Report of the
High Data-Rate Macromolecular Crystallography Meeting
Brookhaven National Laboratory 26-28 May 2016
Report Date: 7 June 2016

This is the first of a series of three meetings in spring and summer 2016 on changes needed to existing major software packages for support of very high data rate macromolecular crystallography. The first meeting was held at Brookhaven National Laboratory, 26 – 28 May 2016, and was organized by Herbert J. Bernstein of Rochester Institute of Technology, Nicholas K. Sauter of Lawrence Berkeley National Laboratory and Robert M. Sweet of Brookhaven National Laboratory.

The report is the collaborative result of the work of the participants in the meetings, many of whom have approved the wording in the earlier drafts. The editor of the text is HJB (yayahjb **at** gmail **dot** com) to whom comments and corrections should be directed.

The meeting had several sponsors: Funding from Dectris, Ltd of Baden Switzerland to Rochester Institute of Technology, from the National Institute of General Medical Sciences of the National Institutes of Health under grant 3R01GM117126-01S1 to Lawrence Berkeley National Laboratory, from the Department of Energy Offices of Biological and Environmental Research and of Basic Energy Sciences grants DE-AC02-98CH10886 and E-SC0012704, and from NIH grants P41RR012408, P41GM103473, and P41GM111244 to Brookhaven National Laboratory. The opinions expressed in this report are those of the meeting participants and not necessarily those of the funding sources.

The attendees at the meeting were:

Name	Institution
On-site Participants	
Mark Hilgart	APS Argonne National Laboratory
Jun Aishima	Australian Synchrotron
Tom Caradoc-Davies	Australian Synchrotron
Kaden Badalian	Binghamton University
Frances C. Bernstein	Brookhaven National Laboratory (ret.)
Andreas Förster	DECTRIS Ltd.
Markus Mathes	DECTRIS Ltd.
Eugen Wintersberger	Deutsches Elektronen-Synchrotron
David Hall	Diamond Light Source
Graeme Winter	Diamond Light Source
Andrew Hammersley	European Synchrotron Radiation Facility
Gerard Bricogne	Global Phasing Ltd.
Clemens Vonrhein	Global Phasing Ltd.
Aaron Brewster	Lawrence Berkeley National Laboratory
Nicholas K. Sauter	Lawrence Berkeley National Laboratory
Jie Nan	MAX IV Lund University
Harry Powell	MRC Laboratory of Molecular Biology (ret.)
Matt Cowan	NSLS-II Brookhaven National Laboratory
Martin Fuchs	NSLS-II Brookhaven National Laboratory
Jean Jakoncic	NSLS-II Brookhaven National Laboratory
Robert Petkus	NSLS-II Brookhaven National Laboratory
Alexei Soares	NSLS-II Brookhaven National Laboratory
Dieter Schneider	NSLS-II Brookhaven National Laboratory
John Skinner	NSLS-II Brookhaven National Laboratory
Bob Sweet	NSLS-II Brookhaven National Laboratory
Kerstin Kleese van Dam	CSI Brookhaven National Laboratory
Xiaochun Yang	NY Structural Biology Consortium
Seetharaman Jayaraman	NY Structural Biology Consortium
Herbert J. Bernstein	Rochester Institute of Technology
Simon Ebner	SLS Paul Scherrer Institut
Ezequiel Panepucci	SLS Paul Scherrer Institut
Justyna Aleksandra Wojdyla	SLS Paul Scherrer Institut
Martin Savko	SOLEIL Synchrotron
Elena Pourmal	The HDF Group
James Holton	UCSF/LBNL/SLAC
Wladek Minor	University of Virginia

Electronic Participants

Nukri Sanishvili	APS Argonne National Laboratory, GMCA-CAT
Kevin Battaile	APS Argonne National Laboratory, IMCA-CAT
Joe Digilio	APS Argonne National Laboratory, IMCA-CAT
Erica Dugrid	APS Argonne National Laboratory, IMCA-CAT
Spencer Anderson	APS Argonne National Laboratory, LS-CAT
Joe Brunzelle	APS Argonne National Laboratory, LS-CAT
Keith Brister	APS Argonne National Laboratory, LS-CAT
Surajit Banerjee	APS Argonne National Laboratory, NE-CAT
David Neau	APS Argonne National Laboratory, NE-CAT
Frank Murphy	APS Argonne National Laboratory, NE-CAT
K. Rajasankar	APS Argonne National Laboratory, NE-CAT
Jon Schuermann	APS Argonne National Laboratory, NE-CAT
James P. Withrow	APS Argonne National Laboratory, NE-CAT
Steve Ginell	APS Argonne National Laboratory, SBC CAT
Chris Lazarski	APS Argonne National Laboratory, SBC-CAT
John Chrzas	APS Argonne National Laboratory, SER-CAT
Albert Fu	APS Argonne National Laboratory, SER-CAT
Zhongmin Jin	APS Argonne National Laboratory, SER-CAT
Daniel Eriksson	Australian Synchrotron
Vesna Samardzic-Boban	Australian Synchrotron
Stefan Brandstetter	DECTRIS Ltd.
Gleb Bourenkov	EMBL
Alexander Popov	European Synchrotron Radiation Facility
Go Ueno	SPring-8
Kazuya Hasegawa	SPring-8
Keitaro Yamashita	SPring-8
Thomas Eriksson	SSRL SLAC National Accelerator Laboratory
Takanori Nakane	The University of Tokyo
Kay Diederichs	Universität Konstanz

35 participants attended on-site the first day, 31 the second day, and 16 the third day for report draft editing.

28 electronic participants attended the presentations on the first day. On the second day fewer electronic participants attended; eight connected and at least two were active participants in the discussion. There were no electronic participants for the report editing session on the third day.

The first day was primarily devoted to the presentations shown on the meeting web site: http://medsbio.org/meetings/BNL_May16_HDRMX_Meeting.html

Statement of the Problem and Charge to the Meeting

Macromolecular crystallography (MX) is the gold standard for the determination of the atomic-resolution three-dimensional structure of large biologically active molecules. MX is becoming a big data science straining the capabilities of computers and networks. New techniques of serial crystallography are allowing new science to be done but they are increasing the heterogeneity of the data that must be handled.

There are issues in data handling:

- We are dealing with much more data than in the past.
- In the short term we need to store a lot of data that is retrievable quickly.
- In the medium term we need to store some version of much of the same data for processing and for users to take home.
- In the longer term we may need to store the publishable data.
- We need to consider issues of compression and background removal.

Eiger 16M detectors produce 2.4 gigapixels per second, 76 raw gigabits per second. This is often more than 10 Gb/s networks can handle, even when compressed 4:1. We face an Increasingly daunting flood of image data We should try to reduce movement of data, reduce transformations of data, and move data in large blocks. Compressions could be improved, but that will not be sufficient. No single compression is ideal. No single compression is sufficient.

Why be concerned?

- For any stochastic system, the delays and lengths of queues rise sharply as the rate of arrival of information to process approaches the rate at which it can be processed.
- For any information processing system, the rate at which you can move information through the system is limited by the capacity of the narrowest bottleneck.
- When you work close to the capacity of a system you are dancing on the edge of a cliff.

This meeting was charged with finding answers to the following questions:

- What stumbling blocks inhibit direct processing of HDF5 data?
- Is there a way to have the data produced by the detector processed by all the major packages without conversion?
- What are best practices to process Eiger images?
    - Using C, Fortran, Python?
    - In large compute clusters at synchrotrons?

- In users' home lab computers?

Discussion Points

The second day started with parallel software and beamlines/controls breakout sessions, which were recombined to discuss joint issues. The combined result was a great deal of agreement, initially as follows:

It was agreed that we will set up, as a community resource, an HDRMX web site that provides pointers and useful information on open-source software for high data-rate MX as a one-stop shopping page (for details, see consensus recommendation #5 below).

The continued discussion in the joint meeting produced the following preliminary best practices recommendations:

Spot finding. For screening purposes, at present allocating one image per process is most effective, and keeping up with an Eiger 16M at full rate requires approximately ten very competent nodes with normal cores. GPUs are not at present appropriate. J. Holton, J. Jakoncic and G. Winter will carefully consider the evidence, with input from the LBNL group, and make a firm best practices recommendation.

Metadata. It is agreed that what is needed is a way to simply and reliably integrate the full equivalent to the CBF metadata into master files. People need to be made aware of the NXmx definitions that were jointly defined by IUCr COMCIFS and NIAC for exactly this purpose. Easier to follow information will be added to the web site by HJB in consultation with H. Powell, E. Wintersberger and other interested people.

The beamlines/controls group noted that Dectris has agreed to work on optimizing a parallel file writer and streamer on the DCU using the 2 x 10 Gb links. There were also requests to possibly do that on a single 40 Gb link. Dectris will not guarantee failsafe performance using the two in parallel.

After further discussion, the discussion points were refined to produce the consensus recommendations supported by almost all participants at the meeting. The following recommendations are appropriate to those for whom speed and efficiency in MX data collection are of great significance.

DIALS Workshops

There was a separate discussion focused on the issues of dissemination of DIALS and best use of DIALS when working with Eigers. The principal request was for DIALS developers to organize workshops in Europe and the US, perhaps also in Japan / China / Australia, to help users to learn the best use of DIALS. Also there was interest in local / smaller facility-based presentations, perhaps for one day, for a greater number of local users.

Future HDRMX Meetings

The attendees were reminded of the currently planned HDRMX meetings in association with the ACA meeting in Denver on 23 July 2016 and as a satellite meeting to the ECM meeting in Basel on 2 September 2016. There was particularly strong interest in attendance at the ACA session.

Consensus recommendations of the meeting

The meeting notes that all major applications (DIALS, HKL, MOSFLM, XDS) have now worked out ways to read Eiger data, and most (DIALS, HKL, XDS) have ways to read it directly from HDF5-formatted files, but improvements are needed in the support documentation and software tools for creating appropriate HDF5 master files and also in the timing.

1. The meeting notes that Python wrappers have become very important in the development of MX workflow pipelines. There is concern that the use of Python rather than C, C++ or Fortran might reduce efficiency by introducing additional data copying. Comments at the meeting were made about the good handling of the data copying issue by numpy. The DIALS project has volunteered to profile use of h5py to make sure it is as efficient as possible and to check that motion is indeed being minimized as has been suggested is the case with numpy arrays.

2. The meeting notes that the DECTRIS/XDS plugin now available, as discussed in Markus Mathes' talk, appears likely to help improve timing for reading of HDF5 images directly in a wide range of applications. The meeting recommends that application developers should try the DECTRIS/XDS plugin in their applications.

3. The meeting notes that an effort is necessary in an increasing number of cases to provide full CBF-based metadata in HDF5 master files. HJB and CV have volunteered to gather and curate the data from all beamlines willing to contribute. The meeting respectfully asks that beamline scientists, please,

for the love of our science, provide full beamline metadata information for posting on the HDRMX web site

4. The meeting implores all concerned to work toward full, unconditional NeXus compliance.

5. The meeting notes that there is a critical need to improve dissemination of both software and best practices and recommends that the HDRMX website be established to provide one-stop shopping for the community to get the open-source resources they need to do software development for processing Eiger data. The meeting notes with gratitude that the necessary permissions have been granted by the owners of the intellectual property in this list, that Dectris has agreed to allow use and extension of relevant portions of their documentation and that Global Phasing has agreed to work towards adding hdf2mini-cbf to the site. On that site we will include

    - an extended version of the Dectris documentation of the Eiger master file / data file structure, especially for multi-axis metadata, including programmer's reference material explaining clearly the relationships among the NXmx-based NeXus/HDF5 format, the imgCIF/CBF format, and the coordinate systems

    - links to the NeXus format documentation (including documentation regarding the NeXus NXmx application definition) and reference copies of the software and guidance to the portions relevant to MX

    - links to the imgCIF/CBF documentation and reference copies of the software (CBFlib) and guidance on the portions relevant to MX

    - links to the HDF5 documentation and reference copies of the software (HDF group version) and guidance on the portions relevant to MX

    - links to LZ4 compression documentation and software (NIAC version)[1]

    - links to BitShuffle compression documentation and software (NIAC version)[1]

[1] Because these compression filters are heavily used, there are many copies held at many different sites with minor configuration differences that have caused difficulties in integration with various packages. Settling on the NIAC version as the reference copy for this community will hopefully allow for smoother integration.

- links to eiger2cbf documentation and reference copies of the software and guidance on the portions relevant to MX

- the dectris-xds-plugin and plugin API that permits direct reading of HDF5 images from applications and provides a framework to help insulate application design from image format issues

- useful scripts (starting with XDS fork scripts)

- guidance for and examples of adding beamline and experiment metadata into an existing Nexus/HDF5 master file as written by Dectris, including both CBF and HDF5 metadata templates, and the necessary software tools

- guidance for and examples of writing and adding beamline and experiment metadata into new master file

- guidance for simulating the output of the Eiger streaming interface starting from existing HDF5 or CBF files (to be eventually followed by full software implementations)

- tools to calibrate and verify beamline metadata, with links to beamline metadata examples from all beamlines willing to contribute

- a repository of example Eiger datasets from different synchrotrons, including example raster scans in HDF5 format preferably with micro and macro cases. Inasmuch as most currently available raster scans are in CBF format, GW will provide CBFs and HJB will convert those to HDF5. EHP will provide some native Eiger raster scans.

HJB will be secretary for the website assisted by CV. As per WM, storage of up to 200TB for data will be provided by Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC, http://proteindiffraction.org). As additional storage sites for data are volunteered, access will be coherently integrated with access to the IRRMC storage.

6. The meeting accept GW ‘s generous offer to work with Pilatus users to try to increase use of full CBF headers.

7. The meeting endorses an effort to create a reference implementation of DIALS spot finder running as a script, not as a server.

8. Inasmuch as the metadata issues discussed above will be providing metadata from multiple sources, Dectris has agreed to endeavor to provide the data that appears in miniCBF headers. The coordinate system is the one defined in the NeXus standard and used in the NeXus NXmx application definition. A link to the default NeXus geometry definitions will be provided on the HDRMX web site. The beamline scientists continue to be responsible for additional geometry data such as complex nesting of axes and non-NeXus-default rotation directions or any vector/axis definitions not covered by the standard.

9. The meeting recommends that for use cases beyond the current Dectris assumptions a working group of this meeting will prepare descriptions of those use cases, forward them to NIAC for appropriate action and the template merge and edit capabilities will be extended to allow beamlines to provide the necessary new master files.

10. The meeting notes that for accepting and processing this data in a timely manner, high bandwidth networks (more than 10 Gb/s) and 10 or more substantial processing nodes are likely to be needed.

11. GW (chair), AB, NKS, WM, TCD, JMH, MS, JN, JJ, MCH, JAW are forming a benchmark committee that will define standard benchmarks, run them and forward results to HJB for the web site. This committee will also investigate the issues surrounding compression of the x-ray data and gather evidence on effectiveness of the various schemes and provide useful examples for the web site. HJB will assist in necessary format conversions and re-bricking data files.

12. The meeting respectfully asks Dectris to investigate if it would possible to have the option of a 40Gb/s interface from the DCU.

We are pleased to note the following useful information provided by Takanori Nakane: “Keitaro Yamashita has adapted Cheetah's spot finding routine (mostly used for SFX at XFEL) to receive frames from EIGER ZeroMQ interface; https://github.com/keitaroyam/cheetah/tree/eiger-zmq/eiger-zmq It is used at SPring-8 BL32XU with EIGER 9M.”

Conclusions

This was a particularly collegial, collaborative and effective meeting that achieved its goals and laid the groundwork for future collaborative efforts that should help to improve the efficiency and effectiveness of high data-rate macromolecular crystallography. Having met in person makes it more likely that people will continue to collaborate efficiently in the future.