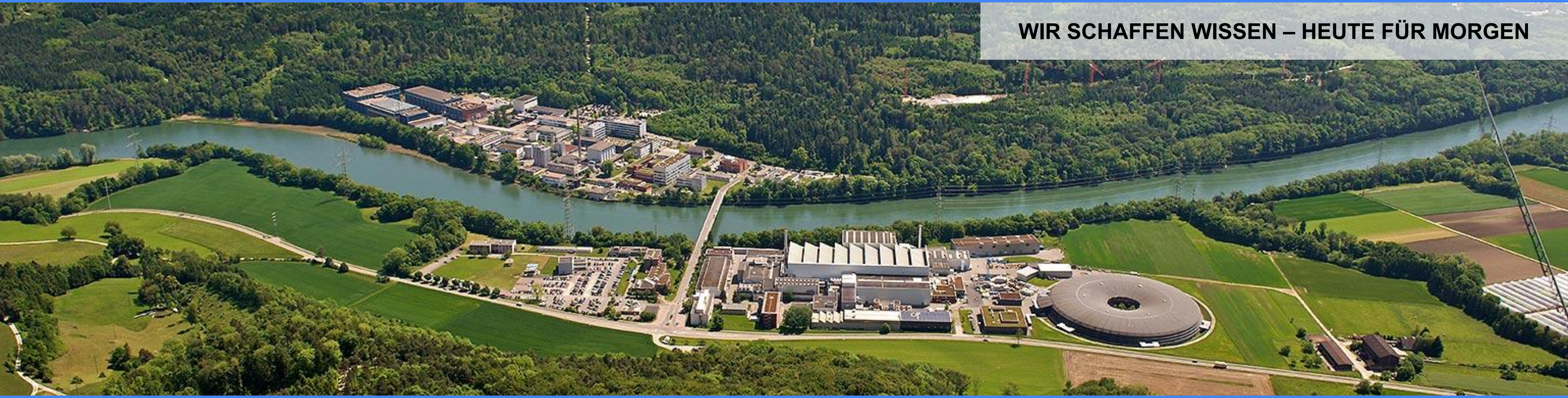
PAUL SCHERRER INSTITUT


WIR SCHAFFEN WISSEN – HEUTE FÜR MORGEN

HDRMX & Eiger

Efficient Handling of Large and Small (Detector) Data at the Paul Scherrer Institute

S. Ebner, J.A. Wojdyla & E. Panepucci

Paul Scherrer Institut (PSI), CH-5232 Villigen PSI, Switzerland

summary

EigerX 16M at the SLS

- most users still using 0.1°, 0.1s per frame – 180-360°
- auto processing via in house pipelines via ADP
- users do not complain about data volume after bs-lz4 compression, before with lz4…
- most users not wowed by it

Data retrieval options

- GlobusOnline:
  - www.globus.org
  - Hardly used by MX users
  - Proprietary customers need to pay
- rsync + ssh
  - Usage increasing
- External hard drive
  - Most used method

PAUL SCHERRER INSTITUT PSI

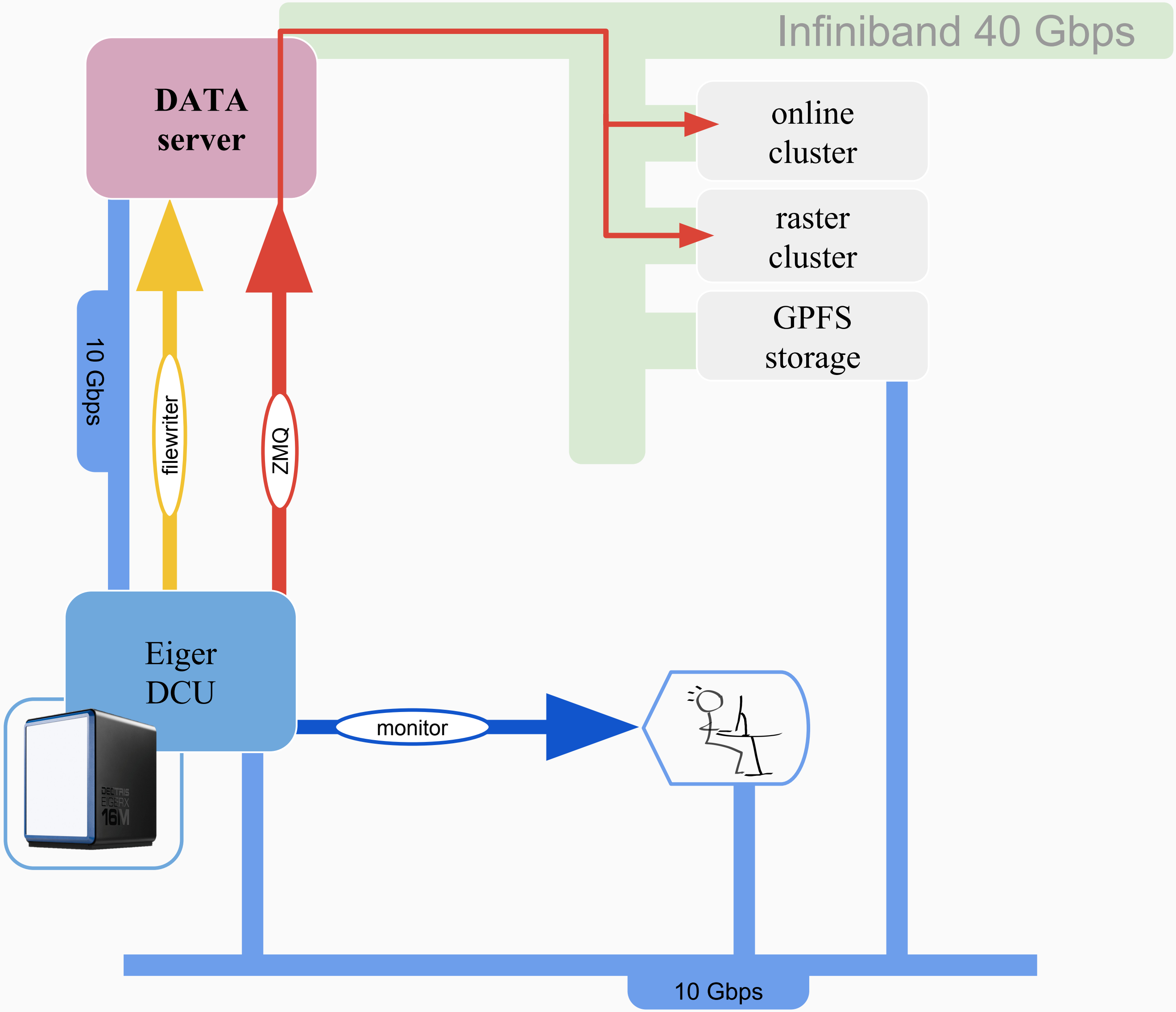
Hardware Infrastructure

- Computing
  - Online-Cluster: 4 nodes: Dual Xeon E5-2697v2 (2.70 GHz), 24 cores, 256GB ram, Scientific Linux 6.4
    - Data reduction
    - Spot finding (raster)
  - Raster-Cluster: 3 nodes: Dual Xeon E5-2697v2, 24 cores, 256GB ram, Scientific Linux 6.4
    - Spot finding (raster)
  - Offline-Cluster: 16 nodes: Dual Xeon E5-2690v3 (2.60 GHz), 256GB ram, Scientific Linux 7.0
    - MX software
    - graphics available via nomachine

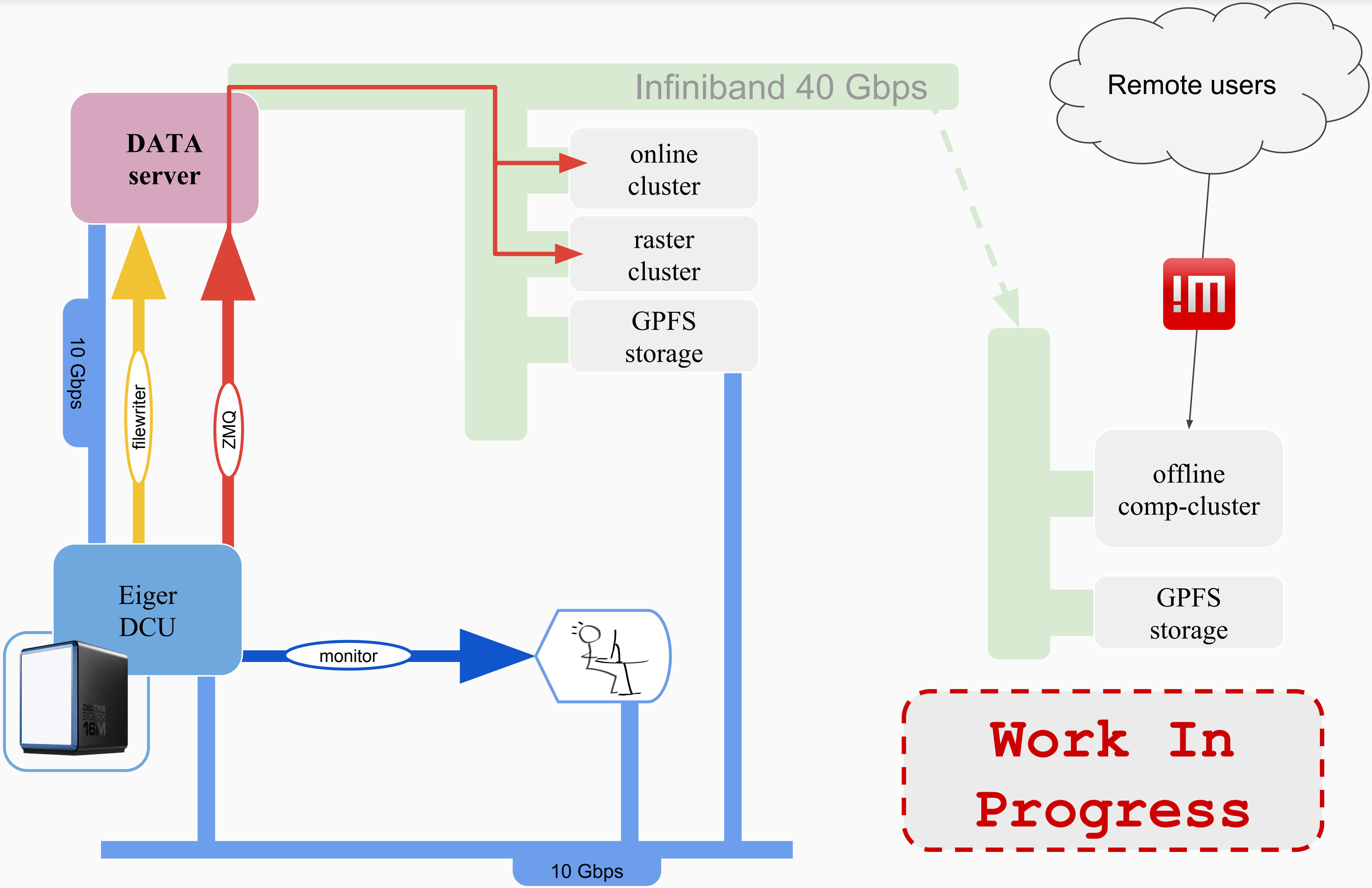
- Storage
  - IBM GPFS version 4.1
  - 1.2 PB Total
  - 175 TB for all MX beamlines

- User console on 10 Gbps
  - improves 16m loading and display for user inspection

Hardware infrastructure – layout now

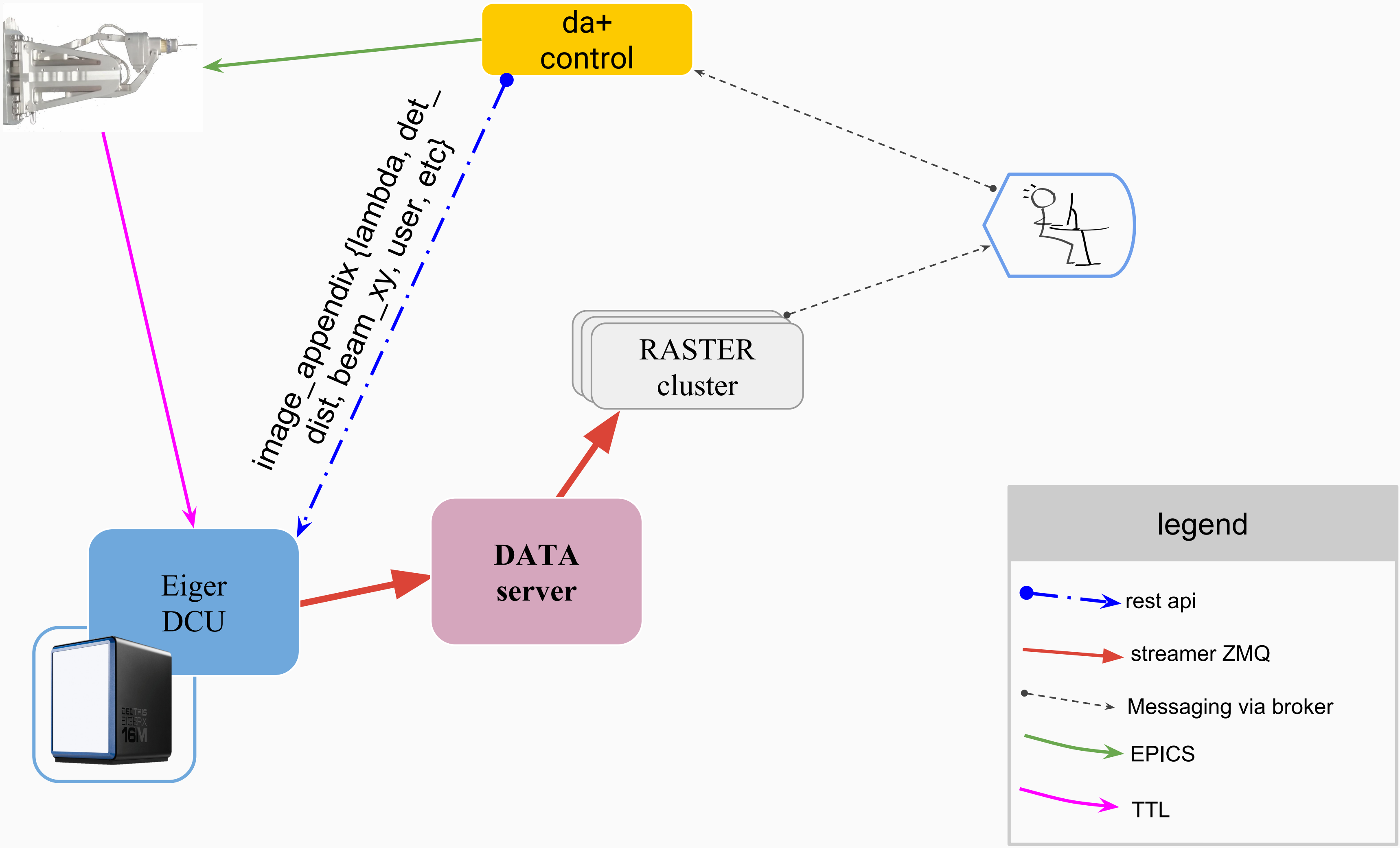


Hardware infrastructure – layout future

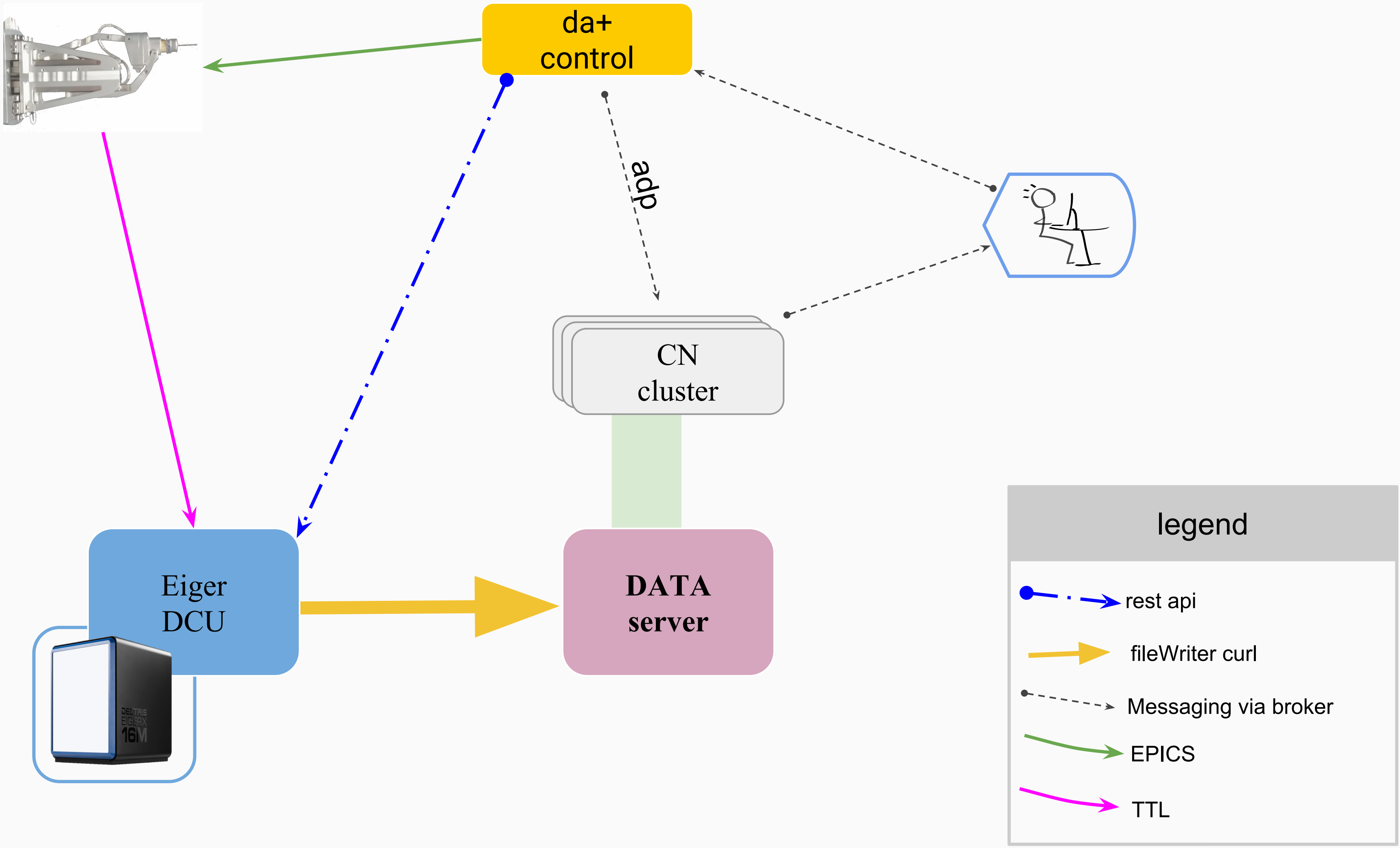


Work In Progress

Data Pathways – stream



Data Pathways – fileWriter



fileWriter vs streamer

- fileWriter alone
  - must rework master
  - no online analysis possible
  - most robust interface so far
- streamer alone
  - online analysis possible
  - need to assemble h5/NXmx
  - FW 1.5.2 not very robust (FW 1.6.2 improves it? to be tested)
- both fileWriter + streamer
  - double bandwidth
  - not officially supported at high rates (still?)
  - limits highest rate for rastering in our case

Grid scanning (rastering) at our site must use both.

PAUL SCHERRER INSTITUT

Automatic data processing

- Grid scan
  - file writer *h5 images displayed in Albula
  - streamed bslz4 data analyzed with spotfinder
- Strategy
  - file writer *h5 images
  - conversion to cbf with eiger2cbf -> indexing & strategy with mosflm
- Dataset
  - file writer *h5 images
  - fast_xds (initialized before full dataset is collected)
  - goeiger.com inhouse processing pipeline (after full dataset is available gpfs)

PAUL SCHERRER INSTITUT PSI

Automatic data processing

- Example for dataset with angular range of 180°
  - Fast_xds:
    - Run 1 with 30° of data JOBS=XYCORR INIT
    - Run 2 with 60° of data JOBS=COLSPOT IDXREF
    - Run 3 with 120° of data JOBS=DEFPIX INTEGRATE CORRECT
  - Goeiger.com (default XDSP1 option):
    - XDS processing of 180° of data in space group P1
    - POINTLESS to determine correct space group
    - Rerun CORRECT step in new space group (and INTEGRATE if necessary)
    - XDSCONV to prepare mtz file(s)
- we have no online raddam monitoring via spot finding must be reliable so users won’t abort data collections thinking their crystal is dead

PAUL SCHERRER INSTITUT PSI

Data Reduction XDS vs DIALS

- Lysozyme dataset 900 images @ 0.1°
  - XDS processing without H5ToXds.script

XDS	h5	cbf	h5	cbf
	4 nodes	4 nodes	1 node	1 node

	JOBS=8
	PROCESSORS=12
		JOBS=8
		PROCESSORS=12
			JOBS=4
			PROCESSORS=6
				JOBS=4
				PROCESSORS=6

XYCORR	1.3	1.3	1.3	1.3
INIT	18.1	12.2	18.1	13.0
COLSPOT	12.3	9.9	42.8	32.0
IDXREF	2.2	2.3	2.0	2.0
DEFPIX	1.5	1.5	1.5	1.4
INTEGRATE	29.7	20.0	87.5	64.7
CORRECT	7.2	7.2	7.7	7.4
TOTAL	**76.6**	**55.3**	**163.3**	**122.6**

DIALS	h5
	1 node
	24 CPUs
import	10.0
find_spots	60.2
index	116
refine	54.1
integrate	142.2
export	4
TOTAL	**386.5**

PAUL SCHERRER INSTITUT PSI

Problems controlling

In situ serial crystallography

- user selects 20-40 xtals
- one arm for all xtals or one per xtal
    - **one**, otherwise **too much arm-time overhead** (2.5 s per arm command)
- typical: 20 xtals selected – 10° total each xtal – 0.1° 0.1s per frame
    - ntrigger=20 nimages=100
    - nimages_per_file=100
    - one trigger per _data_*.h5
    - one master to **confuse** them all: omega in master means nothing
- using filewriter?
    - need to rework master file **before** delivering to user’s folder
- using stream?
    - need to write hdf5 from scratch

other protocols

SAD with inverse beam and small wedges

- ntrigger = number of wedges, both inverse and direct
- nimages = number of frames per wedge
  - can nexus NXmx handle this?
- need to simplify: have to sort the data files and create master file for the direct dataset and one for the inverse

Actually, anything other than a single continuous sweep over a single crystal will either need to be reworked if using fileWriter or assembled from the streamed images.

Both fileWriter and streamer work but with the streamer we have the possibility to have a peek at the diffraction before it ever hits an I/O bottleneck.

Other issues

https://github.com/kiyo-masui/bitshuffle

- OpenMP compilation results in processes that deadlock if running too many in single node
    - decompress time goes from around 50ms to minutes

Our Eiger DCU could not be properly configured to use both 10 Gbps for data.

Eiger webmin could be improved and more control given to the sites.

Hoping for a more robust streamer interface in FW 1.6.2 to be tested next shutdown.

Acknowledgements

Justyna Wojdyla – automatic data processing

Simon Ebner – Streaming concept, implementation

Dectris – for this very nice detector and how quickly it addressed our urgent issues

Leonardo Sala – for the data retrieval setup

Heiner Billich – the hardware infrastructure

MX Team – for spending nights during test shifts trying to understand how to cope with this detector

Our **very** patient users who were willing to suffer LZ4 compressed datasets in the beginning.

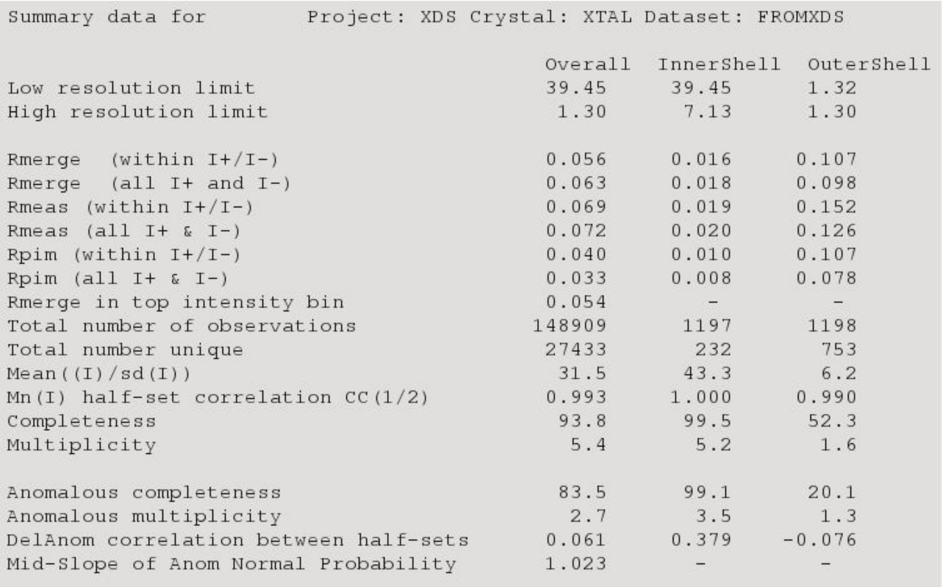
Data Reduction XDS vs DIALS

- Processing with DIALS

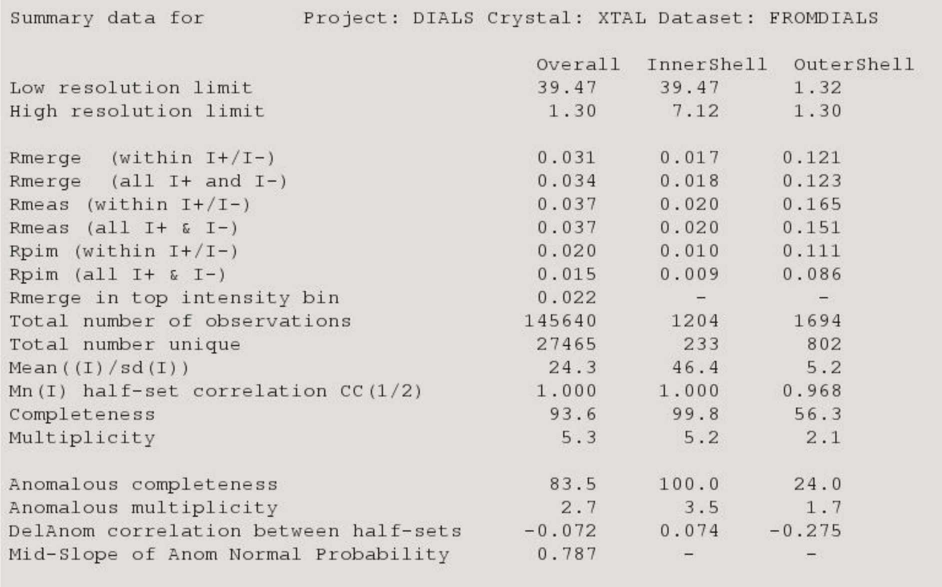
```
dials.import /sls/X06SA/data/e10003/Data10/20160408/testshot/testshot_5_master.h5

dials.find_spots datablock.json spotfinder.filter.min_spot_size=3 spotfinder.mp.
nproc=24 spotfinder.filter.d_min=1.3

dials.index datablock.json strong.pickle indexing.nproc=24 refinement.mp.nproc=24
unit_cell=78.93,78.93,36.94,90,90,90 space_group=P422 d_min=1.3

dials.refine indexed.pickle experiments.json nproc=24 scan_varying=True

dials.integrate refined_experiments.json refined.pickle integration.mp.nproc=24
prediction.d_min=1.3
```

Data Reduction XDS vs DIALS

- Lysozyme dataset 900 images @ 0.1°
  - XDS processing: XDS_ASCII.HKL -> AIMLESS
  - DIALS processing: dials.export -> POINTLESS -> AIMLESS

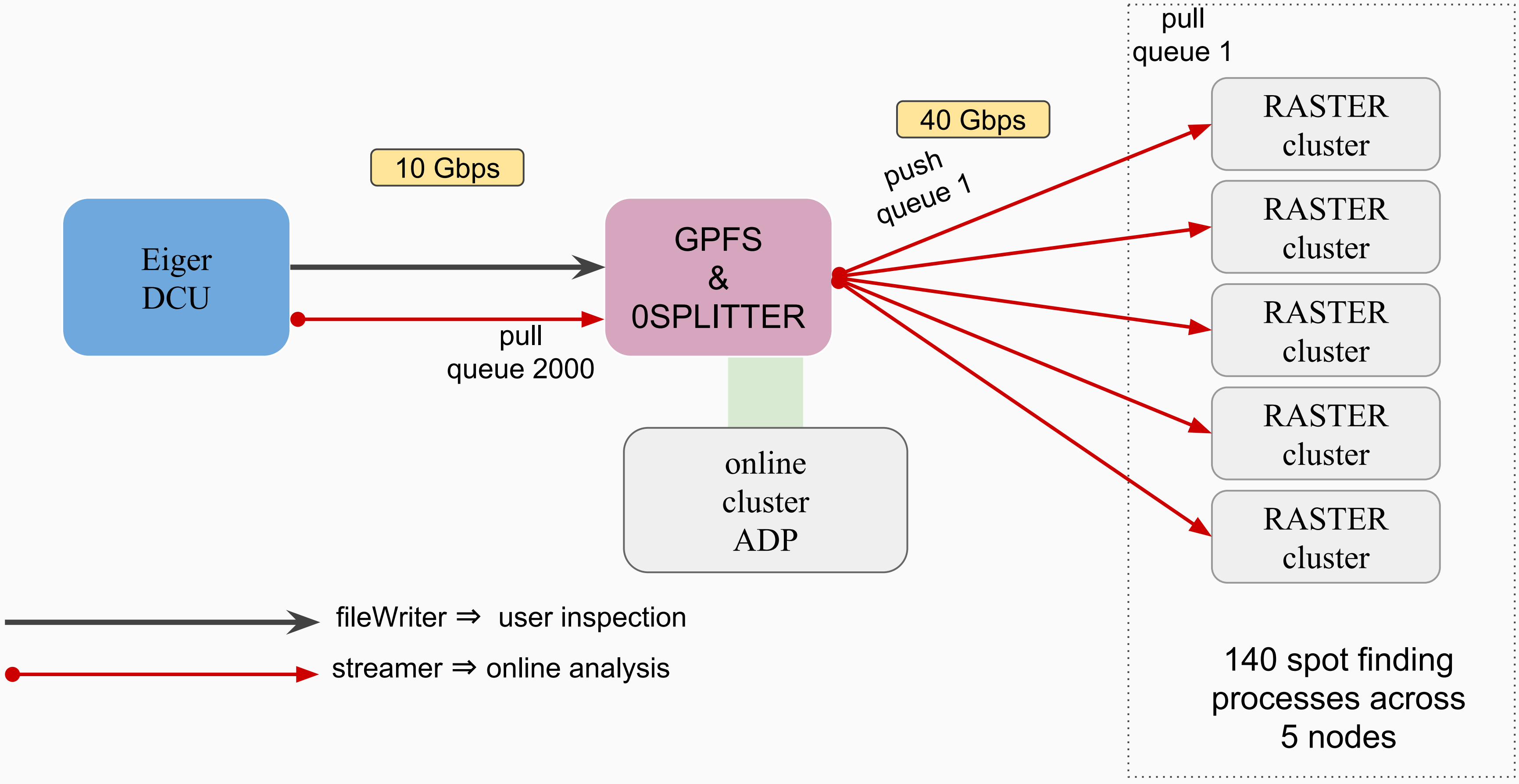


```
Summary data for        Project: XDS Crystal: XTAL Dataset: FROMXDS

                                           Overall  InnerShell  OuterShell
Low resolution limit                        39.45     39.45      1.32
High resolution limit                        1.30      7.13      1.30

Rmerge  (within I+/I-)                      0.056     0.016     0.107
Rmerge  (all I+ and I-)                     0.063     0.018     0.098
Rmeas (within I+/I-)                        0.069     0.019     0.152
Rmeas (all I+ & I-)                         0.072     0.020     0.126
Rpim (within I+/I-)                         0.040     0.010     0.107
Rpim (all I+ & I-)                          0.033     0.008     0.078
Rmerge in top intensity bin                 0.054        -         -
Total number of observations               148909      1197      1198
Total number unique                         27433       232       753
Mean((I)/sd(I))                              31.5      43.3       6.2
Mn(I) half-set correlation CC(1/2)          0.993     1.000     0.990
Completeness                                 93.8      99.5      52.3
Multiplicity                                  5.4       5.2       1.6

Anomalous completeness                       83.5      99.1      20.1
Anomalous multiplicity                        2.7       3.5       1.3
DelAnom correlation between half-sets       0.061     0.379    -0.076
Mid-Slope of Anom Normal Probability        1.023       -         -
```



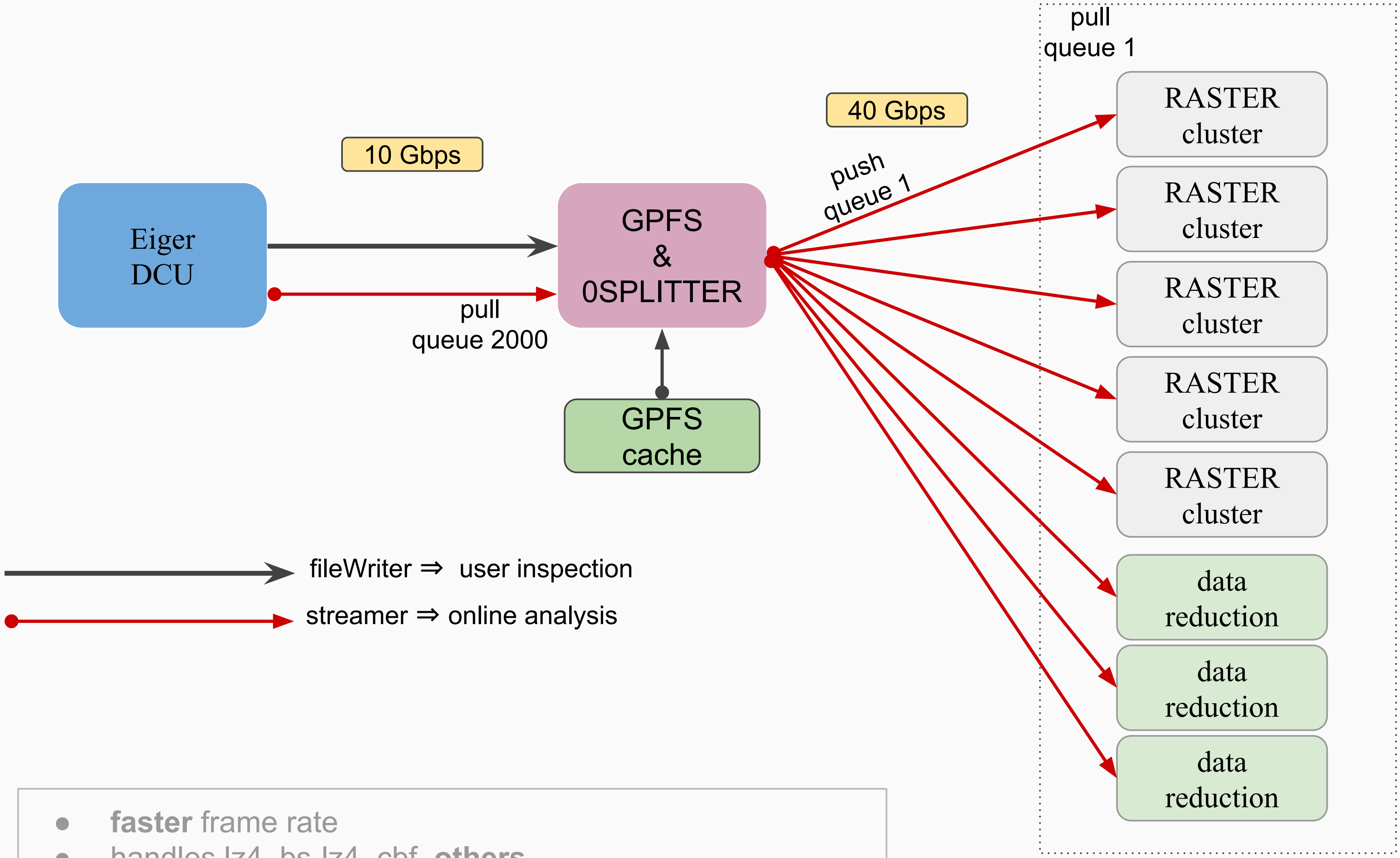
```
Summary data for        Project: DIALS Crystal: XTAL Dataset: FROMDIALS

                                           Overall  InnerShell  OuterShell
Low resolution limit                        39.47     39.47      1.32
High resolution limit                        1.30      7.12      1.30

Rmerge  (within I+/I-)                      0.031     0.017     0.121
Rmerge  (all I+ and I-)                     0.034     0.018     0.123
Rmeas (within I+/I-)                        0.037     0.020     0.165
Rmeas (all I+ & I-)                         0.037     0.020     0.151
Rpim (within I+/I-)                         0.020     0.010     0.111
Rpim (all I+ & I-)                          0.015     0.009     0.086
Rmerge in top intensity bin                 0.022        -         -
Total number of observations               145640      1204      1694
Total number unique                         27465       233       802
Mean((I)/sd(I))                              24.3      46.4       5.2
Mn(I) half-set correlation CC(1/2)          1.000     1.000     0.968
Completeness                                 93.6      99.8      56.3
Multiplicity                                  5.3       5.2       2.1

Anomalous completeness                       83.5     100.0      24.0
Anomalous multiplicity                        2.7       3.5       1.7
DelAnom correlation between half-sets      -0.072     0.074    -0.275
Mid-Slope of Anom Normal Probability        0.787       -         -
```

Eiger Online Analysis – today

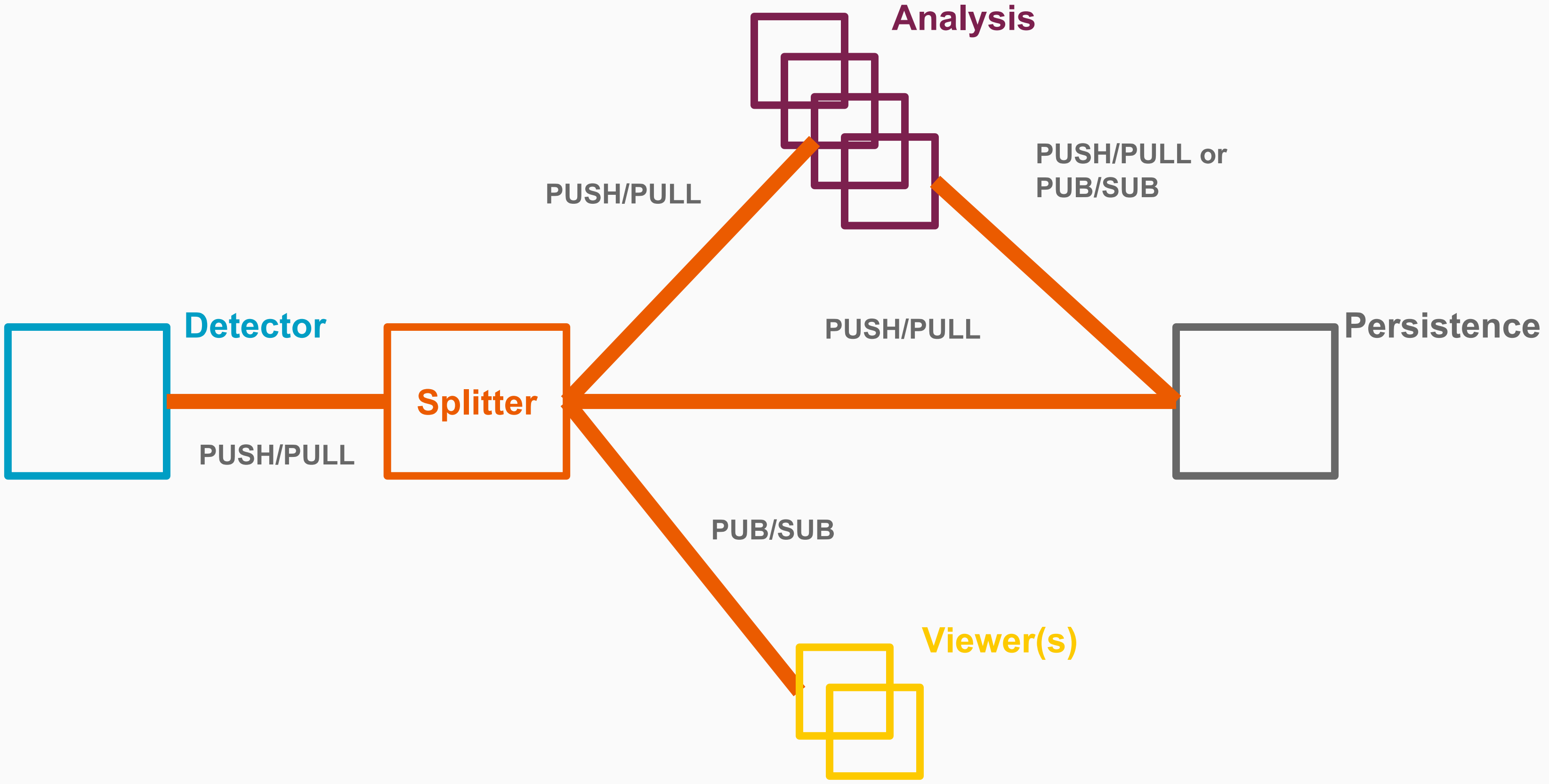


- 2.5 s per frame
- handles lz4, bs-lz4, cbf
- cannot re-analyze if needed

Eiger Online Analysis – future?



- **faster** frame rate
- handles lz4, bs-lz4, cbf, **others**
- **replay streams**


Analysis
PUSH/PULL
PUSH/PULL or
PUB/SUB
Detector
Splitter
PUSH/PULL
Persistence
PUSH/PULL
PUB/SUB
Viewer(s)