High Data-Rate Macromolecular Crystallography
MAX IV Meeting, 15 – 17 March 2017
https://indico.maxiv.lu.se/event/233/

Data Handling II

Herbert J. Bernstein
Rochester Institute of Technology

Work supported in part by Dectris, Ltd.
Meeting supported in part by Dectris and by Max IV

HDRMX

MAX IV

Introduction

The new regime of very high data rates makes the availability of appropriate and efficient data manipulations in beamline infrastructure a much more important issue than in the past. Different experimental protocols, applications and pipelines may require different services. Loads in network, storage, and computational resources need to be carefully considered to maximize return on investment.

- Packaging data and metadata for use at the beamline
- Packaging data and metadata for export to home institutions
- Selection of appropriate subsets of images
- Summing images
- Binning images
- Regions of interest
- Format conversion
- Compressions

HDRMX

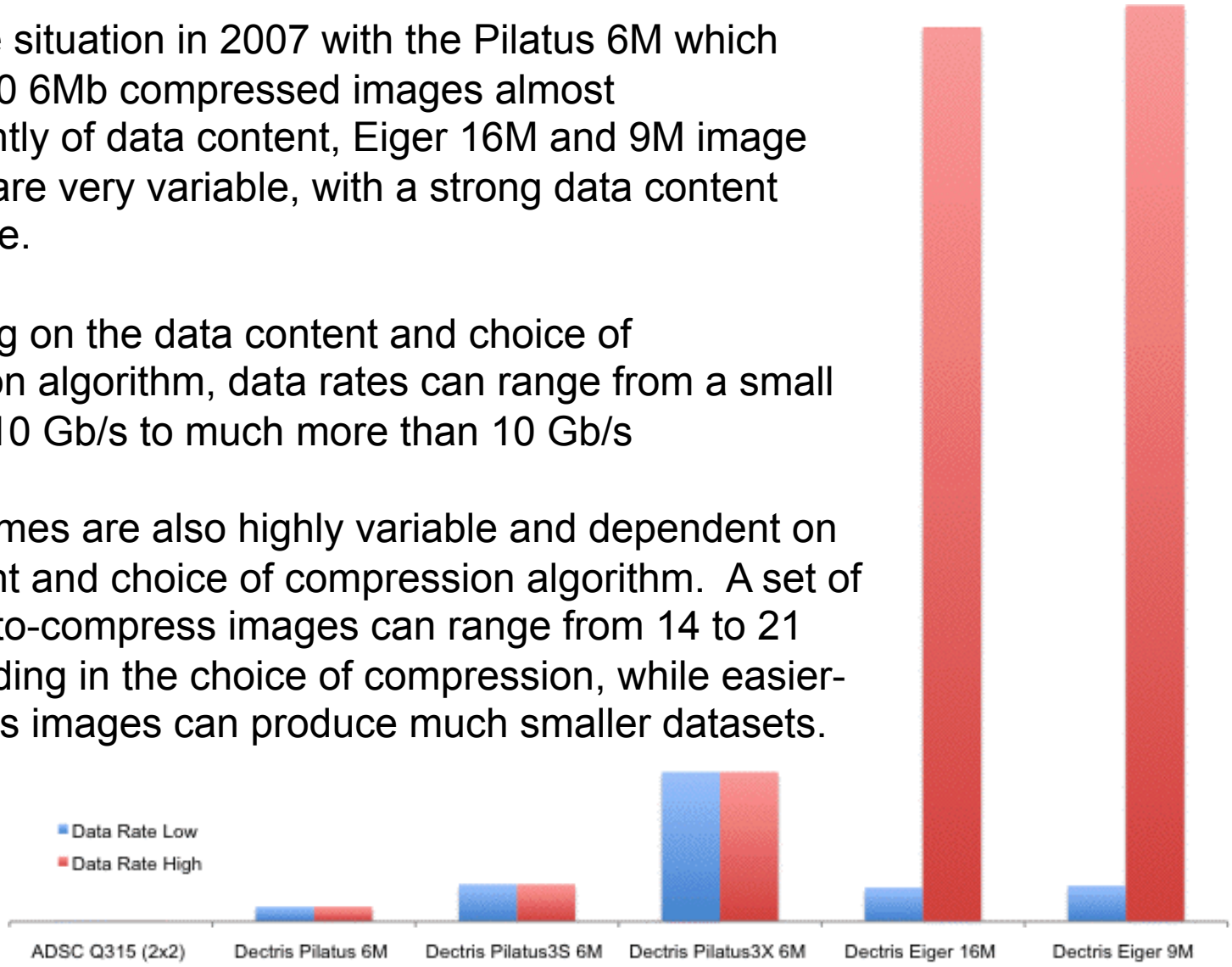
MAX IV

How High are The Data Rates?

• Unlike the situation in 2007 with the Pilatus 6M which delivered 10 6Mb compressed images almost independently of data content, Eiger 16M and 9M image data rates are very variable, with a strong data content dependence.

• Depending on the data content and choice of compression algorithm, data rates can range from a small fraction of 10 Gb/s to much more than 10 Gb/s

• Data volumes are also highly variable and dependent on data content and choice of compression algorithm. A set of 3600 hard-to-compress images can range from 14 to 21 GB, depending in the choice of compression, while easier-to-compress images can produce much smaller datasets.



HDRMX

MAX IV

What We Can Do

- To conserve resources, we need to try to keep the data we actually need to do our experiments, but each change we make has its risks.
- Every time we reformat data to a different compression or different format we consume resources (time, cpus, network bandwidth, disk space) to do the conversion.
- If we sum or bin images, we risk loss of detail in spot profiles.
- If we use ROIs, we risk loss of spots.
- But, if we do nothing we risk running out of network bandwidth and disk space sooner, having to drop images sooner than if we had made them smaller, and we risk falling behind in the race against radiation damage.
- No answer is perfect or universal, but a structure that allows beamlines to choose which tradeoffs work for their users would be helpful.

HDRMX

MAX IV

Good News on Compression

- Facebook provides **Zstandard**, open source, with a patent license

https://github.com/facebook/zstd

- **Blosc** packages Zstandard with bitshuffle as an HDF5 plugin
- **The resulting bszstd compresses much better than bslz4 at a modest additional cpu cost**. Here are recent results for Lysozymes on an Eiger 9M at AMX at NSLS-II:

Set	BSLZ4 CR	BSZSTD 1 CR	BSZSTD 2 CR	Space lev 1 savings	Time lev 1 cost	Space lev 8 savings	Time lev 8 cost
100Hz_494	5.6	6.9	7.1	**18%**	**1%**	21%	25%
100Hz_B_497	5.2	6.1	6.3	**16%**	**6%**	18%	25%
100Hz_C_501	3.8	4.1	4.2	**7%**	**1%**	9%	27%
200Hz_493	6.8	9.2	9.8	**27%**	**1%**	31%	23%
200Hz_B_495	6.3	8.1	8.5	**23%**	**1%**	27%	24%
200Hz_C_499	4.5	5.1	5.2	**11%**	**1%**	13%	27%

HDRMX

MAX IV