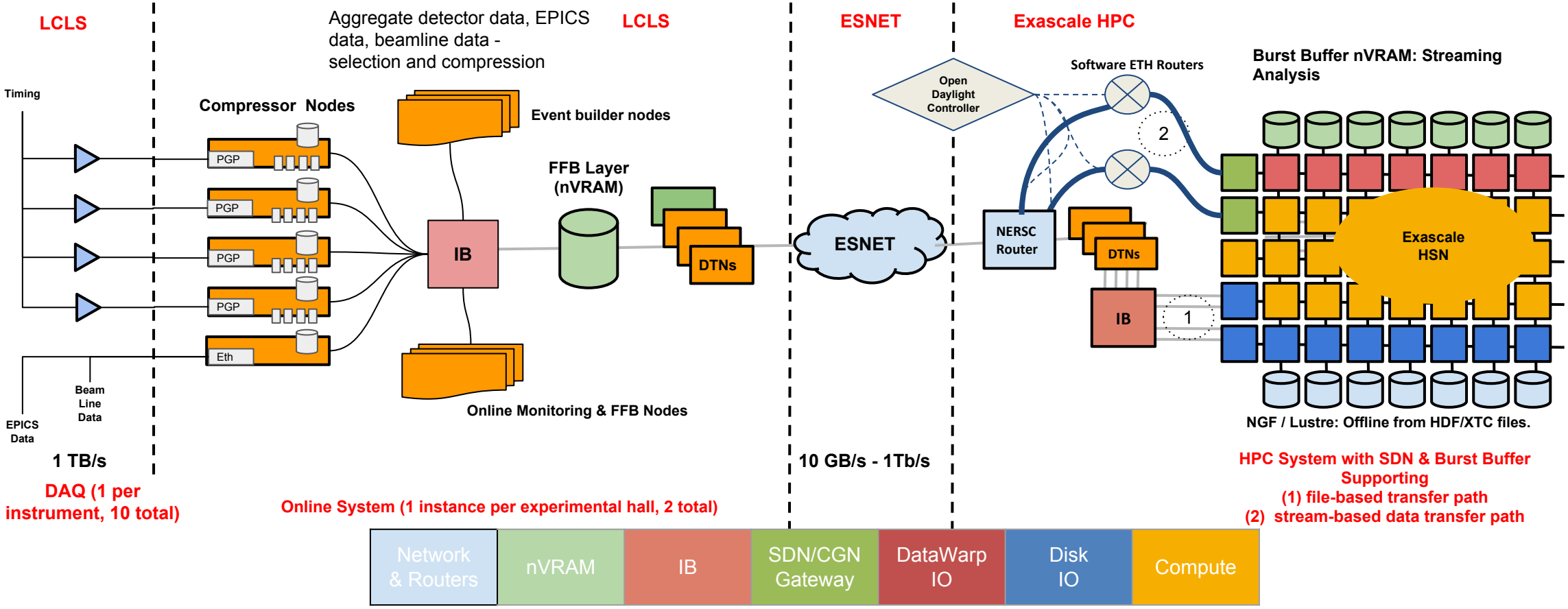
EXASCALE COMPUTING PROJECT

# ExaFEL Project Plan

1. **Algorithmic improvements** for high data throughput experiments
   Develop **exascale optimized algorithms** and port existing algorithms to exascale architectures
2. Port LCLS **data analysis framework** to supercomputer architecture, allow scaling from hundreds of cores (today) to hundreds of thousands of cores
   LCLS analysis framework (psana) handles **parallelization, calibration, input/output** ⇨ science specific algorithms run within psana
   Integration of psana with **Legion** framework for scalability, portability, input/output optimization
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 analysis on the supercomputer
   **Visualize** the results of the analysis back to the experimenters in quasi real time



# ExaFEL Data Flow



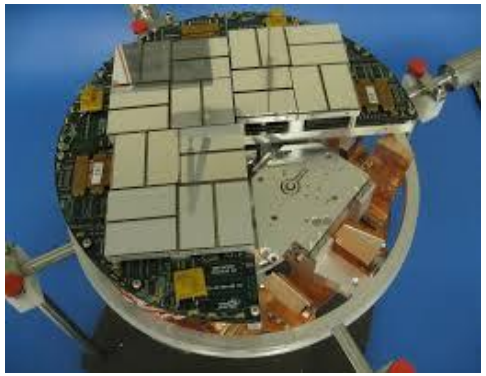
# Example of computing intensive algorithms for ExaFEL: Scaling the nanocrystallography pipeline

- **Avoidance of radiation damage** and emphasis on **physiological conditions** requires a transition to fast (fs) X-ray light sources & large ($10^6$ image) datasets

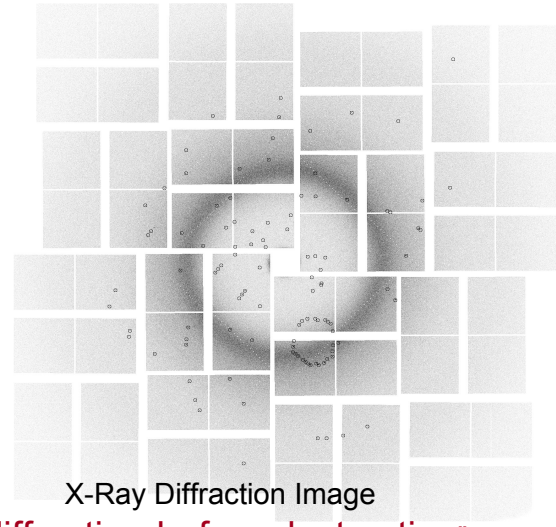
- Main steps in the algorithm are
  (1) identifying the Bragg diffraction spots
  (2) deducing the geometry of the lattice repeat,
  (3) refining the model again
  (4) summing the X-ray signal in each spot for further analysis

- IOTA being added to CCTBX in years 2-3 to improve success rate for (2)

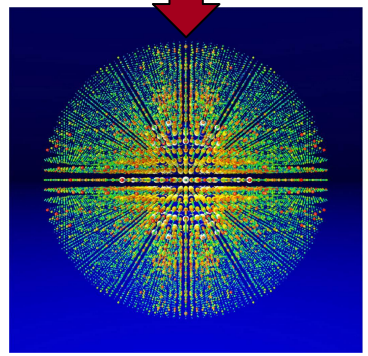
- Ray tracing being added to improve the modeling detail for (4)



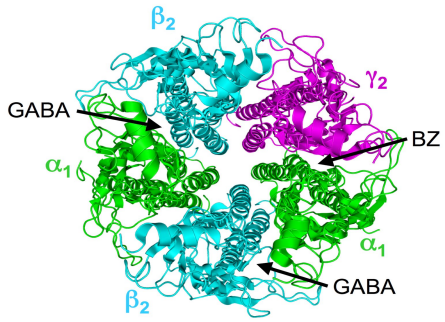
Megapixel detector



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



Intensity map (multiple pulses)

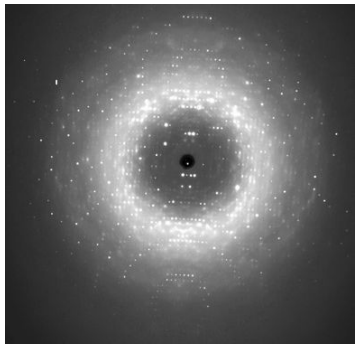



Electron density (3D) of the macromolecule

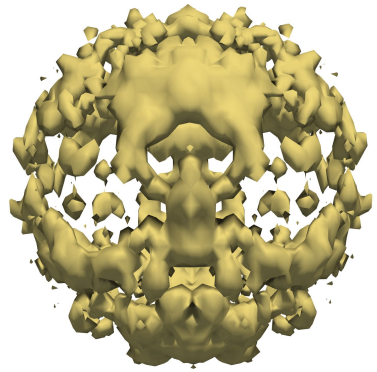
EXASCALE COMPUTING PROJECT

# Departure from a Perfectly Regular Lattice: Diffuse Scattering

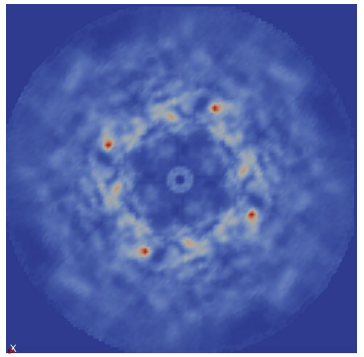
- **Intensity away from the Bragg peaks** that can be obtained simultaneously and used to derive information about **protein dynamics**
- Main differences with respect to nanocrystallography are (1) measure the intensity away from the Bragg peaks; (2) obtain scale factors from diffuse data instead of Bragg peaks; (3) sample on a finer grid than the Bragg lattice; (4) model and interpretation using protein dynamics instead of the average structure
- The most important data processing bottleneck is (2) which involves computing mode filtered diffraction images. Modeling is also computationally intensive.



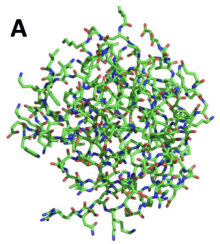
X-Ray Diffraction Image



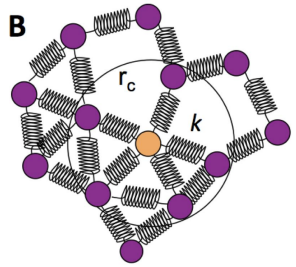
3D Diffuse Intensity Map (Isosurface)



3D Diffuse Intensity Map (Slice)




Atomic Model from Crystallography

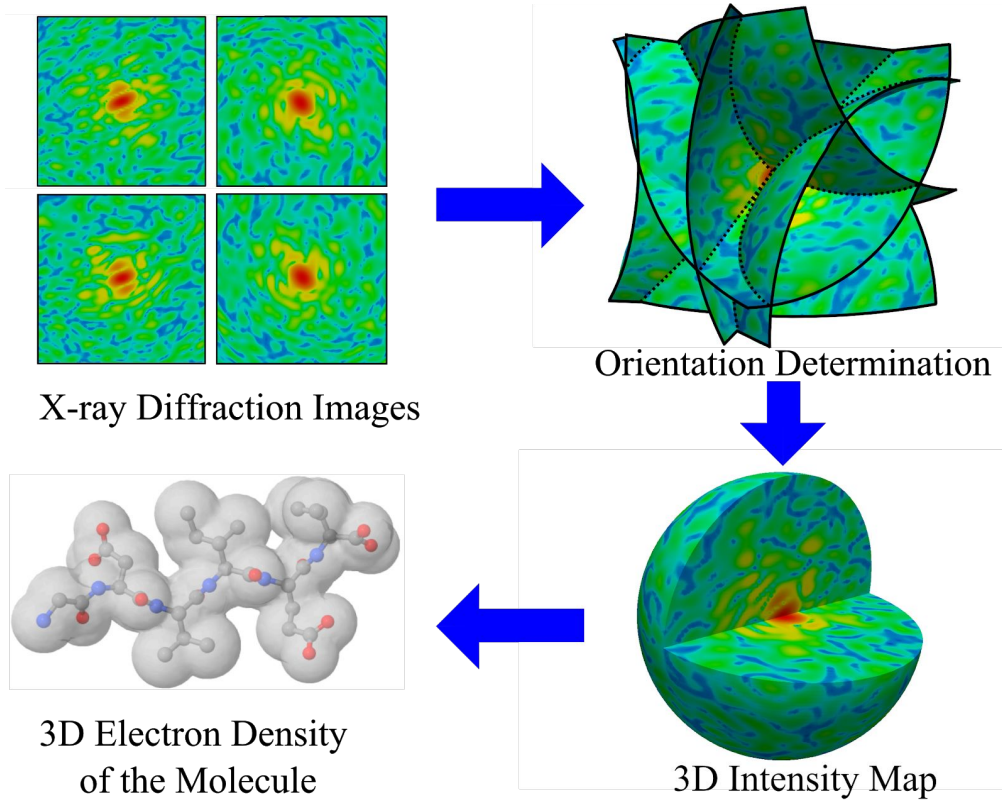



Dynamical Model of the Macromolecule

$$D_{NM}(\mathbf{s}) = \sum_i \sum_j f_i f_j^* e^{-4\pi^2\left(\sigma_i^2+\sigma_j^2\right)s^2}\left(e^{-4\pi^2 s^2 \phi_{ij}} - 1\right)$$

# 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
- Algorithm published in PNAS [1]
- Modular approach allows modifications to model systematic issues in data and/or incorporate additional 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.*

# 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)

# ExaFEL KPPs: Plans & Achievements

Exascale

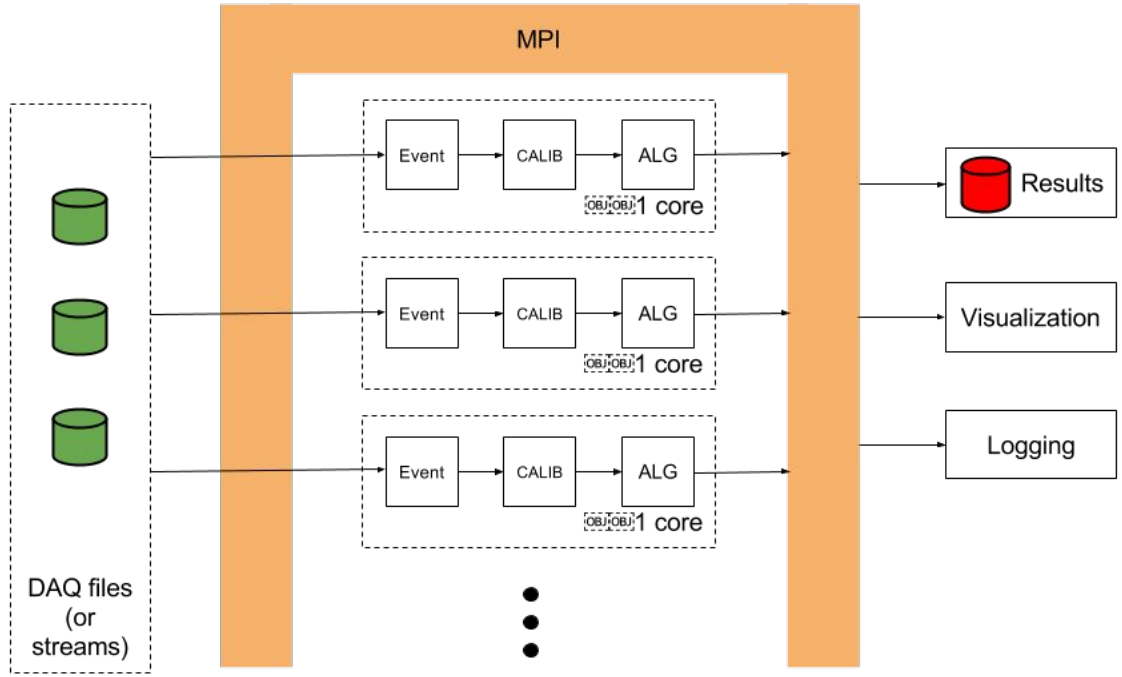
| | | FY17 | FY18 | FY19 | FY20 | FY21 | FY22 |
|---|---|---|---|---|---|---|---|
| **Algorithm** | **Expected Ratio to SFX** | **Cori PII (30 PF)** | **Cori PII (30 PF)** | **Cori PII (30 PF) Summitdev** | **NERSC-9 (>60PF) Summit (200PF)** | **NERSC-9 (>60PF) Summit (200PF)** | **A21 (1EF)** |
| SFX | x1 | **135 Hz (6%)** | **3 kHz (52%)** | 5 kHz | | | |
| SFX with IOTA | x2-x5 | | 100 Hz (10-25%) | | 5 kHz | | |
| SFX diffuse scattering | x2-x5 | | 100 Hz (10-25%) | | 5 kHz | | |
| SFX with Ray tracing | x10-x20 | | | 100 Hz (50-100%) | | 1 kHz | 5 kHz |
| SPI with M-TIP | x10-x20 | | | 100 Hz (50-100%) | | 1 kHz | 5 kHz |

EXASCALE COMPUTING PROJECT

# LCLS Data Analysis Framework

Main features LCLS data analysis framework (psana):

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
4. **Reduces data volumes** to laptop-level sizes in world-standard HDF5 data format




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

Most of the compute goes into the ALG box: what happens in this box is **specific to the experimental technique** (or even to the experiment)

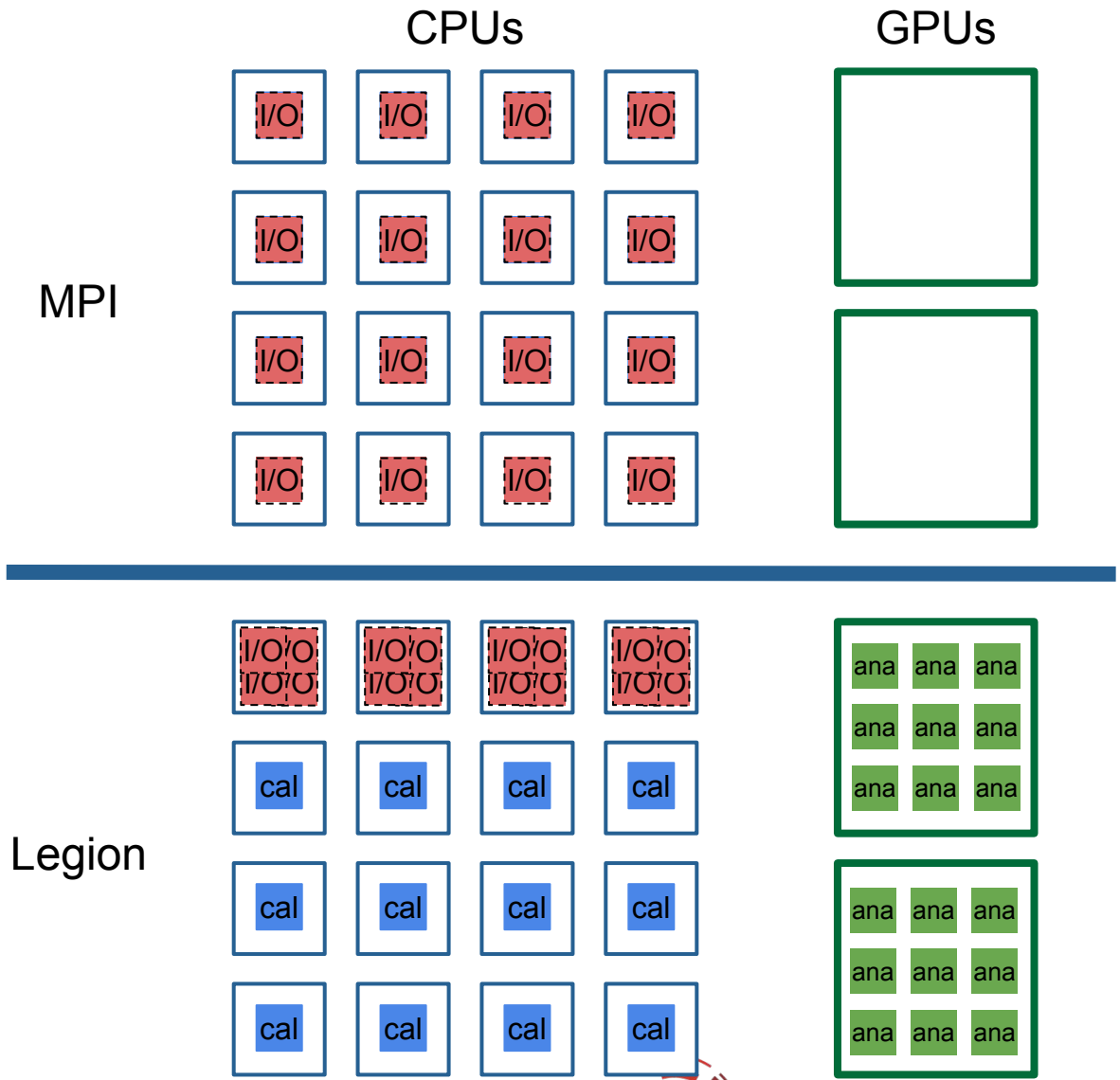
**Scalability is key**: the more time we spend in the ALG box for each event ⇒ the more cores we need to run in parallel to keep up with data taking rates

EXASCALE COMPUTING PROJECT

# Evolution of psana: psana-tasking and Legion

psana-tasking brings data analysis to the Legion task-based runtime system:

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



# Progress and next steps: psana, SFX, SPI, resource orchestration

# Nanocrystallography: Progress and Next Steps

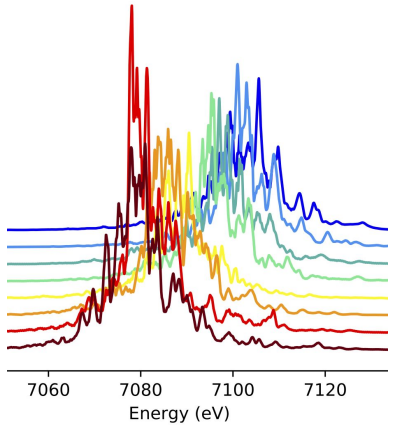
Progress:

- 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 (completion by Oct)
- **IOTA indexing:** higher success rate for processing diffraction patterns (completion by Oct)

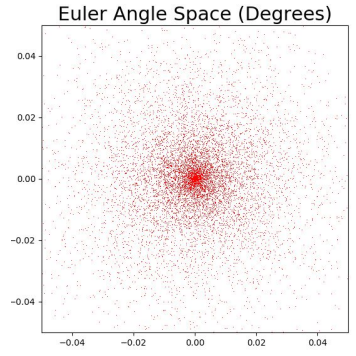
Next steps:

- 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.

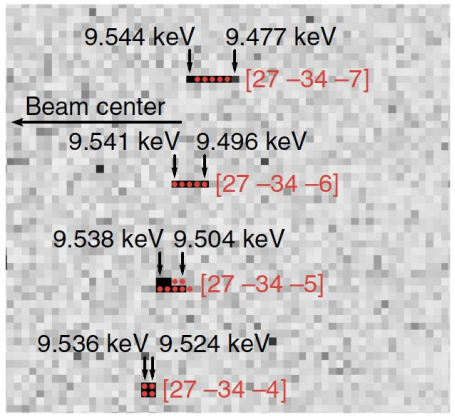
## Physical Modeling




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

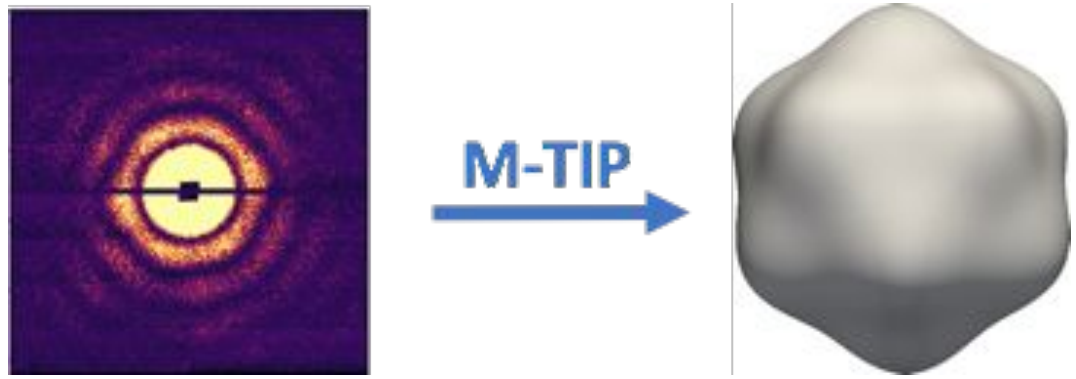
# Single Particle Imaging: Progress and Next Steps

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

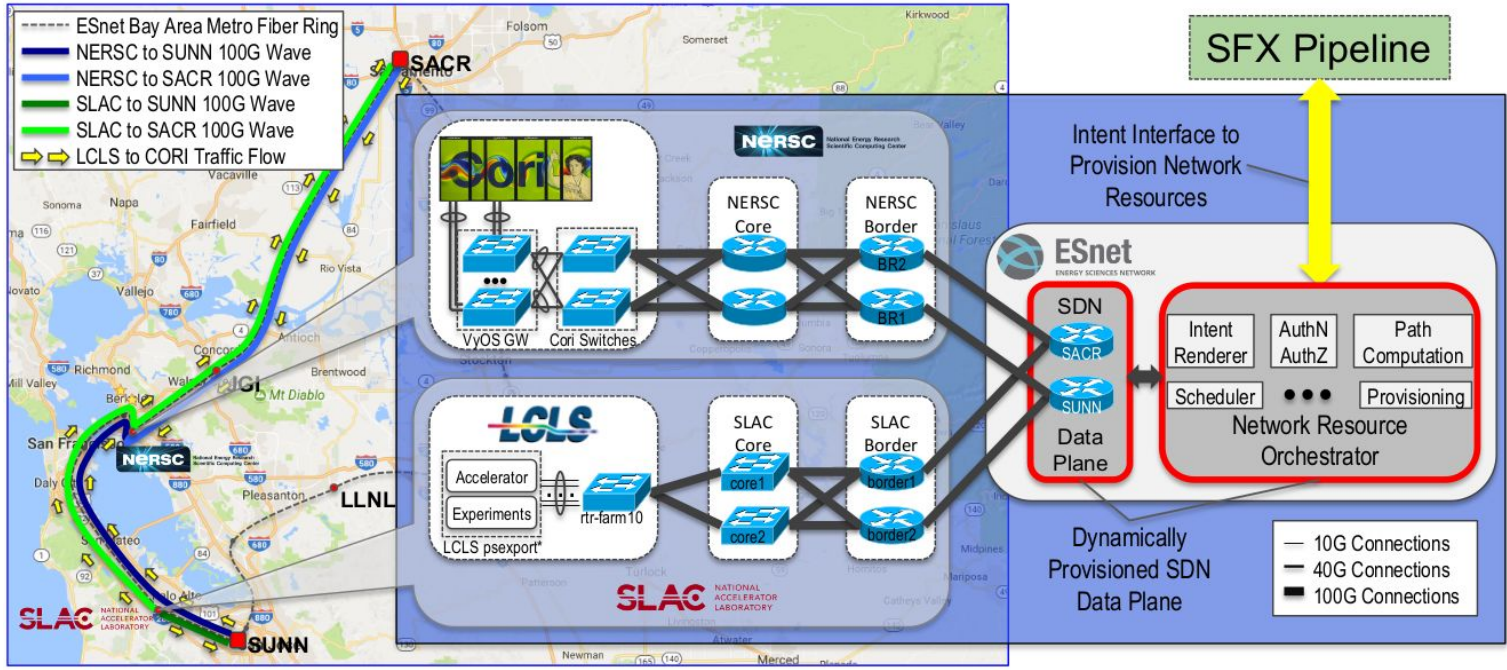
- Network Resource Orchestrator exposes an intent interface to accommodate network bandwidth scheduling requests, allowing science application workflows to interact with the network to support coordination of network, compute, and instrument resources
- SDN (software defined networking) data plane enables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



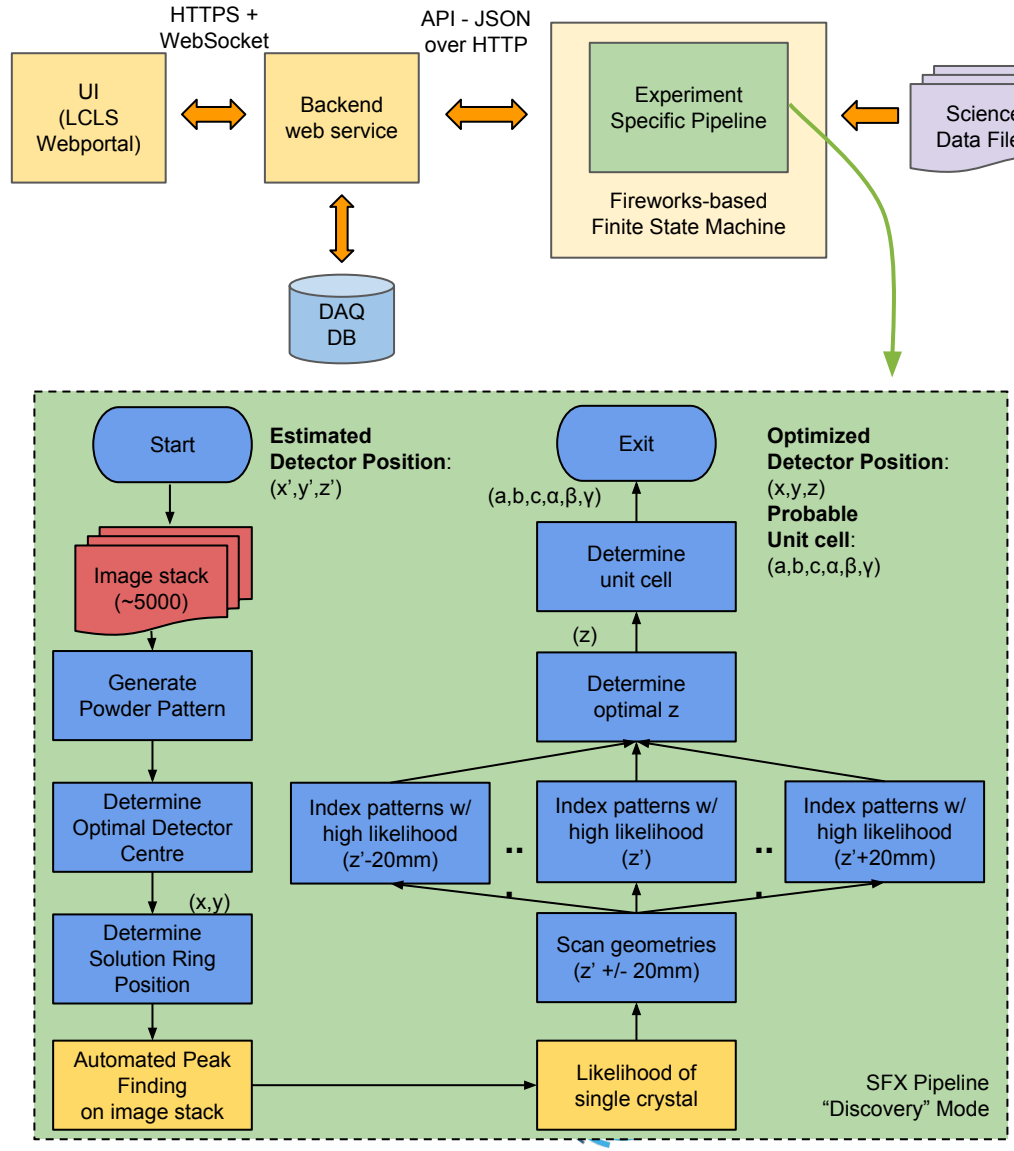
# Resource Orchestration: Progress and Next Steps in Workflow Automation

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**



# Summary

The end of the KNL line has changed our plans a bit, but it also made our path forward clearer:

> The main **focus of Y3 will be improving the strong scaling** (one event processed over many cores) of the more computing intensive algorithms (ray tracing, M-TIP) **on accelerated architectures**

# Backup Slides

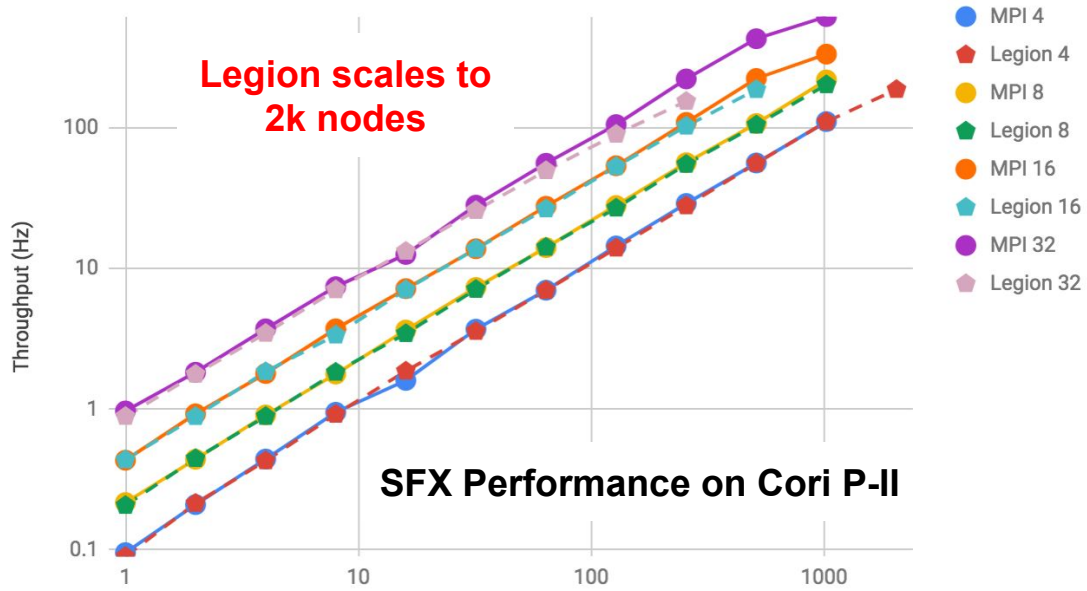
# Psana-tasking: Progress and Next Steps

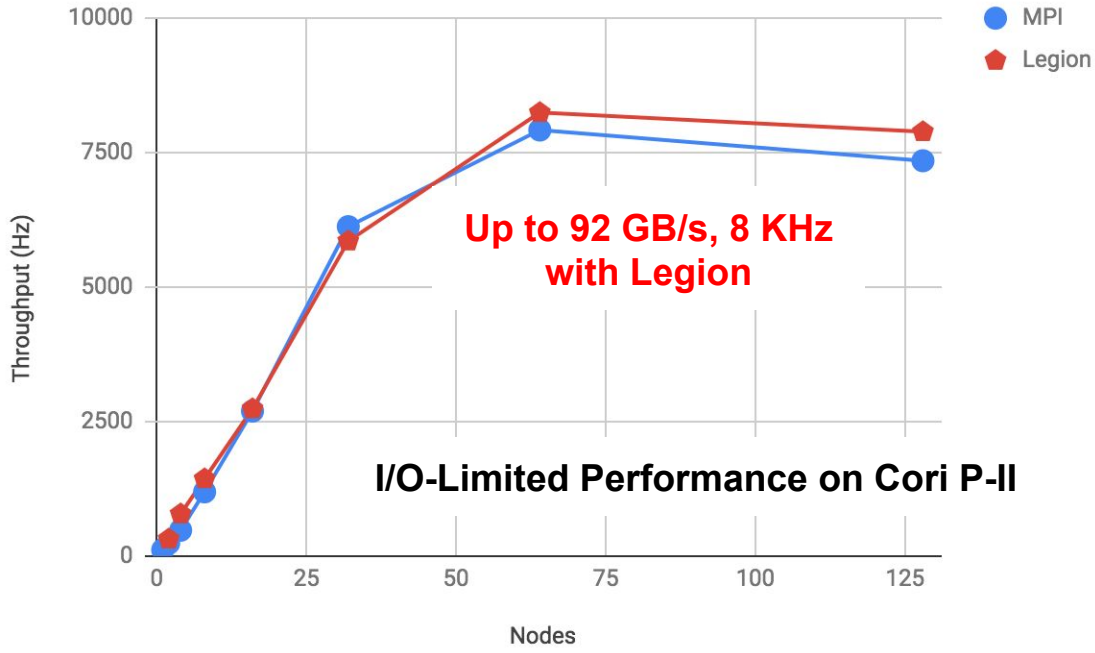
Progress:

- Port to Cori PII and Summitdev
- Contributions to Legion:
  - 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** 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**



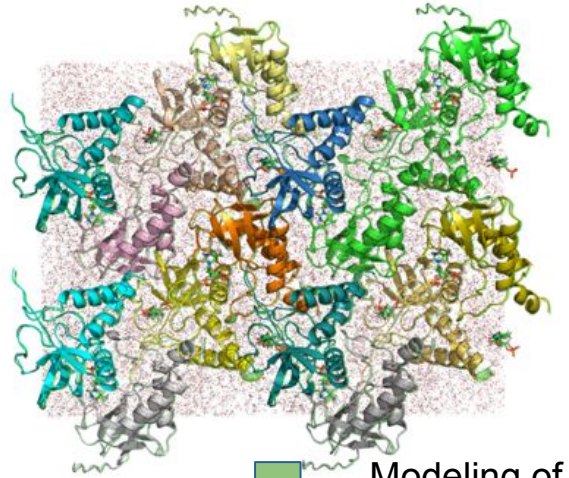
# Diffuse scattering: Progress and Next Steps

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

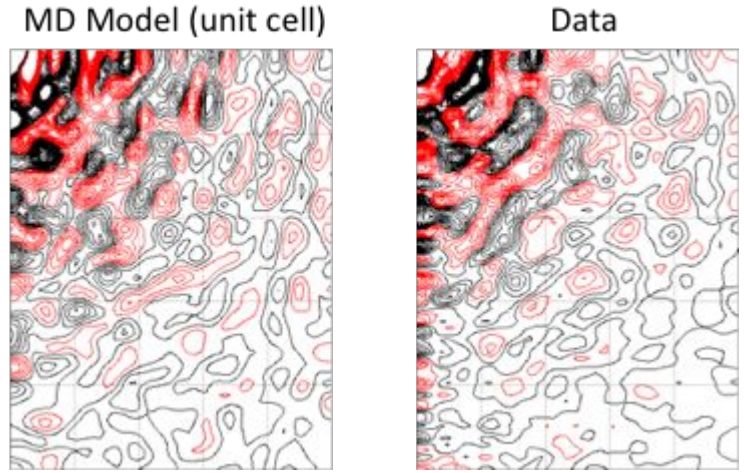
Next steps:

- 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

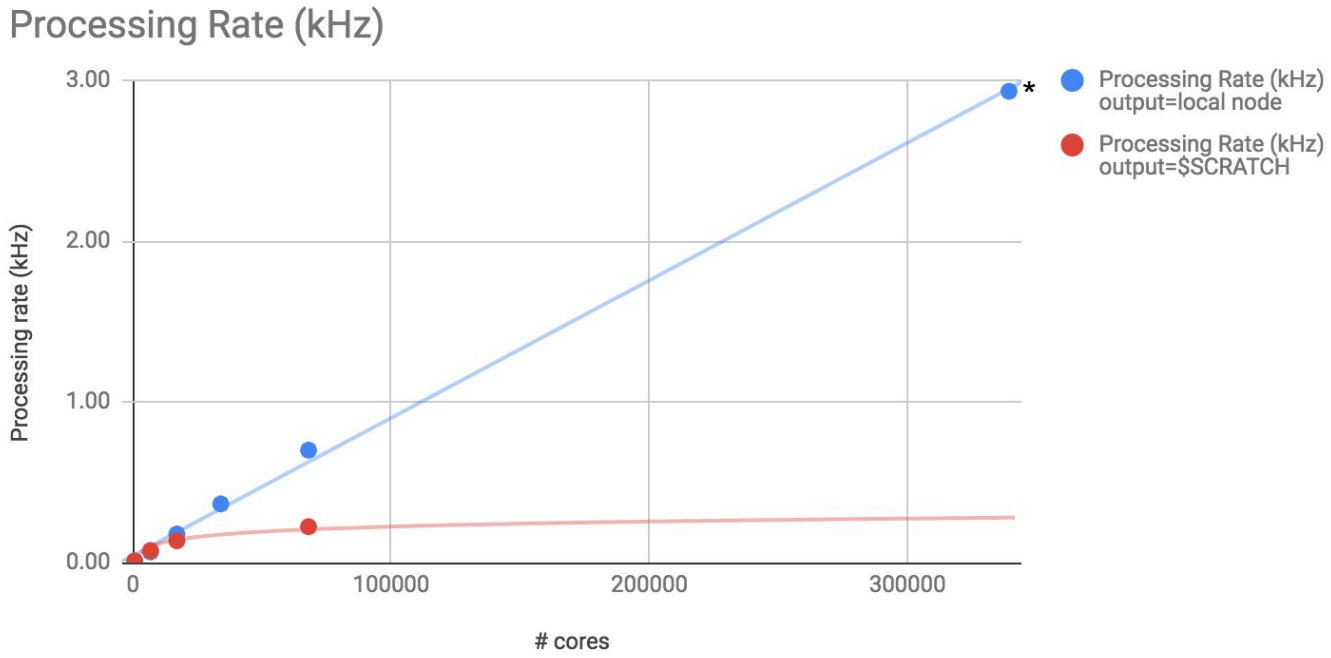



Fourier Transform of Diffuse Intensity

EXASCALE COMPUTING PROJECT

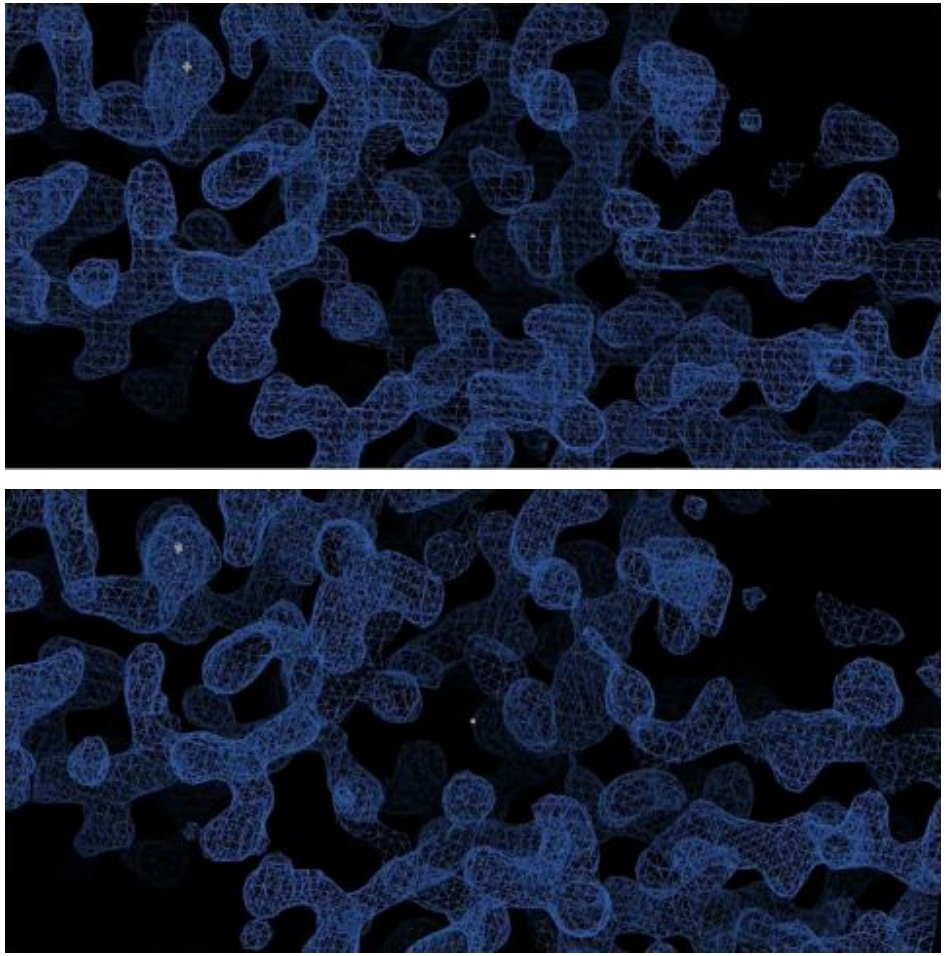
# Psana tasking optimization on Cori PII

**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.

# Lossy Compression (in collaboration with EZ ST team)

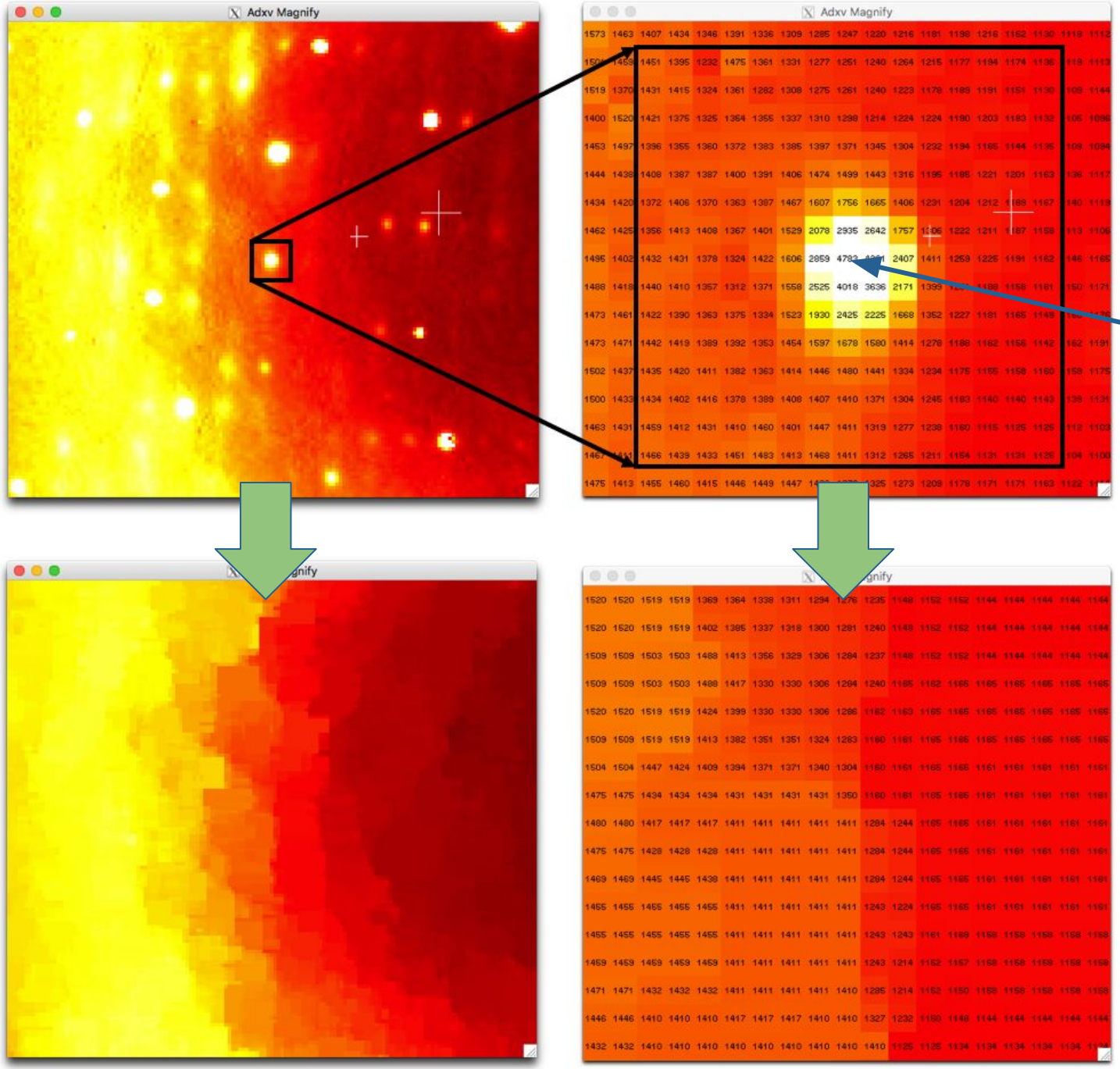


Simulated lossy compression shows Se-SAD can tolerate absolute error bound of 10 ADUs without any problems

**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




# Mode Filter

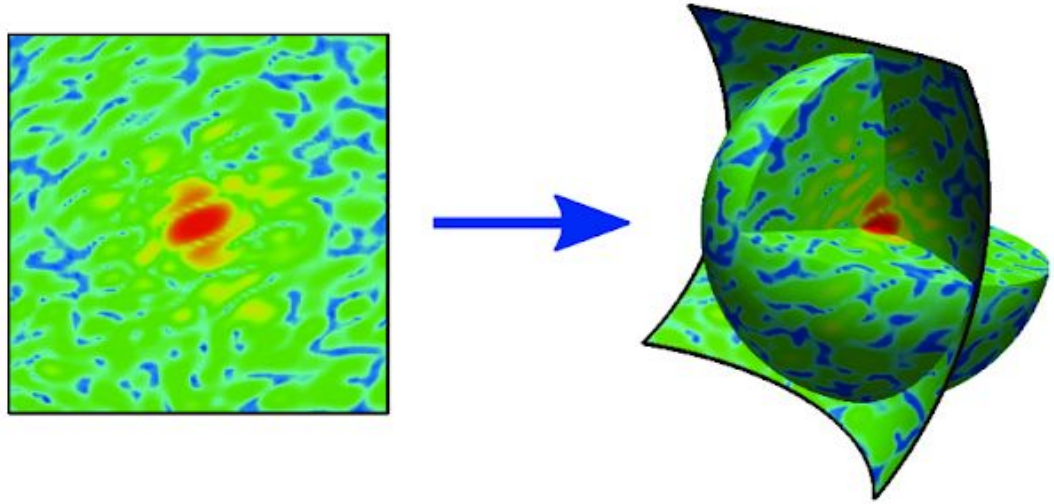
- A method to remove sharp peaks from diffraction images
- Draw a box around each pixel in the image
  - A typical box width is 15-20 pixels
- Replace the central pixel value with the most common value in the black box (the **mode**)
- Use resulting images, e.g., for computing scale factors in merge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

# Single-Particle Imaging Reconstruction Problem

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

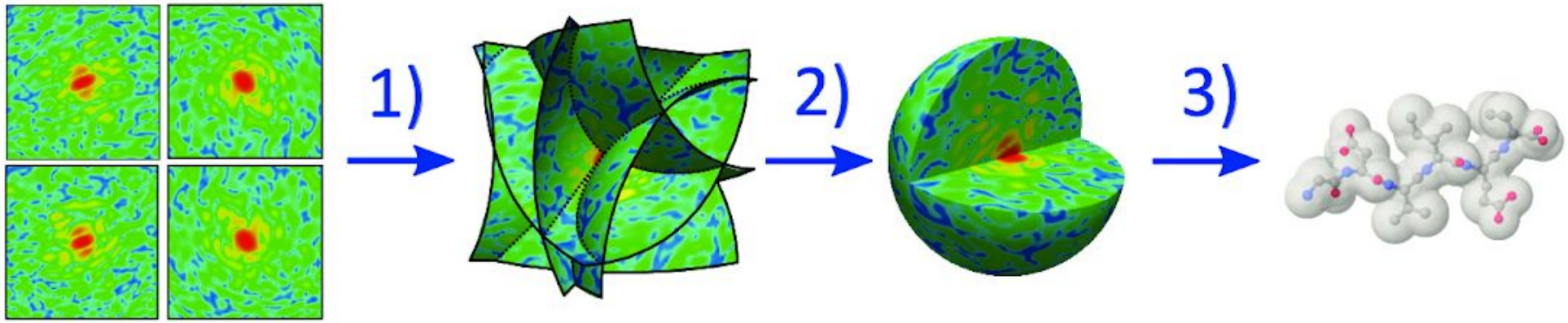
$$J^{(k)}(q,\phi) = I^{(R_k)}(q,\theta(q),\phi),$$

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



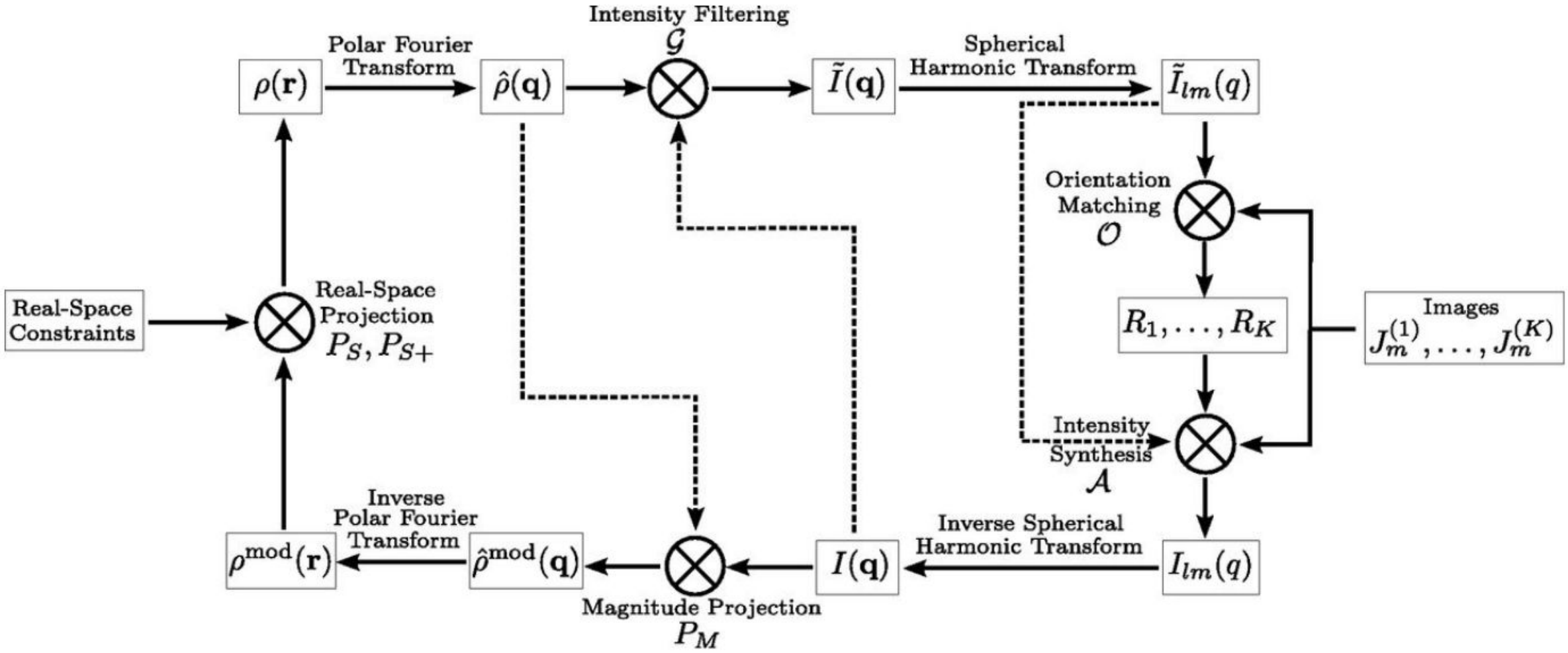
**Challenges:**

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 $\rho$ 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)