Lustre as Data Acquisition File System at Diamond Light Source

Frederik Ferner

Diamond Light Source Ltd

24 April 2012

diamond

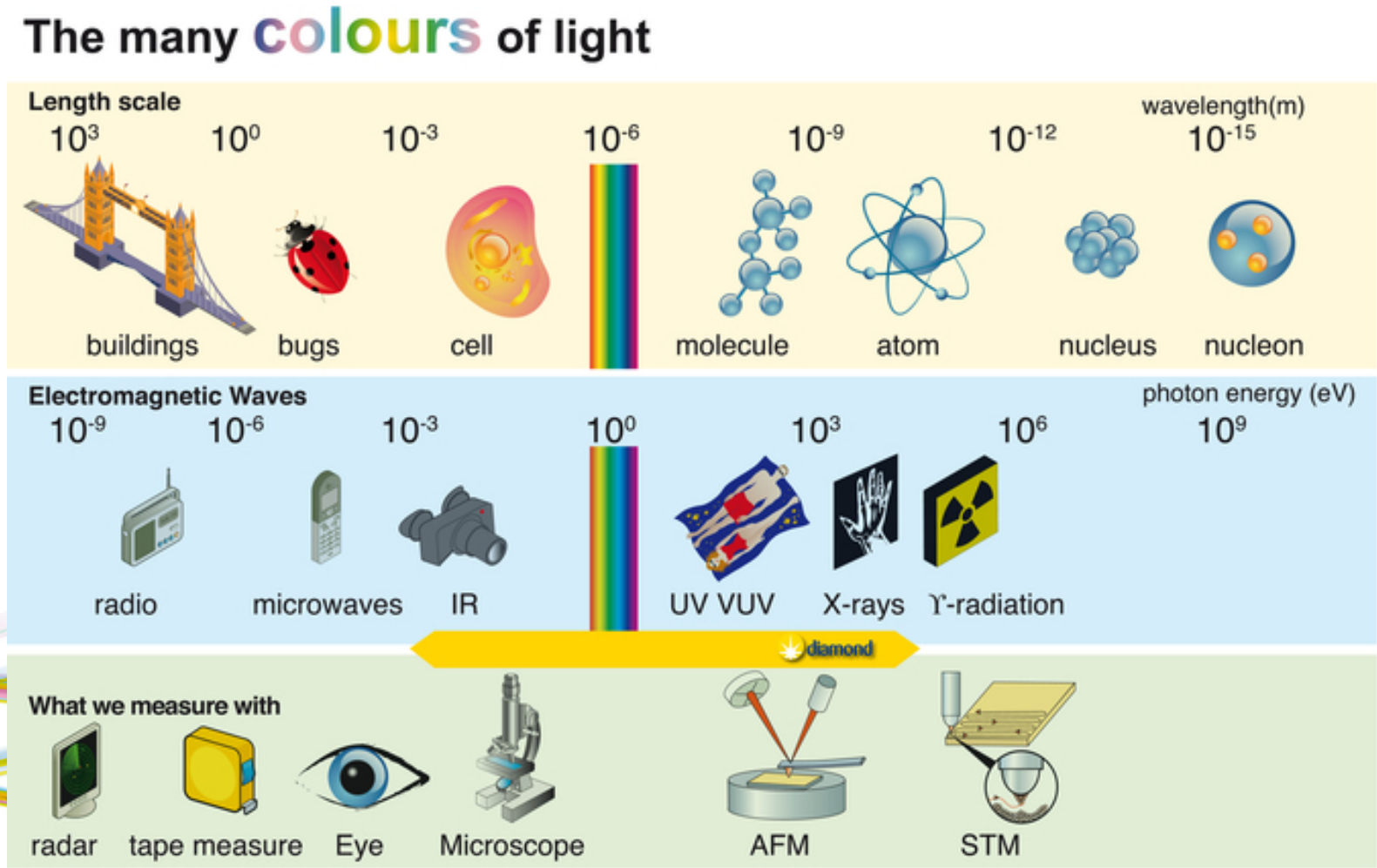
Outline

Diamond Light Source

Diamond Light Source (DLS) is the UK's national synchrotron facility. It is located at the Harwell Science and Innovation Campus in Oxfordshire, UK.

- third generation light source
- 561.6 m circumference storage ring; energy 3GeV
- largest scientific investment in the UK in 45 years
- first users 2007

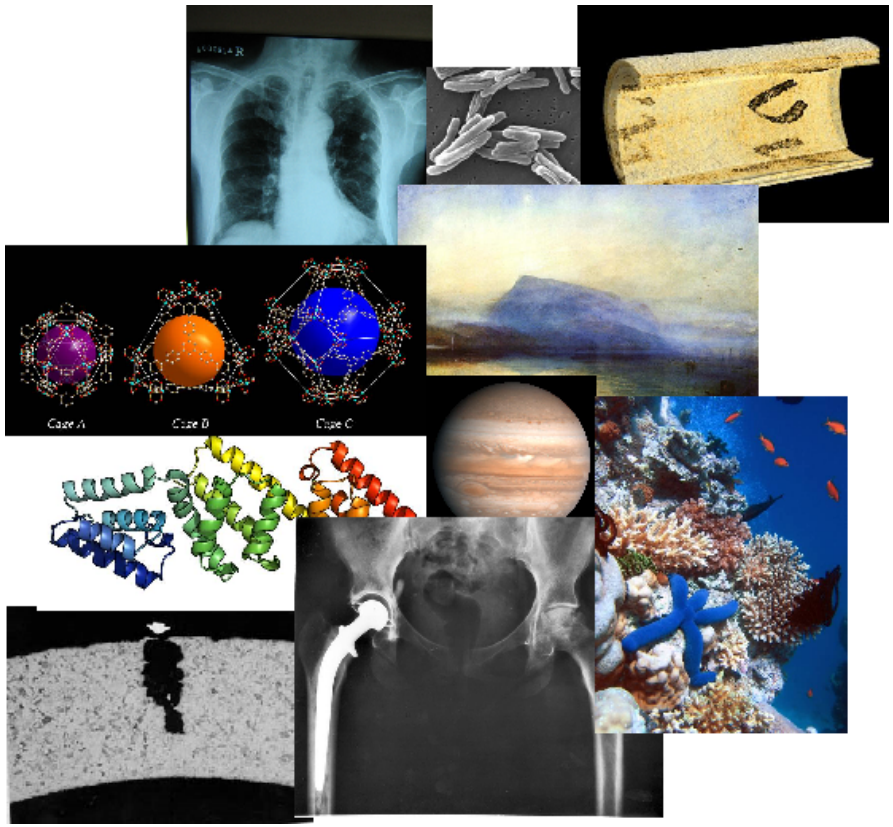
Light spectrum generated at Diamond



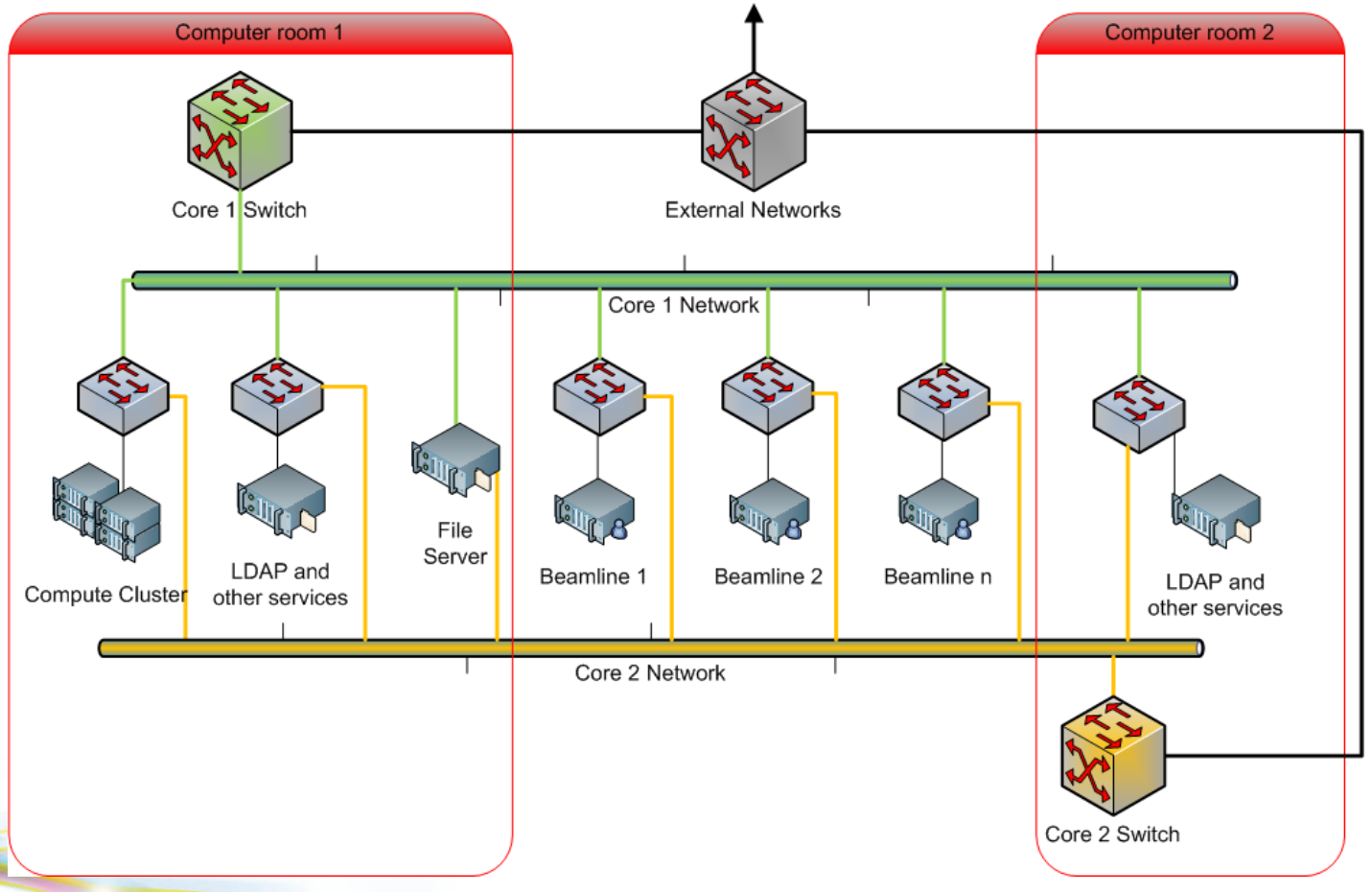
Science done at DLS

Diamond provides facilities for different fields

- Archeological preservation
- Bioscience research
- Climate change
- Nanotechnology
- “Green” Technologies
- Extreme conditions
- Medical science
- Material science

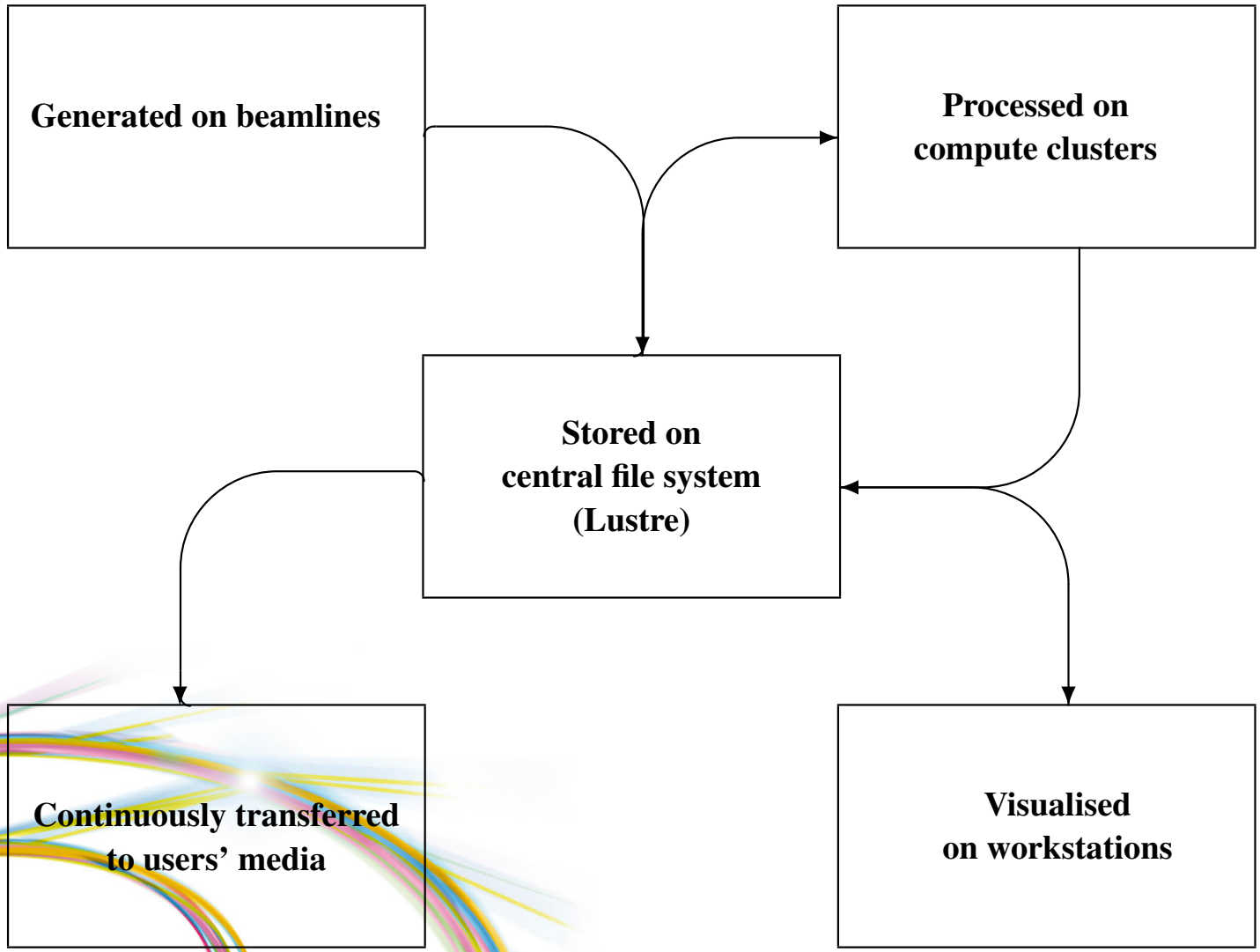


The Science Network



- All beamlines and service networks dual homed
- OSPF and ECMP
- 1GigE or 10GigE uplinks to core

Data Flow


Generated on beamlines
Processed on compute clusters
Stored on central file system (Lustre)
Continuously transferred to users' media
Visualised on workstations

Lustre File Systems, lustre01

First Lustre file system commissioned end of 2008

- 400TB raw (~300TB usable); >60% full
- DDN S2A 9900 for OSTs, MD3000 for MDT
- PE2970 for OSS and MDS
- 6 OSSs in active-active fail-over pairs
- MDS pair as active-passive fail over
- servers connected to core networks via 10Gbit Ethernet
- aggregate write speed ~3.5GB/s

Lustre File Systems, lustre03

New Lustre file system commissioned early 2011

- 600TB raw (~400TB usable)
- DDN SFA 10K for OSTs, EFI 3015 for MDT
- PE R610 for OSS and MDS
- 4 OSSs in active-active fail-over pairs
- servers connected to core networks via 2x 10Gbit Ethernet bonded links
- throughput ~5.5GB/s

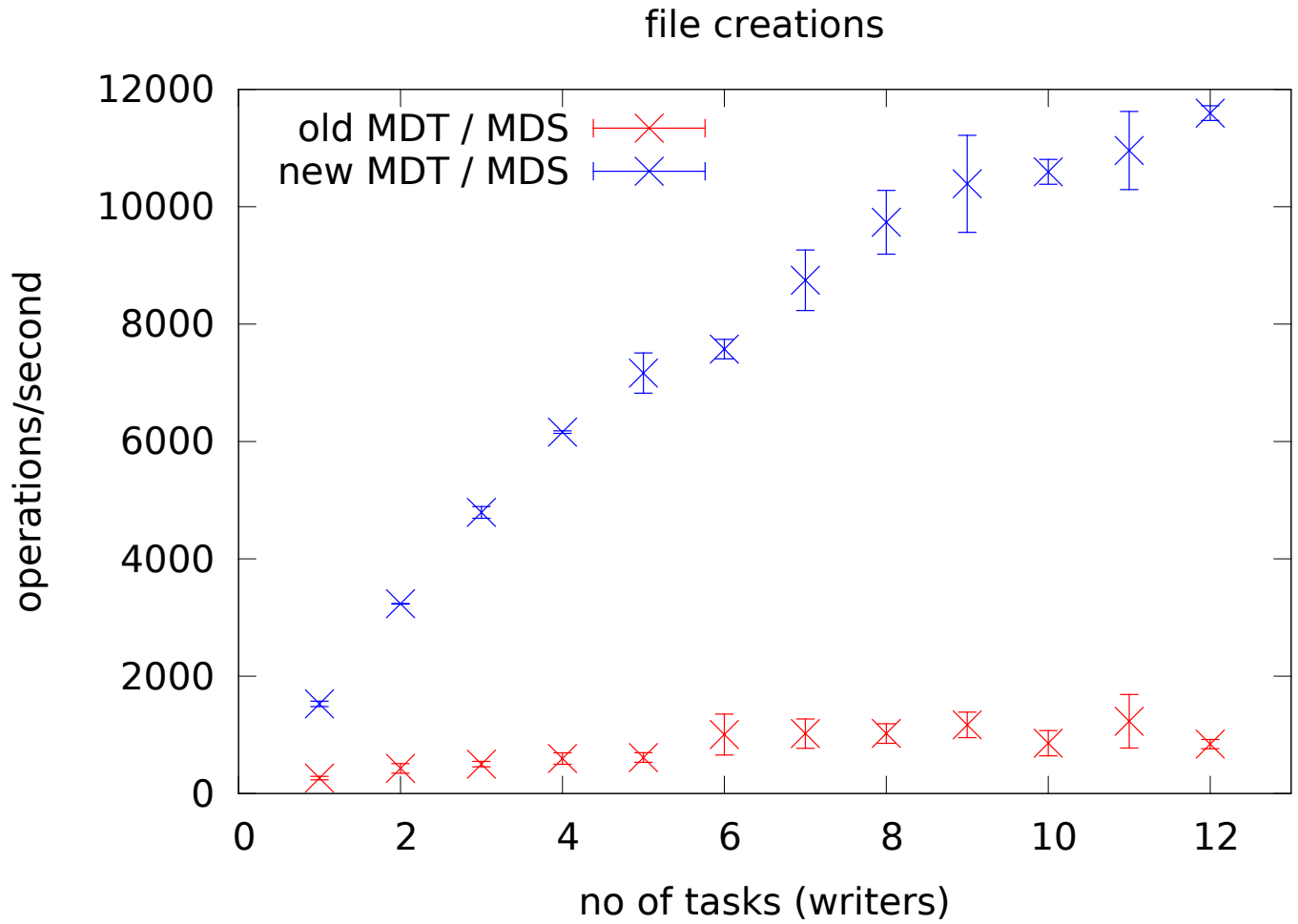
Recent Upgrade: MDT for lustre01

Earlier this year we upgraded our MDS/MDT for lustre01.

- replaced MD3000 with MD3200 as MDT
- upgraded the MDS servers from PE2970s to R610s.
- MDT transferred using dd over the network

diamond

MDT upgrade: mdtest results



MDtest[1] results before/after MDT replacement.

[1] mdtest -I 10 -z 5 -b 5 -i 5 -u -d ${TESTDIR} creating 39060 objects per client

Particular Challenges through Data Acquisition use

- clients distributed across the building
- variety of different applications (relative small files, large files), current detectors have throughput range from $30\frac{MB}{s}$ to $200\frac{MB}{s}$
- any interruption can result in lost data
- users are impatiently waiting for their data ('watch ls -ltr' not uncommon)
- strict access control (many ACLs)
- access from Windows required

Data Acquisition Example: Pilatus Detector

One common
detector type at DLS: Pilatus

- generated file size: 6MB
- frame rate: up to 25Hz currently
- used for long scans
  taking 10000+ images per scan
- online data
  processing on compute clusters

Data Acquisition Example: Tomography

Tomography...

- frame size 20MB
- frame rate 5Hz (next generation 70Hz)
- Windows based cameras
- individual tiff files (move to HDF5 files planned)
- next generation will use parallel HDF5 and one output file, through in house application to capture data
- data processing on GPU cluster

Lustre Monitoring

- health checks via Nagios/Zenoss
- performance monitoring
  - ganglia
  - collectl
  - lmt under investigation
- users

Lustre feature wishlist

Feature wishlist

- improved file system access from Windows
  - both mostly read (users workstations)
  - and mostly write (detector control machines)
- increase number of ACLs per file/directory
- NFSv4 ACLs (maybe?)
- monitoring (snmp?)

What can we do to help?

Thank You!

Many thanks also to the whole team at Diamond:
Tina Friedrich, Greg Matthews, Nick Rees