
High Data-Rate Macromolecular Crystallography
NSLS-II Brookhaven National Laboratory
26 – 28 May 2016

# Statement of the Problem and Charge to the Meeting

Herbert J. Bernstein
Rochester Institute of Technology

Work supported in part by Dectris, Ltd.
Meeting supported in part by Dectris, LSBR, NIGMS, DOE

# Statement of the Problem

- 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.

# Statement of the Problem (cont.)

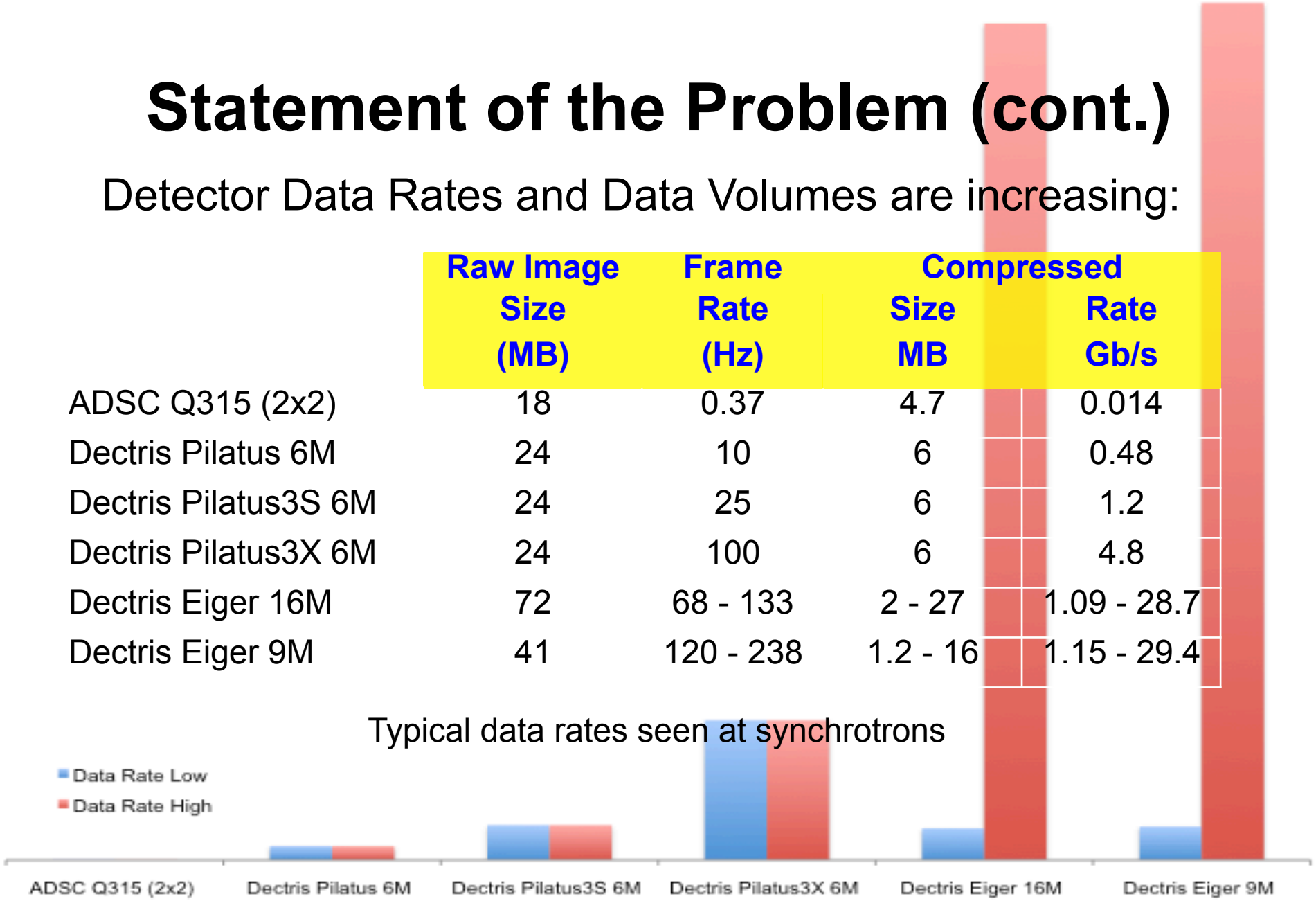
- Issues in Archiving
  - We are dealing with much more data than in the past.
  - In the short term we need to store a lot of data that is retrievable quickly.
  - In the medium term we need to store some version of much of the same data for processing and for users to take home.
  - In the longer term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

# Statement of the Problem (cont.)

Detector Data Rates and Data Volumes are increasing:

| | Raw Image Size (MB) | Frame Rate (Hz) | Compressed Size MB | Compressed Rate Gb/s |
|---|---|---|---|---|
| ADSC Q315 (2x2) | 18 | 0.37 | 4.7 | 0.014 |
| Dectris Pilatus 6M | 24 | 10 | 6 | 0.48 |
| Dectris Pilatus3S 6M | 24 | 25 | 6 | 1.2 |
| Dectris Pilatus3X 6M | 24 | 100 | 6 | 4.8 |
| Dectris Eiger 16M | 72 | 68 - 133 | 2 - 27 | 1.09 - 28.7 |
| Dectris Eiger 9M | 41 | 120 - 238 | 1.2 - 16 | 1.15 - 29.4 |

Typical data rates seen at synchrotrons



# Statement of the Problem (cont.)

- Eiger 16M detectors produce
  - 2.4 gigapixels per second
  - 76 raw gigabits per second
  - often more than 10 Gb/s networks can handle
  - even when compressed 4:1
- Increasingly daunting flood of image data
- We should try to
  - Reduce movement of data
  - Reduce transformations of data
  - Move data in large blocks

# Statement of the Problem (cont.)

- Compressions could be improved, but that will not be sufficient

| | bzip2 | bslz4 | nibble-offset | Byte-offset | packed | Lz4**2 |
|---|---|---|---|---|---|---|
| Avg Ratio | 5 | 4 | 4 | 3 | 3 | 2 |
| Max Ratio | 37059 | 114 | 16 | 4 | 647 | 333 |
| Min Ratio | 3 | 3 | 3 | 1 | 2 | 2 |
| Avg fps | 11 | 56 | 33 | 53 | 26 | 20 |

Compression ratios in the presence of varying amounts of water

- 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 bottle neck.
- **When you work close to the capacity of a system you are dancing on the edge of a cliff.**

# Charge to the Meeting

We hope for 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?