High Data Rate MX (HDRMX) at NSLS-II

Dale Kreitler, NSLS-II, Brookhaven National Laboratory
IUCr XXV HDRMX Workshop 8/14/2021
Beamline Characteristics

Specifications

<table>
<thead>
<tr>
<th>AMX</th>
<th>FMX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavelength range</td>
<td></td>
</tr>
<tr>
<td>0.7 – 2.5 Å</td>
<td>0.4 – 2.5 Å</td>
</tr>
<tr>
<td>Flux at focus at 12.7 keV</td>
<td>3.5×10^{12} ph/s</td>
</tr>
<tr>
<td>&gt;4×10^{12} ph/s</td>
<td>1.5 × 1 μm²</td>
</tr>
<tr>
<td>Focal spot min (H×V)</td>
<td></td>
</tr>
<tr>
<td>7 × 5 μm²</td>
<td>Eiger 16M (100Hz)</td>
</tr>
<tr>
<td>Detector</td>
<td></td>
</tr>
<tr>
<td>Eiger 9M (200Hz)</td>
<td></td>
</tr>
</tbody>
</table>

Challenges: data storage, transfer, processing and backup.
Requires advanced software and computing cluster.

Raster-based centering

Raster every sample
Vector (helical) data collections
Vector-based rastering and data collection

- Every raster is a ‘vector’
- Raster center everything
- Vector collection by row
  - data.h5 file written for each row
  - Only rectangular rasters (single detector arm)
  - Python LSDC server generates dozor input file based on metadata from database request object
- Dozor processing triggered on remote node (per row, immediately after data acquired)
- dectris-neggia
- Also can reprocess with dials.find_spots
- FileWriter interface

Dozor: G. Bourenkov, A. Popov
LSDC: J. Skinner, J. Aishima

50 Hz; usage typically 50-200 Hz
500 cell raster, ~2.5-10 sec total exposure
Simple raster queueing

Click 4 corners
AMX/FMX Infrastructure

- @ beamlines (2x): 3 WS, 1 data backup WS, 1 OC WS (single processes) and CAs SRVs … (all on 10 Gbs and on GPFS)

- @ NSLS-II central facility: 23 nodes: ~720 cores (56 Gb/s IB to GPFS; 10 Gbs/ node to node)

- @ NSLS-II central facility: 2 x 16 TB SSD fast buffer (NSDs) GPFS

Data transfer to users with globus (https://www.globus.org), no hard drives
Network architecture for raster processing

- **Processing nodes**
  - CPU-004
  - CPU-005
  - CPU-006
  - CPU-026

- **Processing workstation (5GHz)**

- **Infiniband switch**
  - 56Gb/s ingress
  - 56Gb/s egress

- **Processing node:** Intel Xeon Gold 6154 (36 cores, 3GHz)

- **AMX/FMX switch**
  - 40Gb/s ingress
  - 10Gb/s egress

- **Server**
  - Running areaDetector (Eiger Driver)
  - Trigger mode: external series
  - FileWriter Interface

- **Eiger 9M/16M**

- **Eiger DCU**

- **MNDB**
  - CA1

- **Beamline workstations (running GUI clients)**

- **GPFS**
  - 1 PB HD (long term storage)
  - 2x16 TB SSD (.h5, .cbf) RAID1+0

- **Processing trigger (10Gb/s)**

- **Results as Numpy array**

- **10Gb/s ethernet**

- **Typically ~100 Hz @ AMX**

Kreitler: HDRMX at NSLS-II 8/14/21 National Synchrotron Light Source II Office of Energy, Brookhaven National Laboratory
Database statistics of AMX/FMX collections

- ~270k vector collections in 2020
- ~2 rasters/dataset
- 10 sec raster time improvement → ~55 h of annual beamtime
Pipelines: In-line processing with fast_dp, dimple

- Autoprocessing of all standard, vector datasets triggered after collection (XDS/fast_dp)
- Typical dataset ~10 seconds (180 deg, 0.2 deg, 100Hz)
- In automated mode, ~1 unipuck/hour
- Well diffracting samples (ligand binding or dynamics) — ~4 min/dataset including sample exchange
- 2 dedicated 36 core nodes for spotfinding, indexing, integration
- pointles/amless run on overclocked workstation
Data processing timing with fast_dp

- From lysozyme start up data
- 0.2 deg, 150 deg (750 frames)
- Eiger 9M
- 1 Angstrom resolution
- 8 nodes (XDS):
  - Intel Xeon Gold 6154
  - 36 cores, 3 GHz
- AMX workstation (pointless/aimless):
  - Intel Core i7-8700K
  - 6 cores
  - overclocked @ 5 GHz
- Dimple (tetragonal lysozyme; 12-36 seconds)
- If achieving 1000 samples/day need, processing must keep pace (1 dataset/min)
To do list

In-line processing:
indexing/clustering of heat maps
Detection/scoring of multiple lattices for improved heat maps

Off-line processing:
Additional pipelines (autoPROC, DIALS, KAMO)
Model building (phenix, Rosetta)
Considerations for the future

- How to keep pace with throughput without compromising quality?
- Wealth of new information, but avoid analysis paralysis
- Automation vs. manual collection
- New types of experiments enabled by HDMRX: clustering, dynamics, time-resolved, etc.
- How do we use these new tools?
- Precision of sample stages/goniometers, stable optics required by faster detectors
Acknowledgments

- John Lara
- Jean Jakonic
- Dieter Schneider
- Stu Myers
- Tom Langdon
- Shirish Chodankar
- Bruno Seiva Martins
- Vivian Stojanoff
- John Skinner
- Hugo Slepicka
- Kun Qian
- Sean McSweeney
- Martin Fuchs
- Babak Andi
- Dale Kreitler

This work is supported by the US National Institutes of Health and the US Department of Energy

Software:

- Dozor -- Gleb Bourenkov, Alexander Popov
- fast_dp
- developers/maintainers Graeme Winter, Markus Gerstel, Richard Gildea, Nicholas Devenish, and others