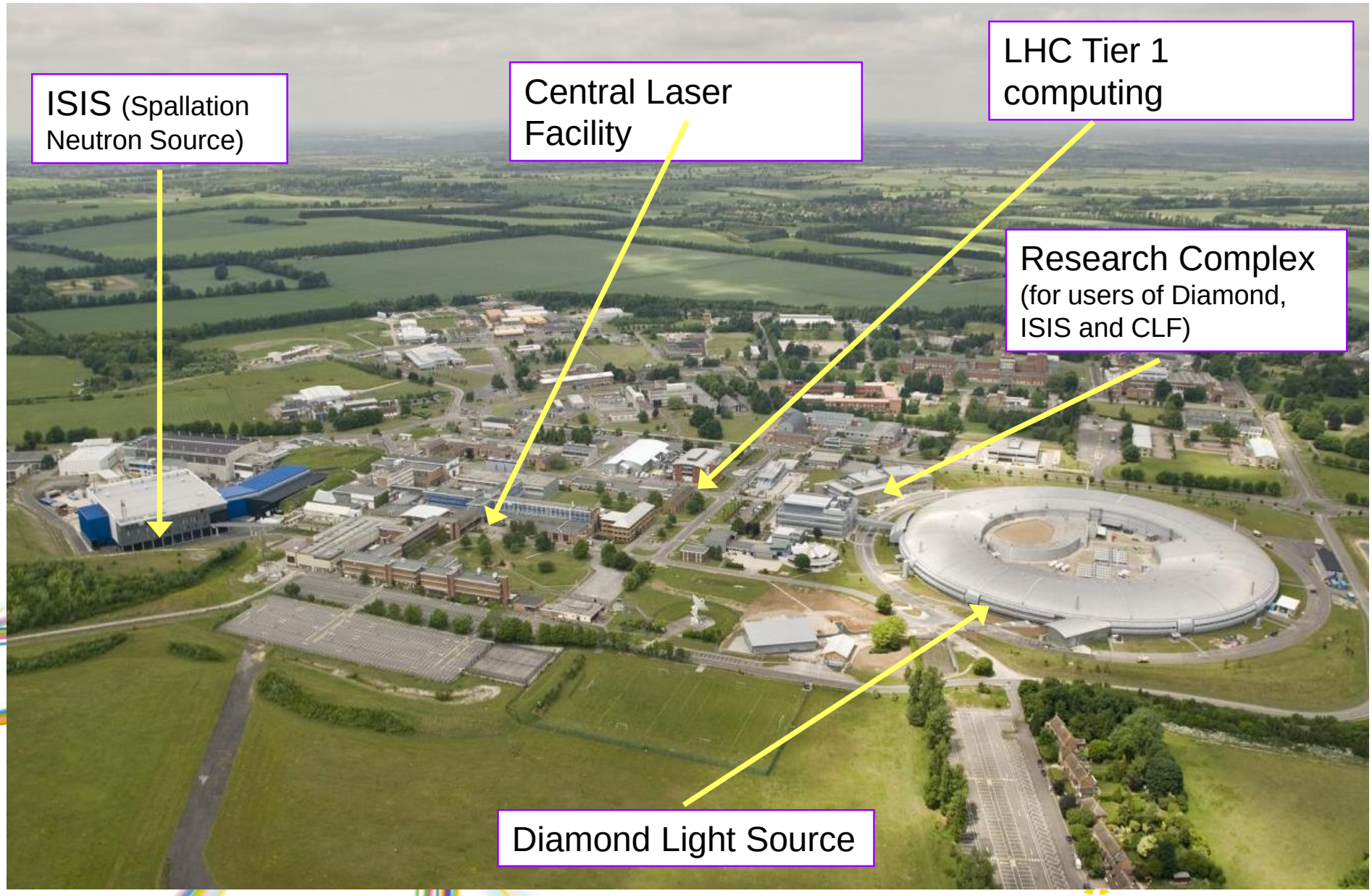
Data (I/O) Challenges at Diamond

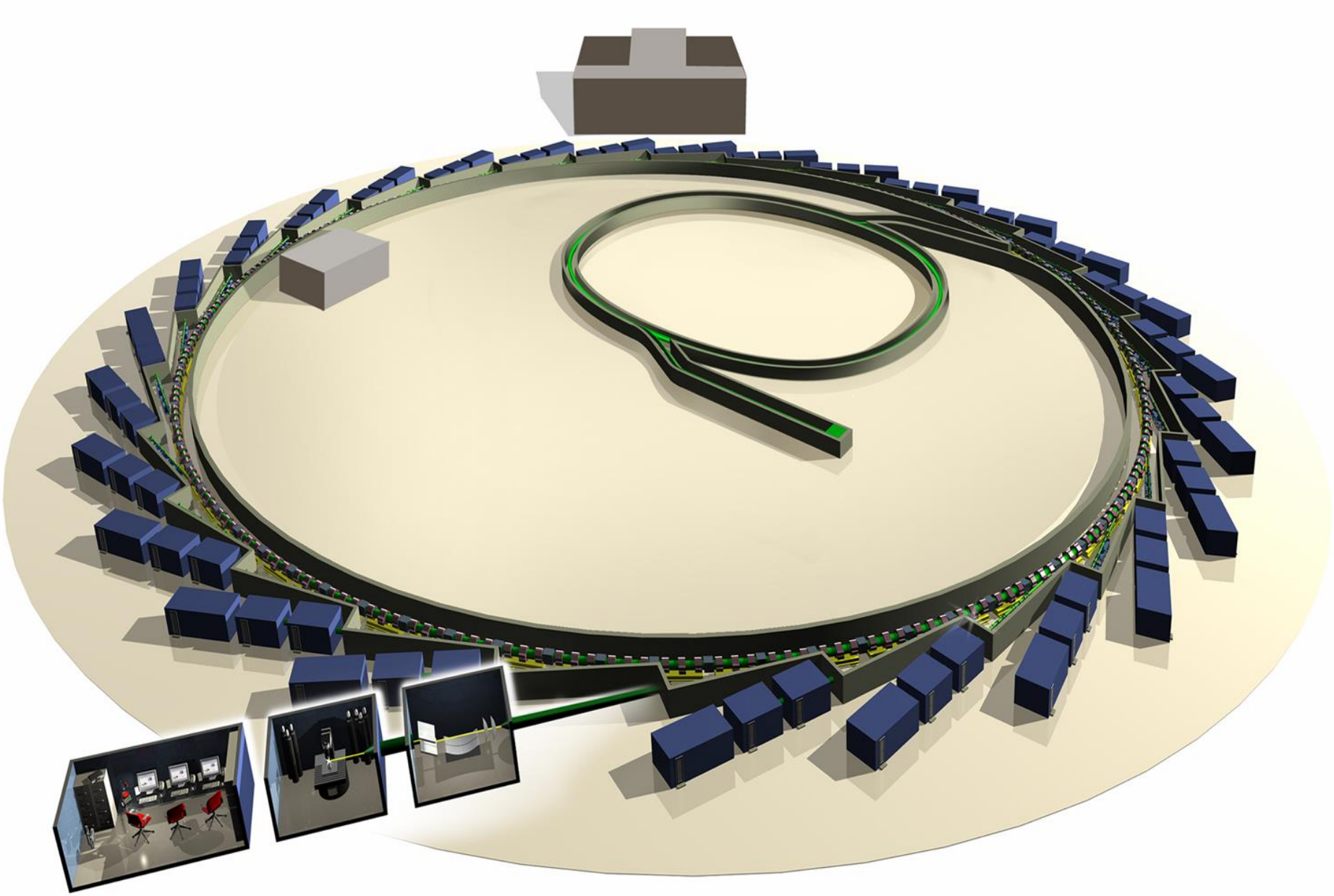
Dr Andrew Richards
Head of Scientific Computing
Diamond Light Source Ltd

andrew.j.Richards@diamond.ac.uk

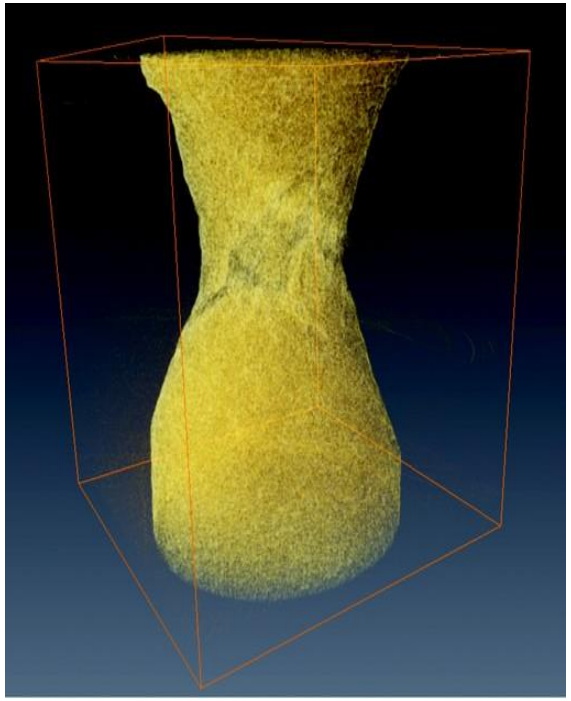
Harwell Science and Innovation Campus



Diamond Light Source

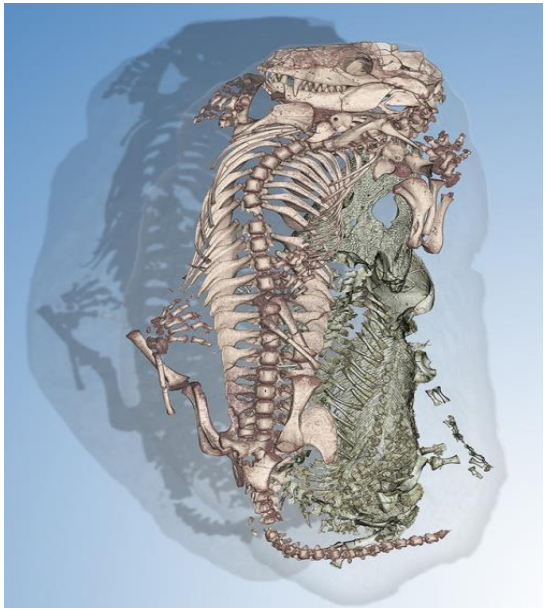


Science SR Examples

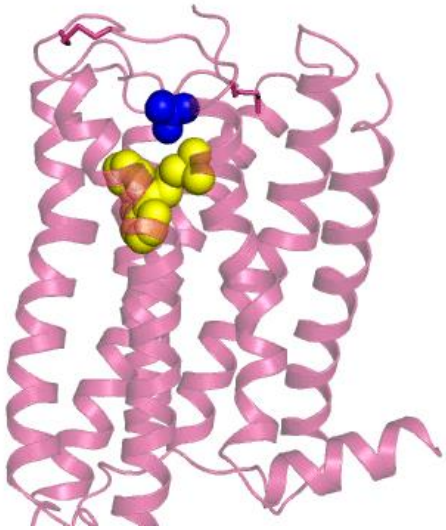


Casting aluminium

Pharmaceutical manufacture &



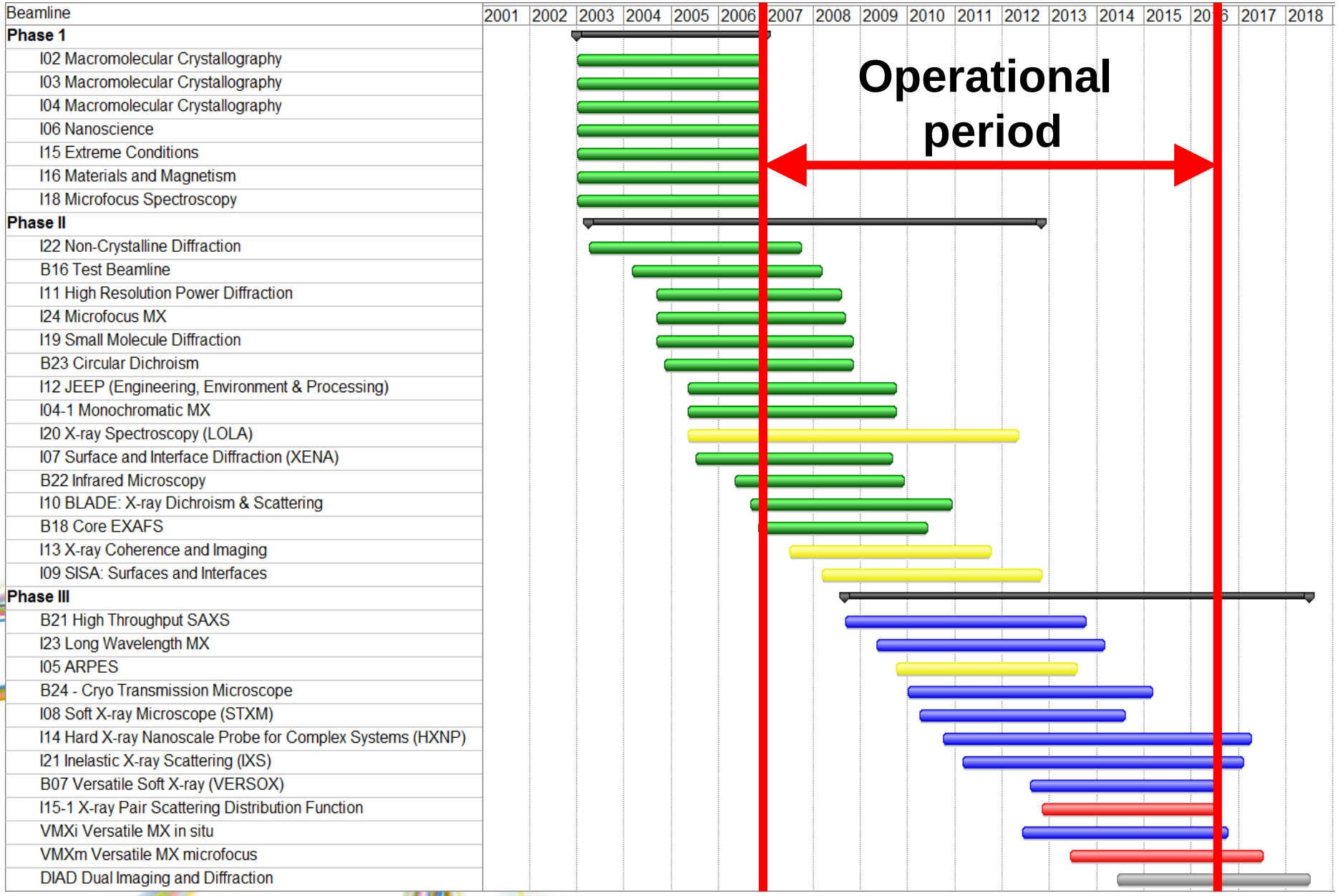
Non-destructive imaging of fossils



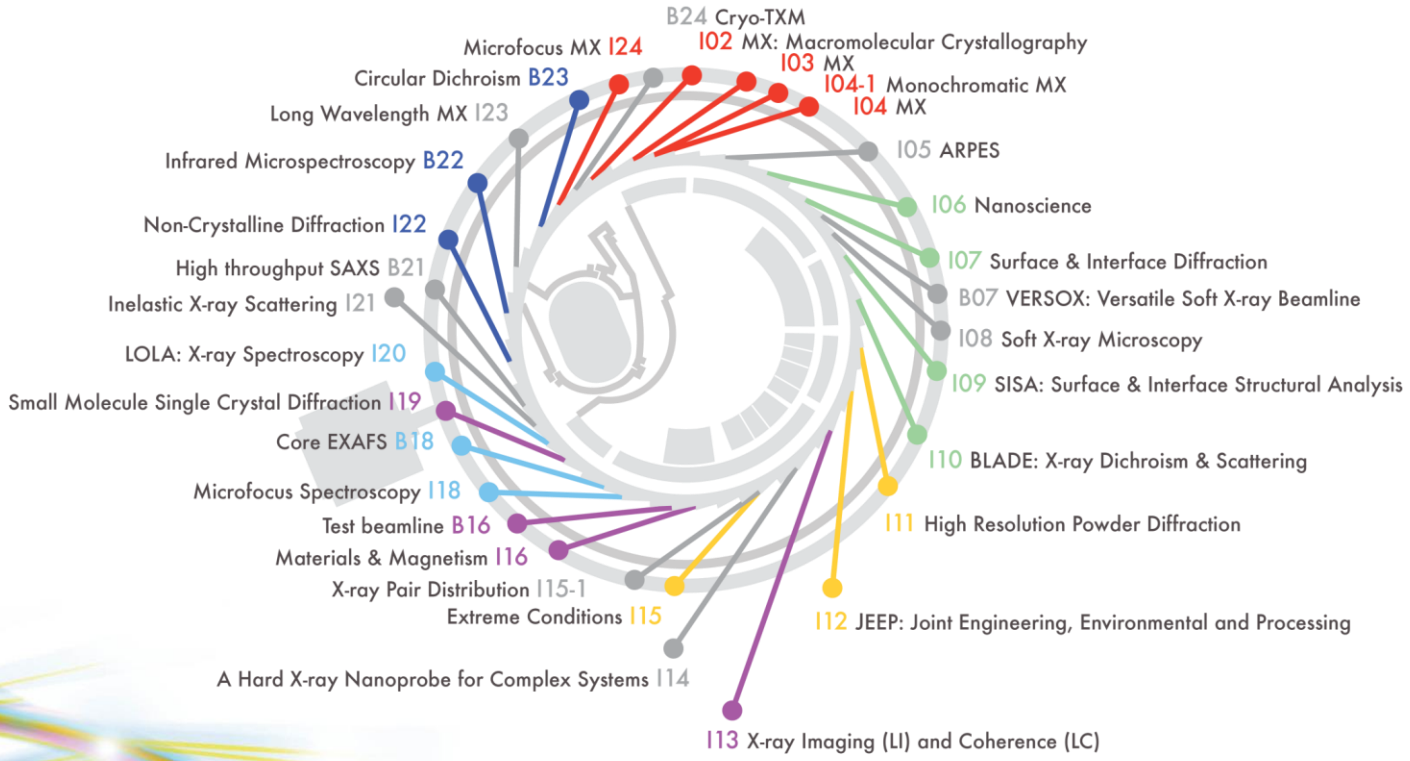
Structure of the Histamine H1 receptor

diamond

Beamlines or Instruments



Beamlines by Village



- Macromolecular Crystallography
- Soft Condensed Matter
- Spectroscopy
- Materials
- Engineering and
- Environment Surfaces and Interfaces

diamond

A National User Facility for Biological Electron Cryo-microscopy (eBIC)

Wellcome Trust Strategic Award/MRC/BBSRC, applicants:
Helen Saibil, Kay Grünewald, David Stuart, Gerhard Materlik

- Funded initially by the Wellcome Trust, MRC and BBSRC at level of £15.6 M over 5 years, augmented to ~£25 M by additional investment by the Trust in 2016
- The facility currently includes:
  - 4 high-end 300kV automated cryo EMs (Titan Krios FEI)
  - 200 keV automated feeder instrument (Talos Arctica)
  - Cryo focussed ion beam instrument (SCIOS)
  - Sample prep incl. vitreous sectioning
  - Correlative fluorescence/EM
  - FEI Polara @OPIC Oxford for CAT 3 samples

Institute of Structural Molecular Biology ISMB

Birkbeck UNIVERSITY OF LONDON

diamond

OPIC Oxford Particle Imaging Centre

University of Oxford

diamond

Electron Microscope

- Life Science EMs
  - 2x Titan Krios Electron Microscopes
  - Gatan Quantum Detector 600MB/s
- 2x Physical Science EMs to come
- 2x further Life Science EMs to come

diamond

Data Flows

- 24/6 Operation, ~5000 hours beamtime per year

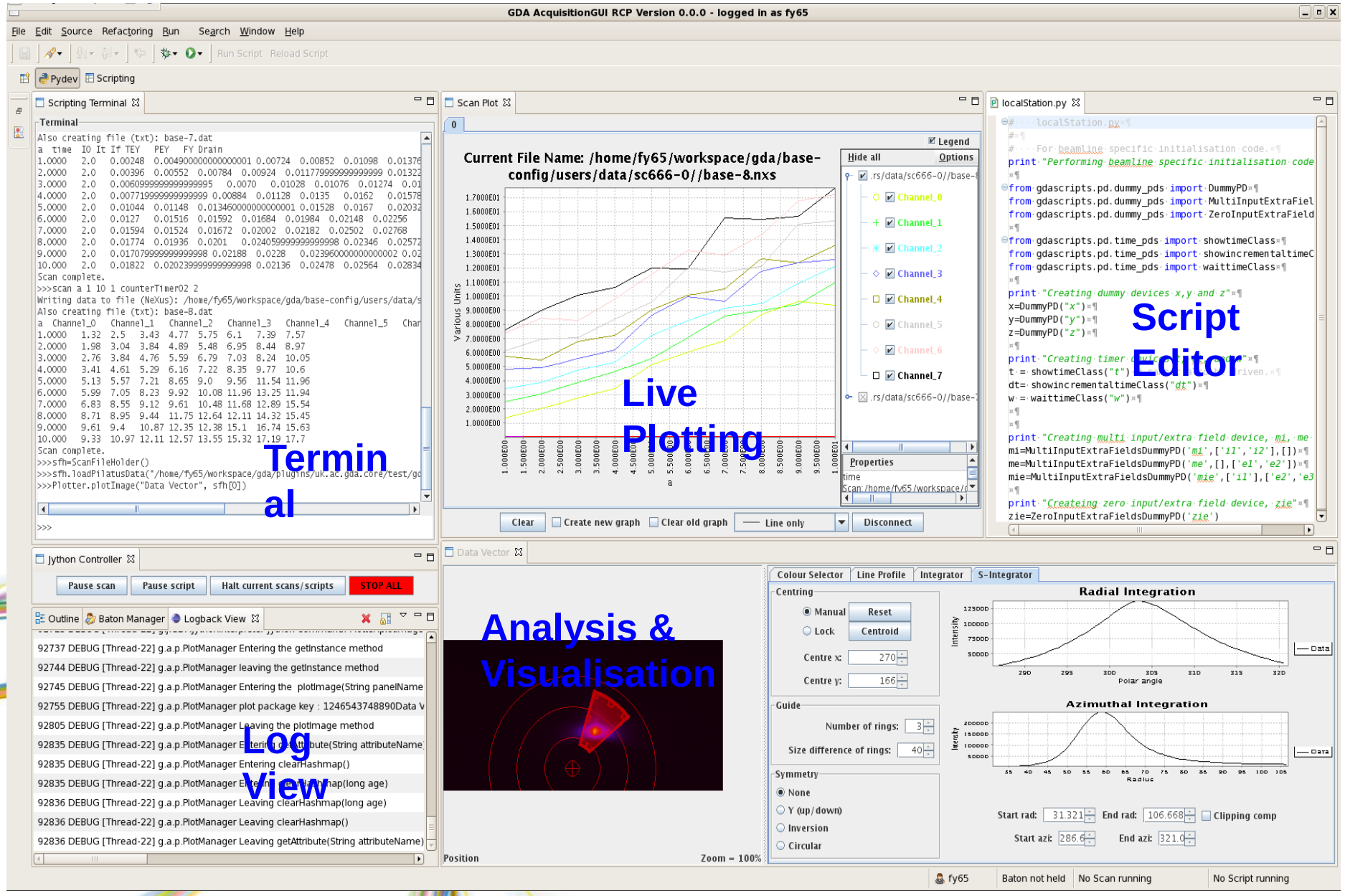
- ~30 Operational Beamlines
- ~6 Operational EM

  - Single stream performance > 900MB/s
  - Newest filesystem :
    - Aggregate performance > 60GB/s R/W
  - Existing file system(s)
    - 40GB/s

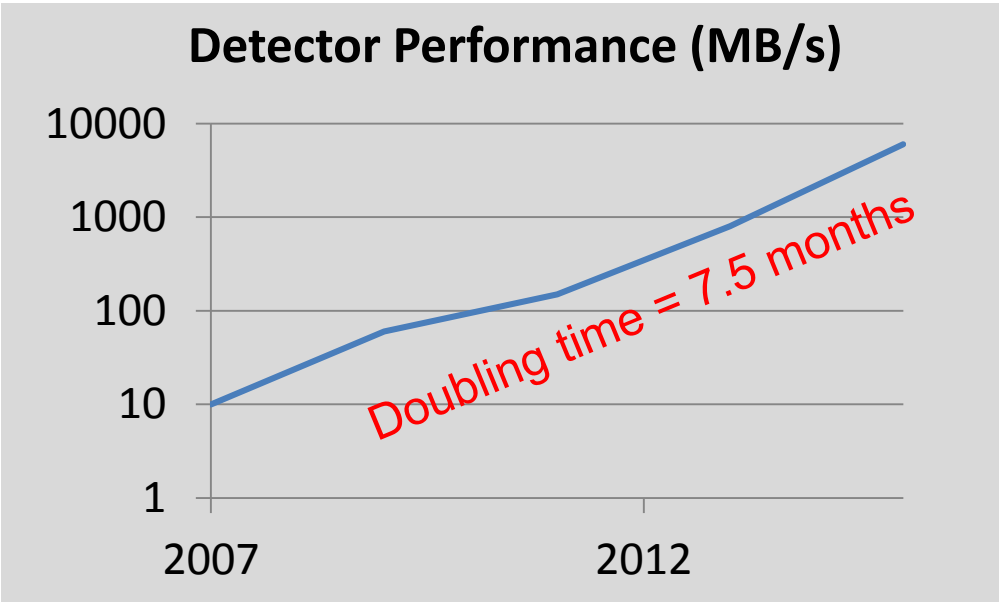
diamond

Typical User Setup

GDA – User Interface

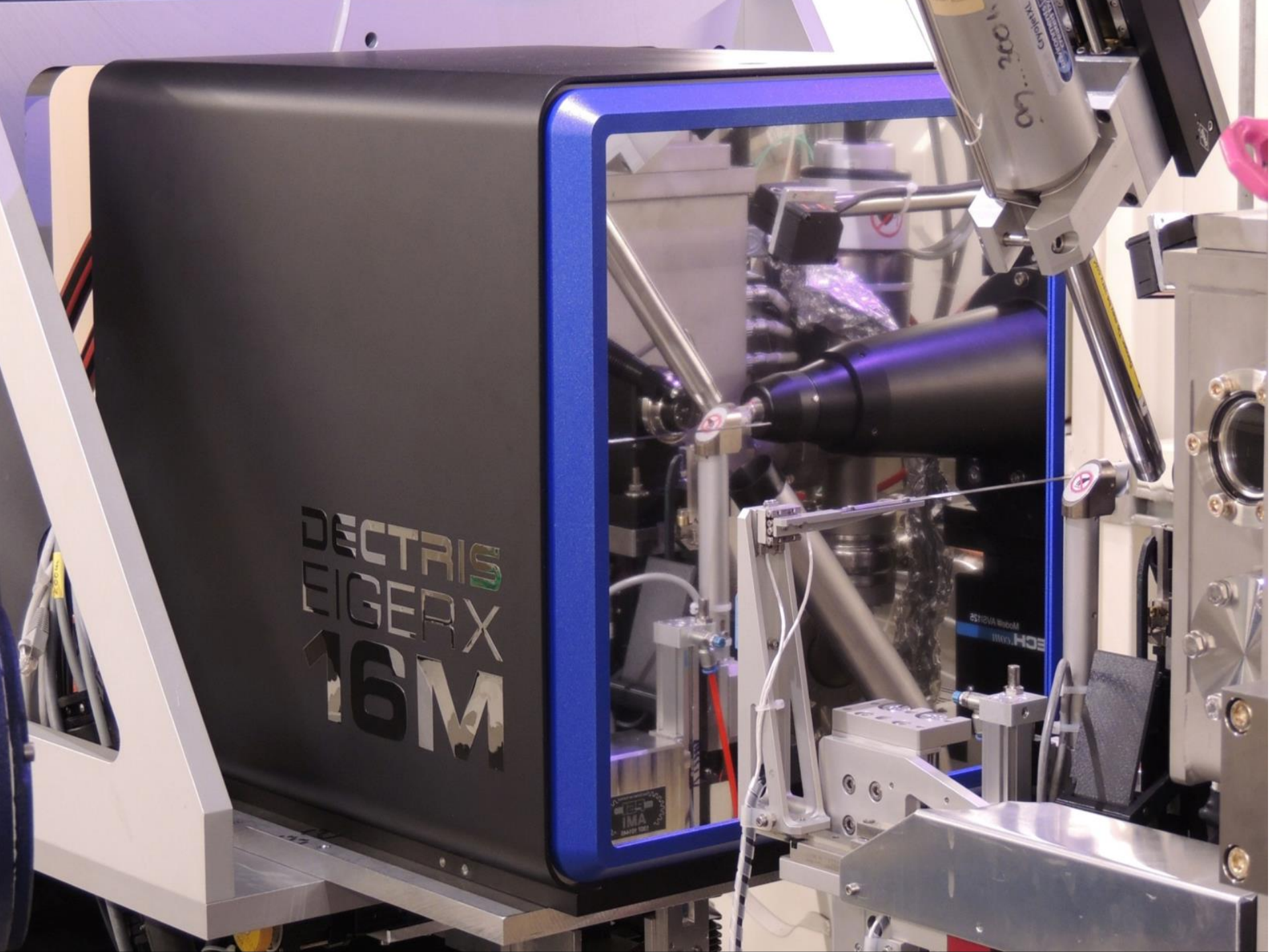


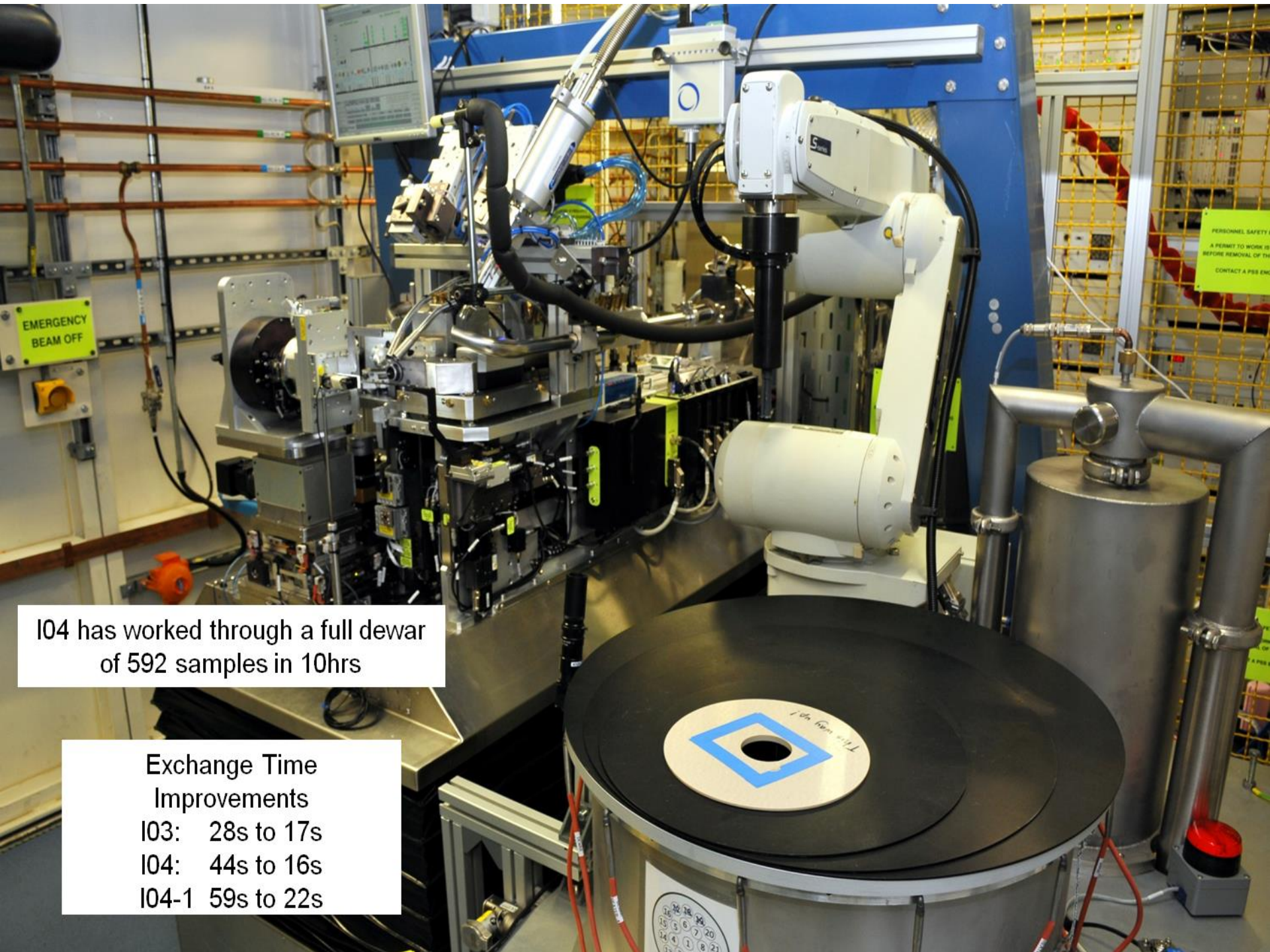
Data Rates



- 2007 No detector faster than ~10 MB/sec
- 2009 Pilatus 6M system 60 MB/s
- 2011 25Hz Pilatus 6M 150 MB/s
- 2013 100Hz Pilatus 6M 600 MB/sec
- 2013 ~10 beamlines with 10 GbE detectors (mainly Pilatus and PCO Edge)
- 2016 Percival detector 6GB/sec

diamond


DECTRIS
EIGER X
16M



I04 has worked through a full dewar of 592 samples in 10hrs

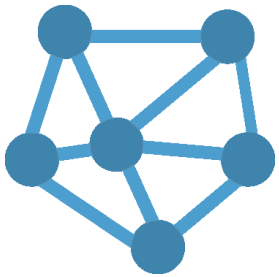
Exchange Time Improvements

I03: 28s to 17s

I04: 44s to 16s

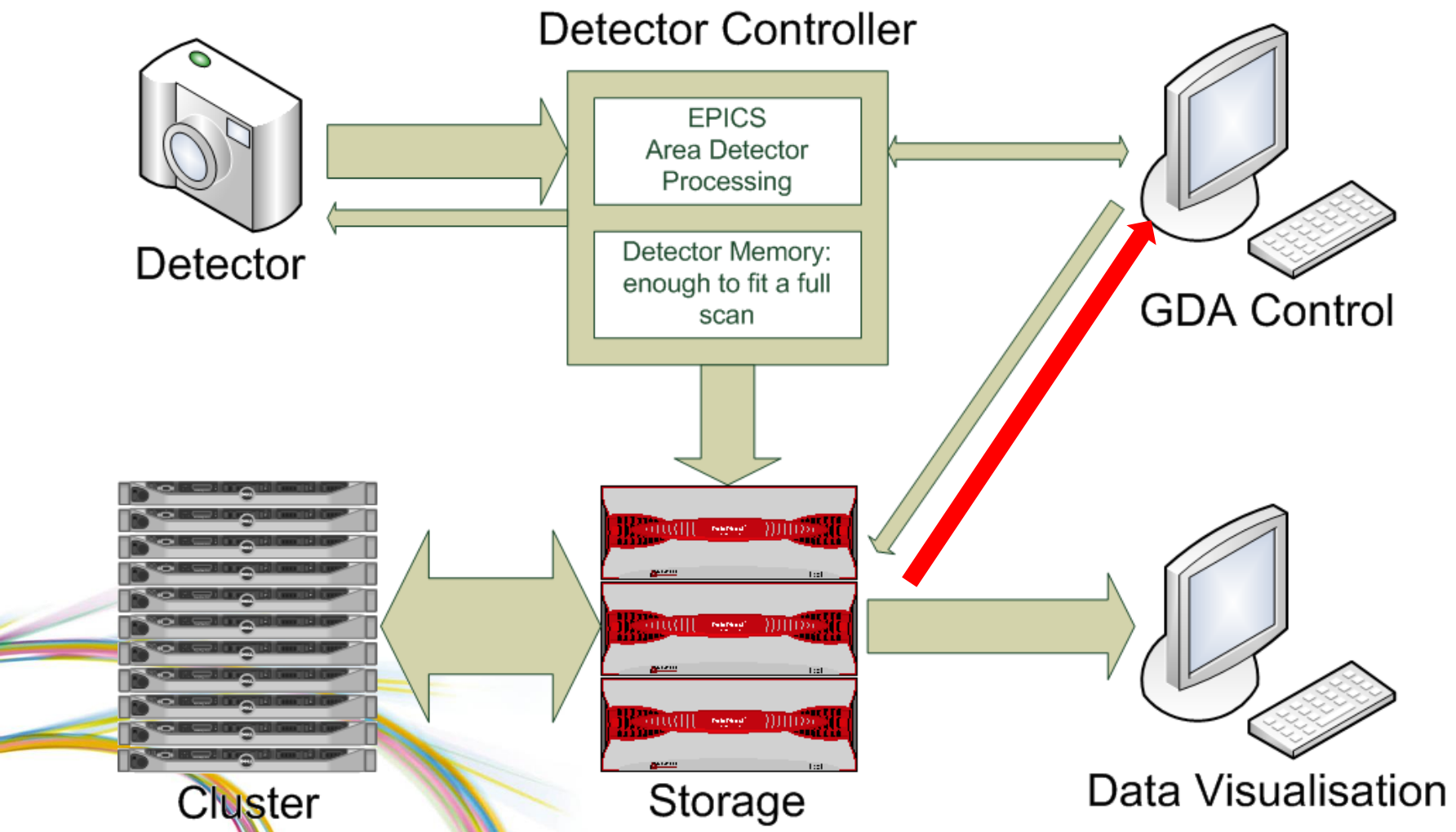
I04-1 59s to 22s

Scientific Computing and Infrastructure at Diamond

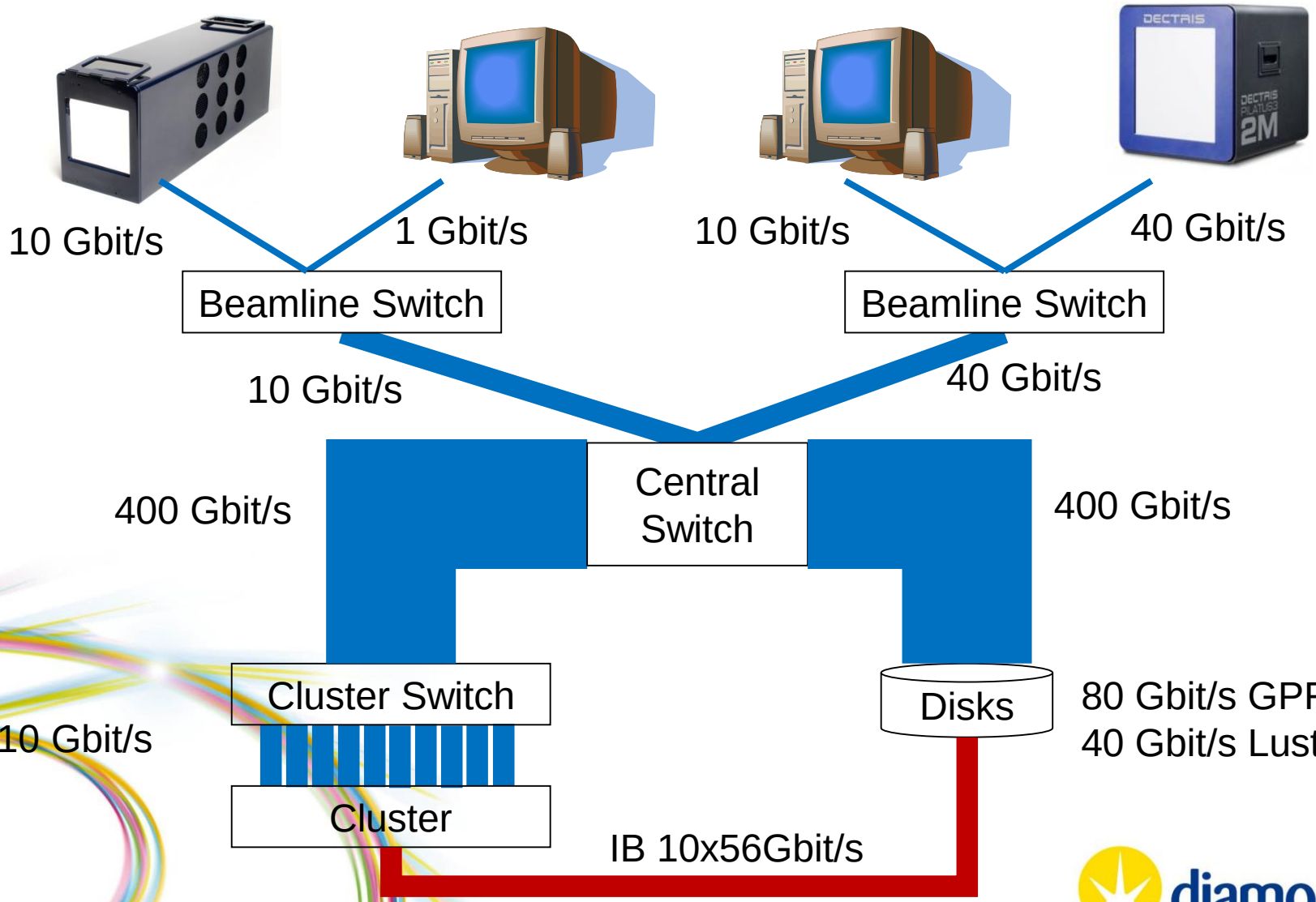


diamond

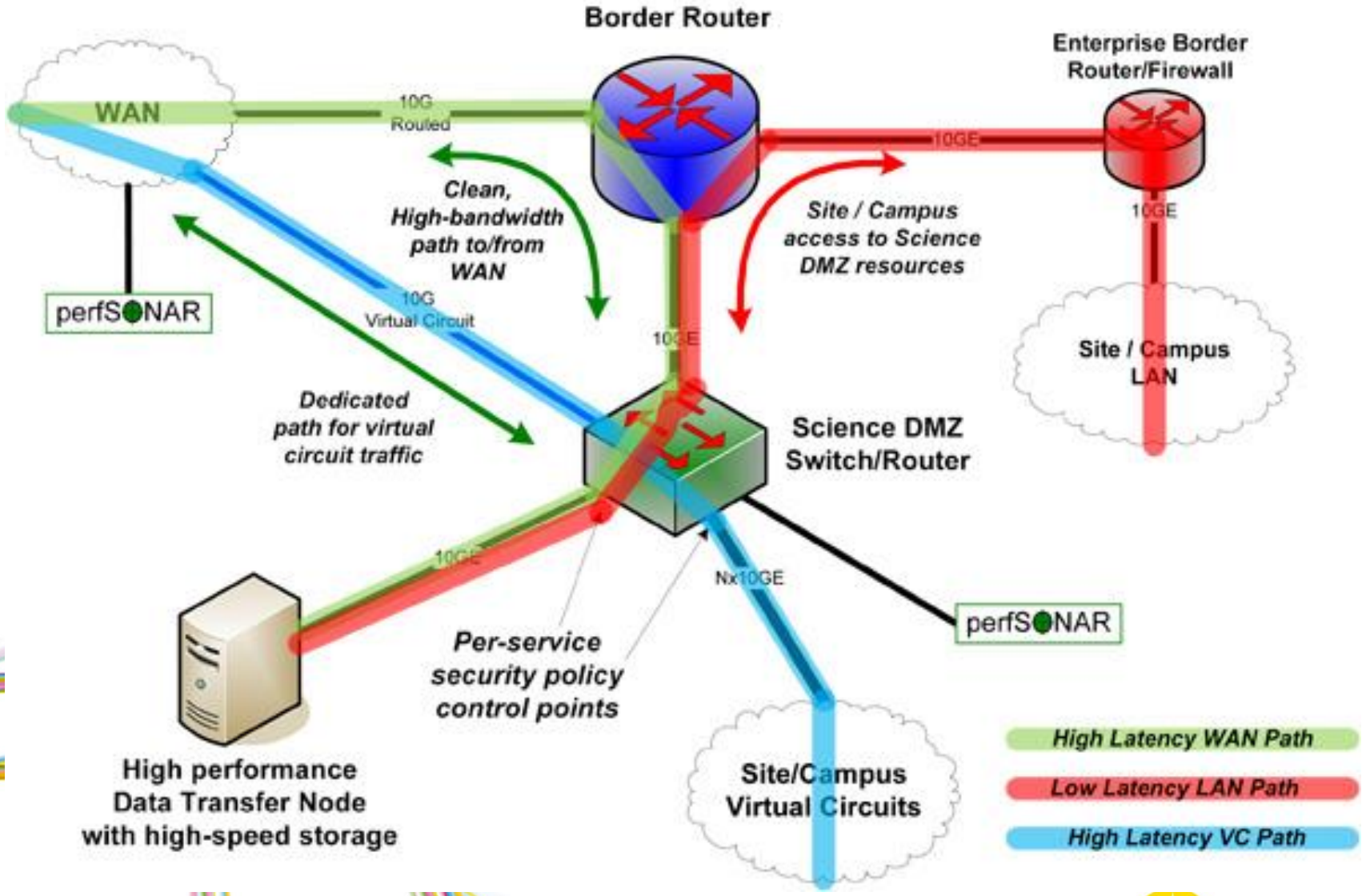
Data Flow


Detector Controller
EPICS
Area Detector
Processing
Detector Memory:
enough to fit a full
scan
Detector
GDA Control
Cluster
Storage
Data Visualisation

Network Bandwidth Balance



Moving Data Off Site a Science DMZ



diamond

Scientific Computing Infrastructure

- HPC / HTC Cluster (~3500 cores)
  - *X86, Nvidia GPU (K80, P100)*
- High Performance Storage (~7.5PB)
  - *Lustre03, Lustre04, GPFS01, GPFS02*
- Network infrastructure
  - *10Gb/s, 40Gb/s to some beamlines*
- User Gateways, Visualisation, Data Transfer
  - NX Service, Globus endpoint
- Support
  - Predominantly Linux infrastructure,
  - BUT also Windows support to beamlines/EM/etc and VM platforms
  - Relies on working with Corporate IT and other groups in Controls and Scientific Software

diamond

Statistics: Data

Target	Available	Used	Performance
XFS	50 TB	47 TB	< 1GB/s
Lustre03	470 TB	370 TB	6 GB/s
Lustre04	140 TB	70 TB	2 GB/s
GPFS01	1 PB	700 TB	15 GB/s
GPFS02	4.7 PB	3.5 PB	40 GB/s
NEXT ???	***~6-9PB***	***-***	***>60GB/s***
STFC Archive	n/a	12 PB	12 TB – 50 TB per day ingest

diamond

Current File systems

- 2x Lustre systems sub 1PB. (not doing a great deal now but worthy of note)
  - DDN based systems SFA10K and SFA12K

- 2x GPFS
  - (1) 877TB usable (GPFS01)
  - 4 nsd servers
  - 2 protocol servers
  - *Over the last 6 months actual data
    - *Read 2GB/s
    - *Write 6GB/s

  - (2) 4.7PB usable (GPFS02)
  - 7 NSD servers, 4 protocol nodes
  - Netapp E series and EF for metadata
  - IB connected to storage
  - IB and 10G out the front
  - *40GB/s benchmarked throughput. Now in excess of as we have added disks

‘Big’ Data Lifecycle challenges

- How much do you mean by BIG?
- How ‘FAST’ do you need to analyse the data?
- What data can be THROWN AWAY?
  - (and at what stage?)
- How LONG do you need to keep the data?
- And WHERE? Where do you want to transfer the data to/from?

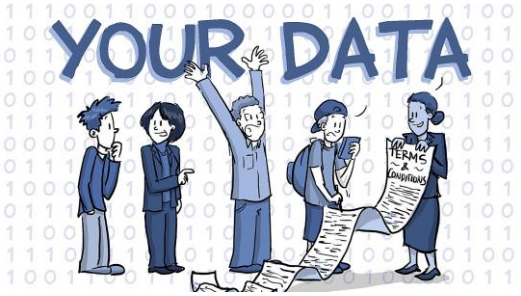
- And WHERE do we best do Post-Processing?

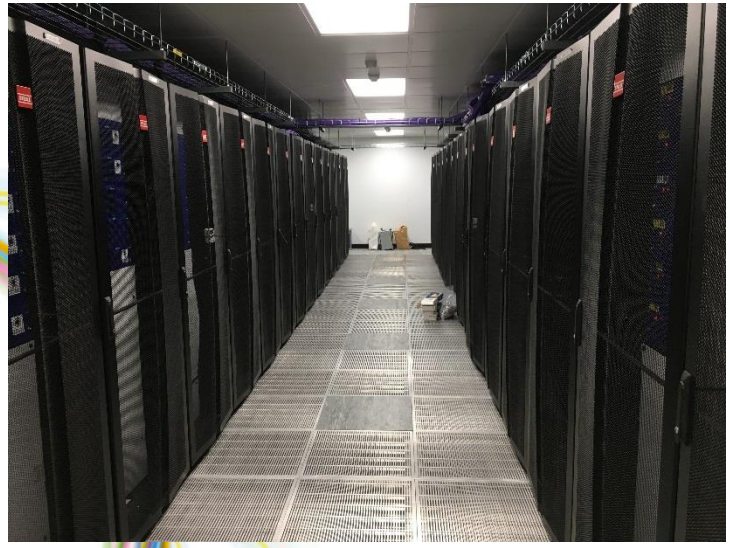
diamond

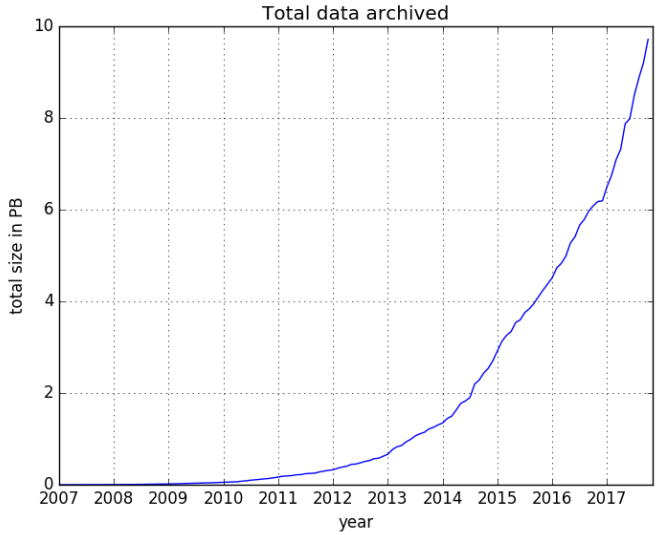
Scientific Computing

New Computer Room

(CSCR3 – Inner courtyard)







12 PB of Archived Data

diamond

Software Developments / Future ?

- HDF5 Virtual Dataset (VDS) to map multiple files into a single, coherent dataset "view" (i.e. a VDS)
  - No longer use MPI-IO
- SWMR –Single Write Multiple Read
    - HDF5 extension
- ZeroMQ increasingly used for data streaming
- FPGA?
  - Smarter data collection / filtering at source
  - New detectors and data rates exceed disk bandwith

diamond

Detector Developments

Percival

- Main challenge: 6GB/s data rate, sustained forever(ish).
- Secondary processing challenge: pixel-per-pixel bit descrambling and pixel-per-pixel calibration algorithms prior to writing an "image" to disk.

Excalibur

- Data-rate 750MB/s * 2 channels.
- real problem: ptychography and tomo-ptychography - very large scans with TB size datasets.
- Secondary processing challenge: block-by-block (i.e. each sensor chip) is angled and need to be rotated 90deg to turn into an "image". Not terribly difficult because of the low(ish) data-rate.
- Secondary processing challenge: on-the-fly compression. Data is sparse so we expect it to be relatively easy to get a good compression ratio.

Eiger and Eiger2

- Data-rate: variable, depending on compression ratio which depends on SNR. Typically the compressed stream fits in a single 40gbps NIC.
- Raw data rate: 4M pixel * 2bpp * 750Hz
- Pre-compressed data can not be processed on the fly before hitting disk.
- Main challenge: near 100% reliability and uptime! MX beamlines are data factories!

- All systems require a degree of parallel DAQ/processing and so that is a key challenge that we face in the software development of all of these systems. We are implementing a single DAQ framework to allow re-use of software for all of these systems.

diamond

New Eiger Detectors

- Eiger on VMXi, we currently see:
- Sustained (10 minutes+/indefinite) acquisitions at 500Hz, with typical compressed data size of 2.2Mbs per frame. This gives about ~ 9 Gbits/second.
- Full speed acquisition burst (30 – 60 seconds) at 750Hz, with typical compressed data size of 2.2Mbs per frame. This gives ~13 Gbits/second.
- This is using 4 writing nodes split across 2 servers.
- Files contain 1000 frames, so in big acquisitions, it's possible to get reasonably large numbers of files being created in the same directory. (We once did a 500000 frame acquisition which created 500 files, though this is not a typical use case).

diamond

Monitoring IO

Monitoring IO

File system Challenges

- Complexity of overall setup
  - E.g. >30 + multi clusters in current GPFS setup
    - Trust relationships
    - Metanode issues ?
    - Other unkown issues
- Need for parallel I/O to cluster
  - Good single stream / single node performance
    - RDMA (typically over IB) need over ETH?
      - Currently restricts Lustre performance for example
  - NFS/CIFS etc presentation to clients

diamond

Thank you