Processing data in serial crystallography on-the-fly: what kind of raw data do we want to store?

Alexandra Tolstikova
DESY Photon Science – Scientific Computing

Workshop on “Raw diffraction data reuse: the good, the bad and the challenging”

IUCr 2023 Congress
Serial crystallography

- One exposure per crystal
- Steady stream of crystals
- Each image processed independently
- A lot of images (millions) (5GB/s, 100TB per experiment)
- Data processing after experiment (month, even years)

Figure: Stellato et al., IUCrJ 1 (2014) p204–212
Data processing in SX

- **Experiment**
  - Live feedback
  - Storage
    - HDD: 200 k€ / 4 PB / 5 yrs
    - SSD: As above, \times10
    - Tape: 100€ / TB / 10 yrs

- **"Now"**
  - Data processing

- **" Later "**
  - Result
Data processing in SX – storage costs

HDD: 200k€ / 4 PB / 5 yrs
SSD: as above x10
Tape: 100€ / TB / 10 yrs

2 PB of storage ↑

Raw data from EuXFEL instruments

Cumulative size
Daily production
Moving average (6 months)
Data processing in SX
Benefits of real-time data processing

- Faster results and publication
- No need to store raw data
- Better situational awareness during experiment
- Faster diagnosis of experiment problems
- Less scope for self-delusion
- Energy efficiency – processing data only once
Real-time data processing pipeline at P11

- 133 frames/sec, 16 megapixel, 16 bits/pixel
- all frames processed with below 1s latency

CrystFEL: https://www.desy.de/~twhite/crystfel
OM: https://www.ondamonitor.com
ASAP::O: https://asapo.pages.desy.de/asapo

Project led by Thomas White
Requirements for real-time processing: detector calibration

We must get the data processing right first time.

Yefanov et al., Optics Express 28459 (2015)

Automated real-time geometry calibration is currently under development.
Requirements for real-time processing: computing resources

- A lot of performance improvements in CrystFEL since the start of the project

- 133 frames/sec, 16 megapixel – 32 CPUs on one node (at 40% hit-rate)
What kind of data do we want to store?
What kind of data can we store?

- All raw data
- Calibrated data (converted to photons)
- Only hits
- Only indexed frames
- Unmerged intensities
- Final result – merged intensities
Reasons to store data

- Fraud prevention
- Hope for a better software/analysis methods
- "Unobtainium" sample
Reasons to store data

- Fraud prevention
- Hope for a better software/analysis methods
- "Unobtainium" sample

Storage costs:

- HDD: 200k€ / 4 PB / 5 yrs
- SSD: as above x10
- Tape: 100€ / TB / 10 yrs

Is it really cheaper to store the data, compared to re-running experiment? What if you (as a user) have to pay for it?
Possible compromise solutions

- Store hits only
- Store indexed frames only
- Lossy compression:
  - pixel binning (2x2, 3x3, 4x4...)
  - peaks only?
  - other methods*
- Store data only when it yields result
- Store random sample of the data
- Send data straight to archive (tape)

*Talk by O. Yefanov (Session A118, Wed 23 Aug)
Acknowledgments

DESY:
  FS-SC: Thomas White, Tim Schoof and Anton Barty
  CFEL: Dominik Oberthür, Alessandra Henkel, Bjarne Klopprogge, Julia Maracke, Philipp Middendorf, Ivan de Gennaro Aquino
  P11: Johanna Hakanpää, Helena Taberman, Guillaume Pompidor
  IT: Martin Gasthuber, Juergen Hannappel, Sergey Yakubov

LCLS: Valerio Mariani

Want to try it?
  ZeroMQ data interface: already in CrystFEL 0.10.0
  ASAP::0 interface: in CrystFEL 0.10.2
  See doc/articles/online.rst and doc/articles/speed.rst in CrystFEL directory

Want to hear more?
  Novel Data Methods Workshop tomorrow, Wed 23 Aug @ 14:00
  https://sites.google.com/view/newdatamethods-iucr2023