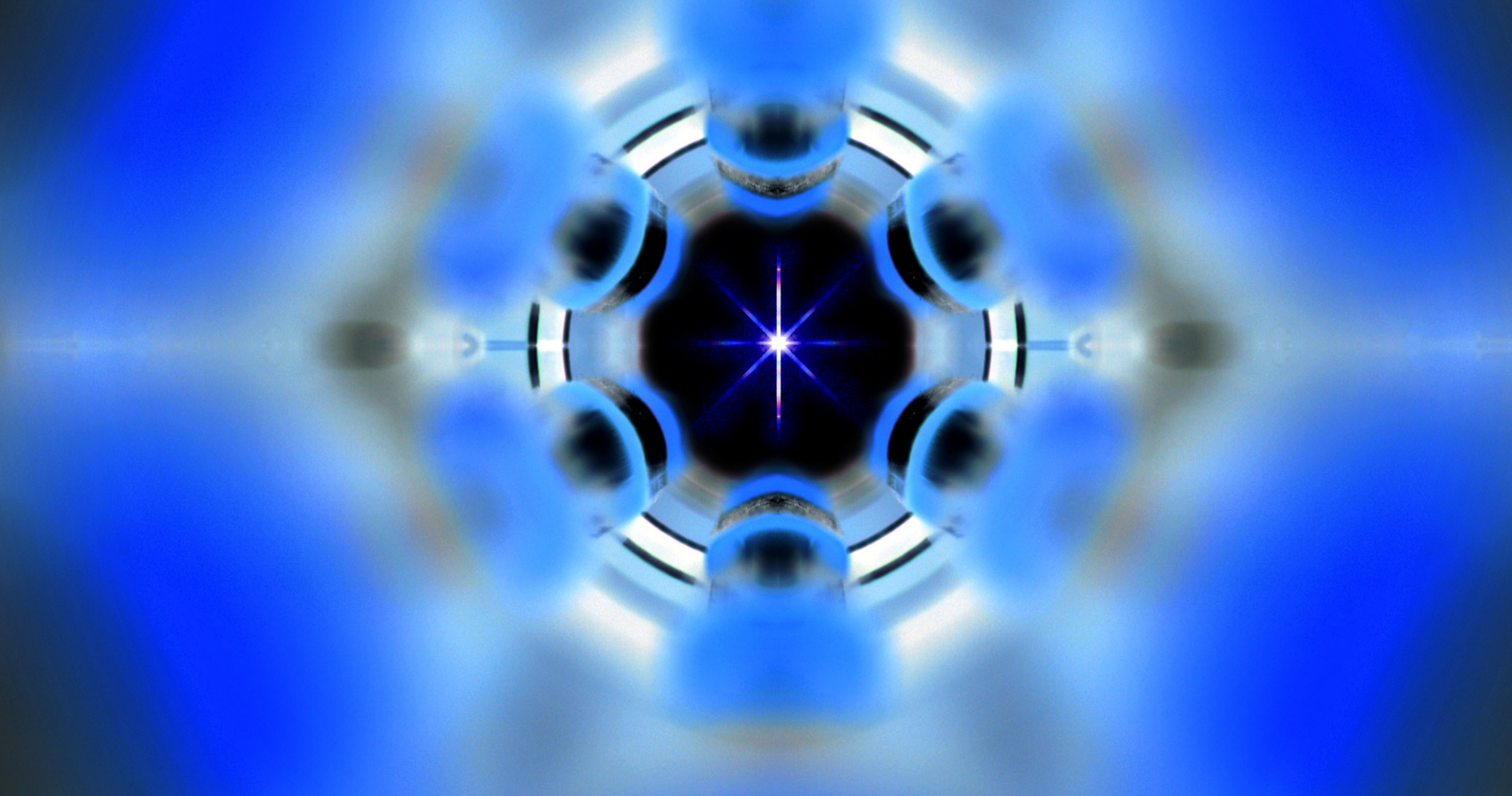
LAD 14
DIAMOND LIGHT SOURCE



Dave Bond

diamond

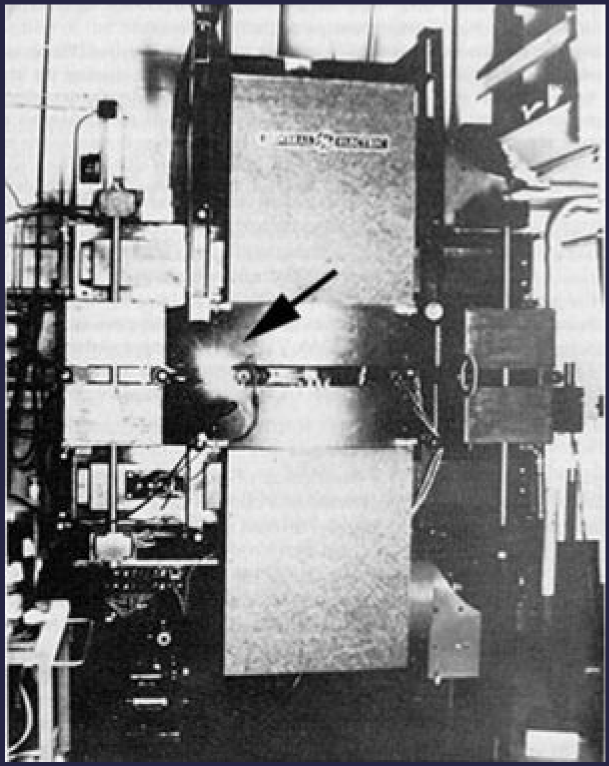
So, what do we actually do?

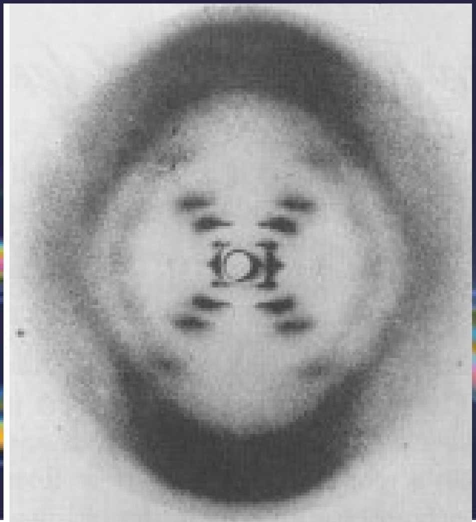
- The Diamond machine is a type of particle accelerator
- CERN = high energy particles smashed together and analyse the “crash”!
- Diamond = accelerate electrons to produce synchrotron light
- Use this light to study matter – like a “super microscope”

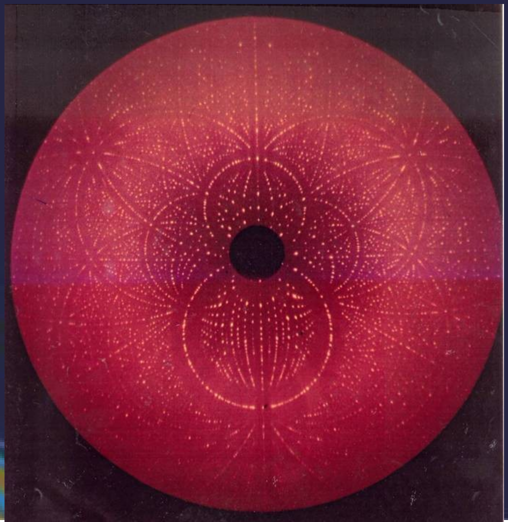
diamond

A Brief History of Synchrotrons

- Early particle accelerators produced "waste" in the form of light
- However X-rays were increasingly used as diagnostic tools, such as crystallography.
- 1953 – Structure of DNA realised using X-ray crystallography
- 1956 – first "parasitic" experiments carried out at a synchrotron







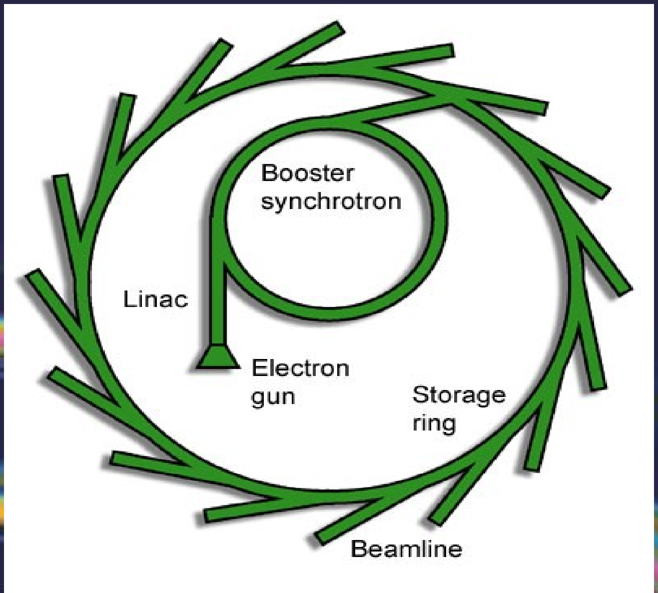
diamond

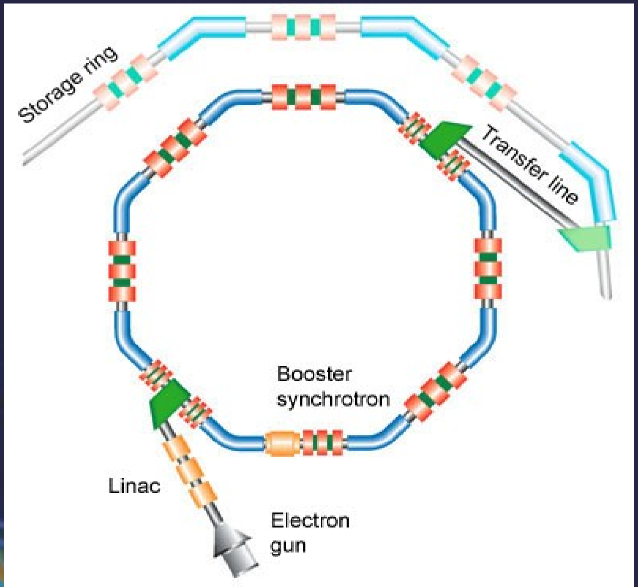
A Brief History of Synchrotrons

- 1960s – Synchrotron light used for atomic and molecular spectroscopy
- 1970s – increasing number of SR experiments, including crystallography
- 1980 - first dedicated synchrotron light source was built in 1980 in the UK - SRS
- Diamond has replaced the SRS which closed in August 2008

The Diamond machine

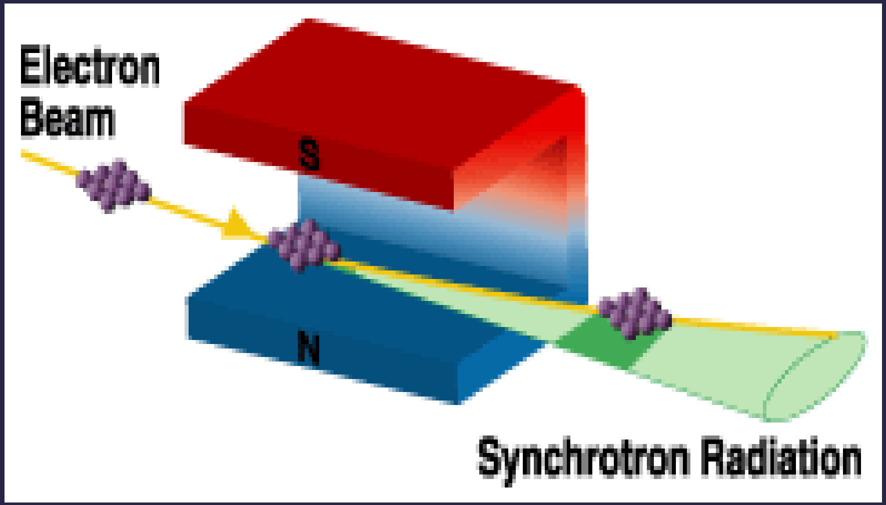
- **Three particle accelerators:**
- **Linear accelerator**
- **Booster Synchrotron**
- **Storage ring**
- **(48 straight sections angled together, 562m long)**

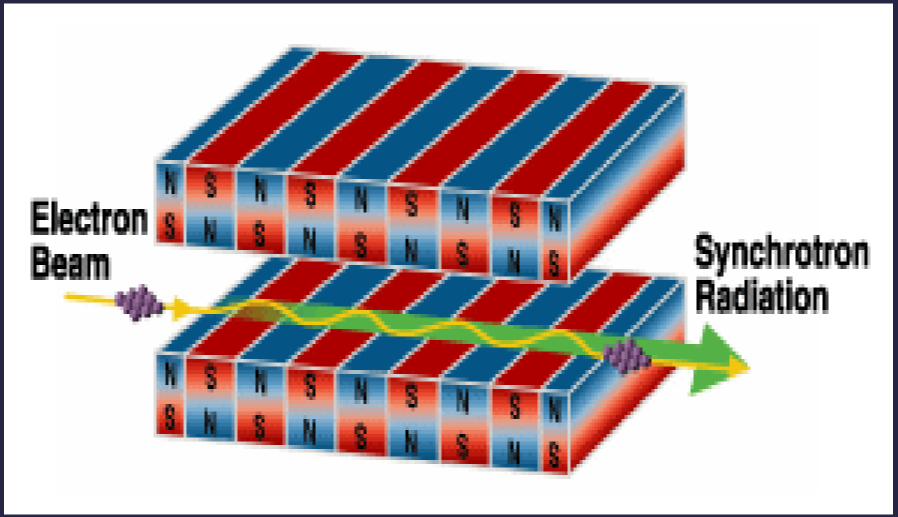




It’s all done with magnets

We use magnets to focus and direct the beam, and to make the light we use in experiments





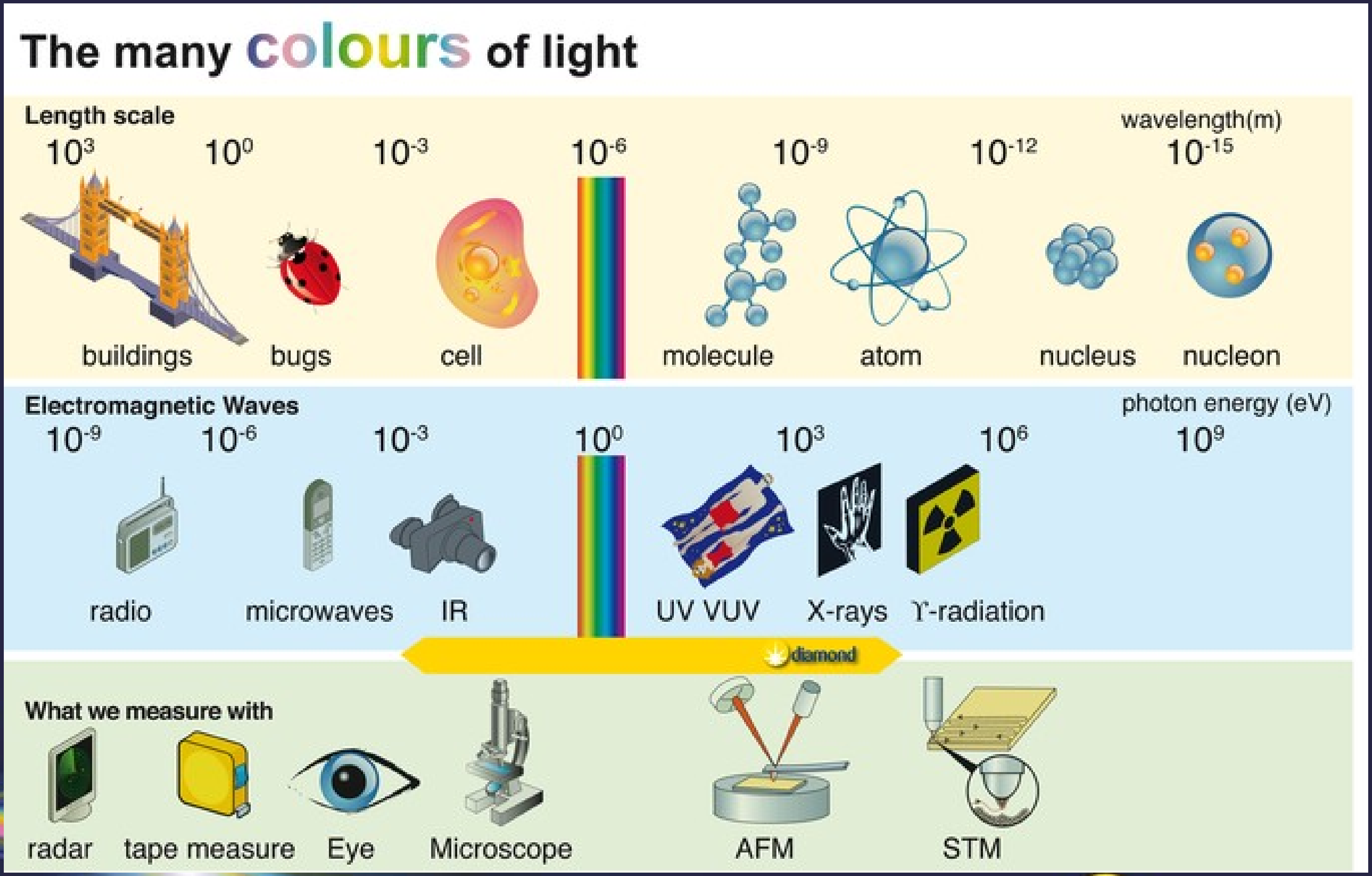
B = 1.4 T (20,000 x Earth’s magnetic field, or 100 x typical bar magnet).

24 straight sections in the storage ring. 22 straights available for insertion devices (IDs)

These enable us to produce X-rays that are high energy, more tightly focussed and tuneable.

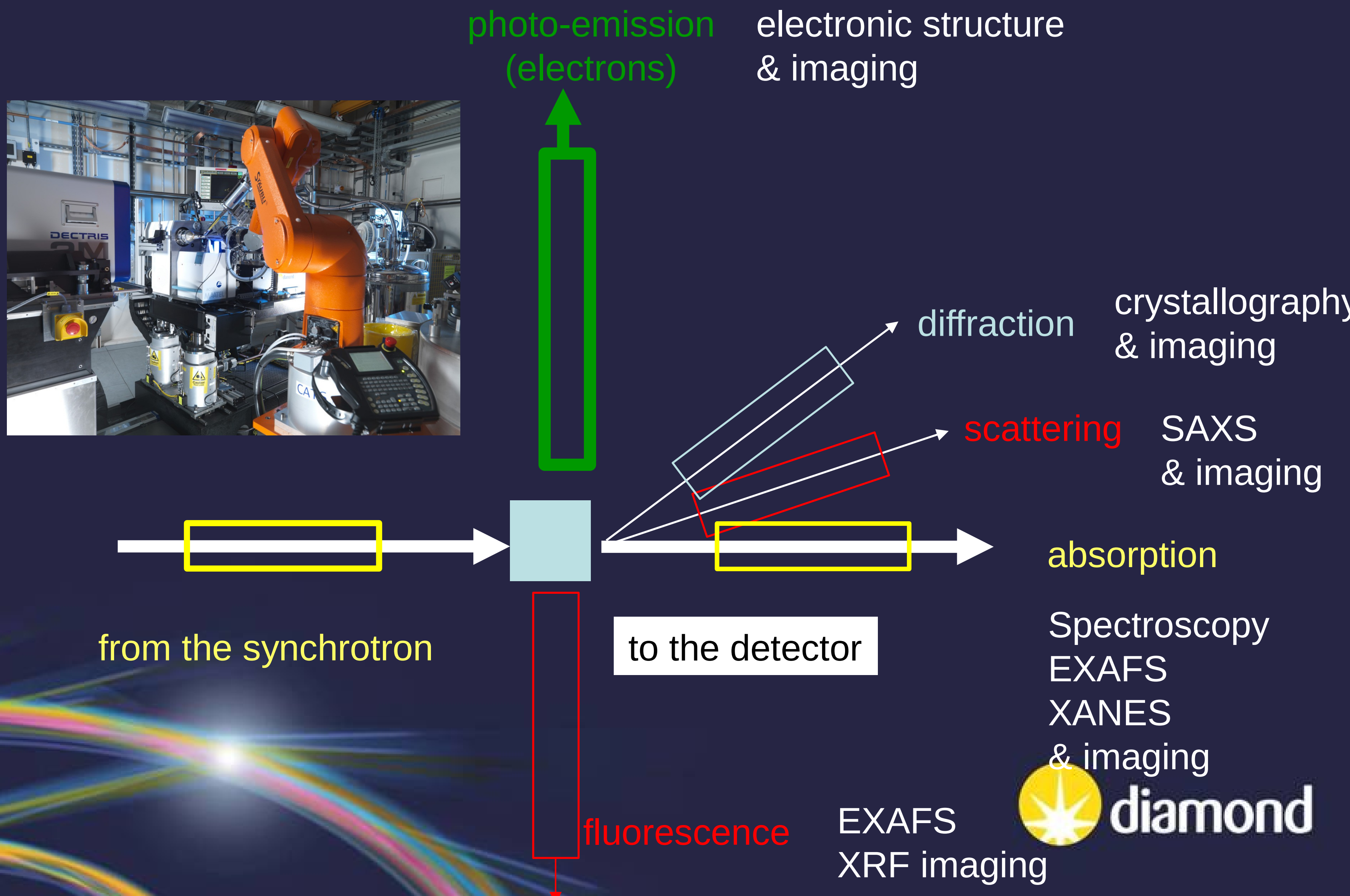
diamond

What is synchrotron light?



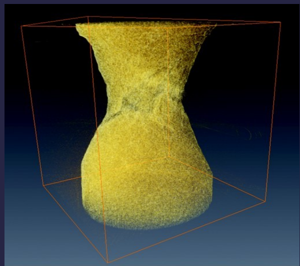
diamond

Types of experiment



diamond

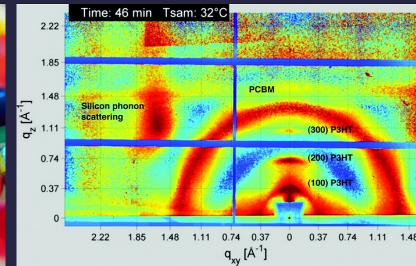
Research into .... and much more



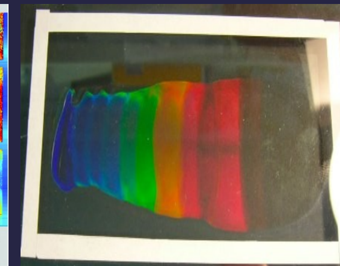
Casting aluminium

Understanding the corrosion process

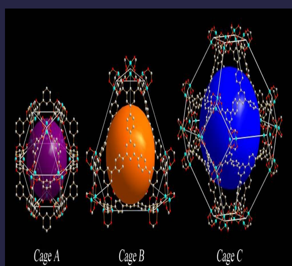
Pharmaceutical manufacture and processing



Organic photovoltaics

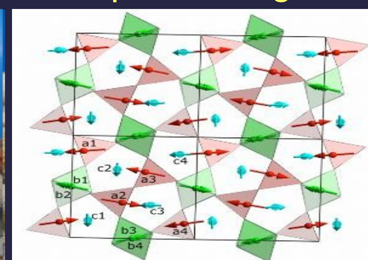


Tunable polymers

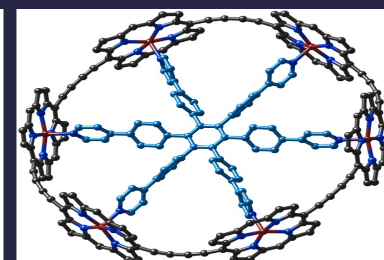


MOFs for hydrogen storage

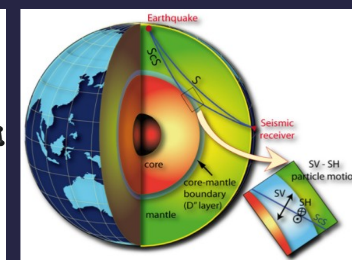
Bio-mimetics



Multiferroics – electronic storage and memory



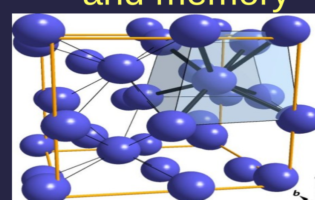
Harry's wheel – complex templates for new materials



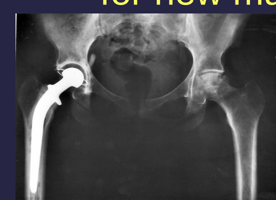
Earth science

Environmental science

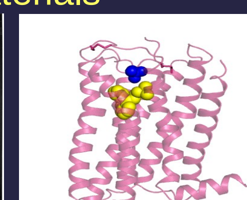
Cultural heritage and conservation



Fundamental research



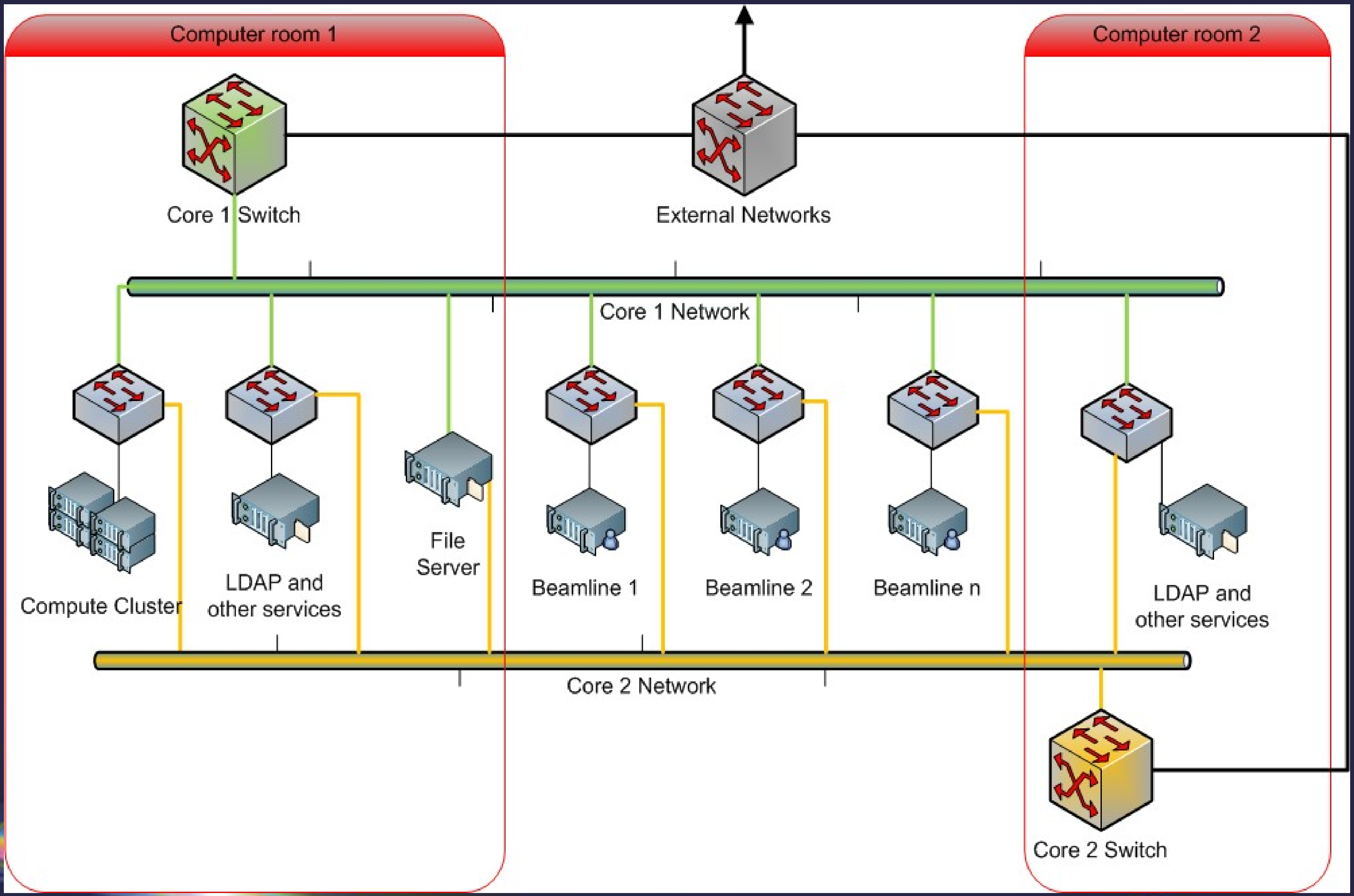
Understand rejection in hip implants



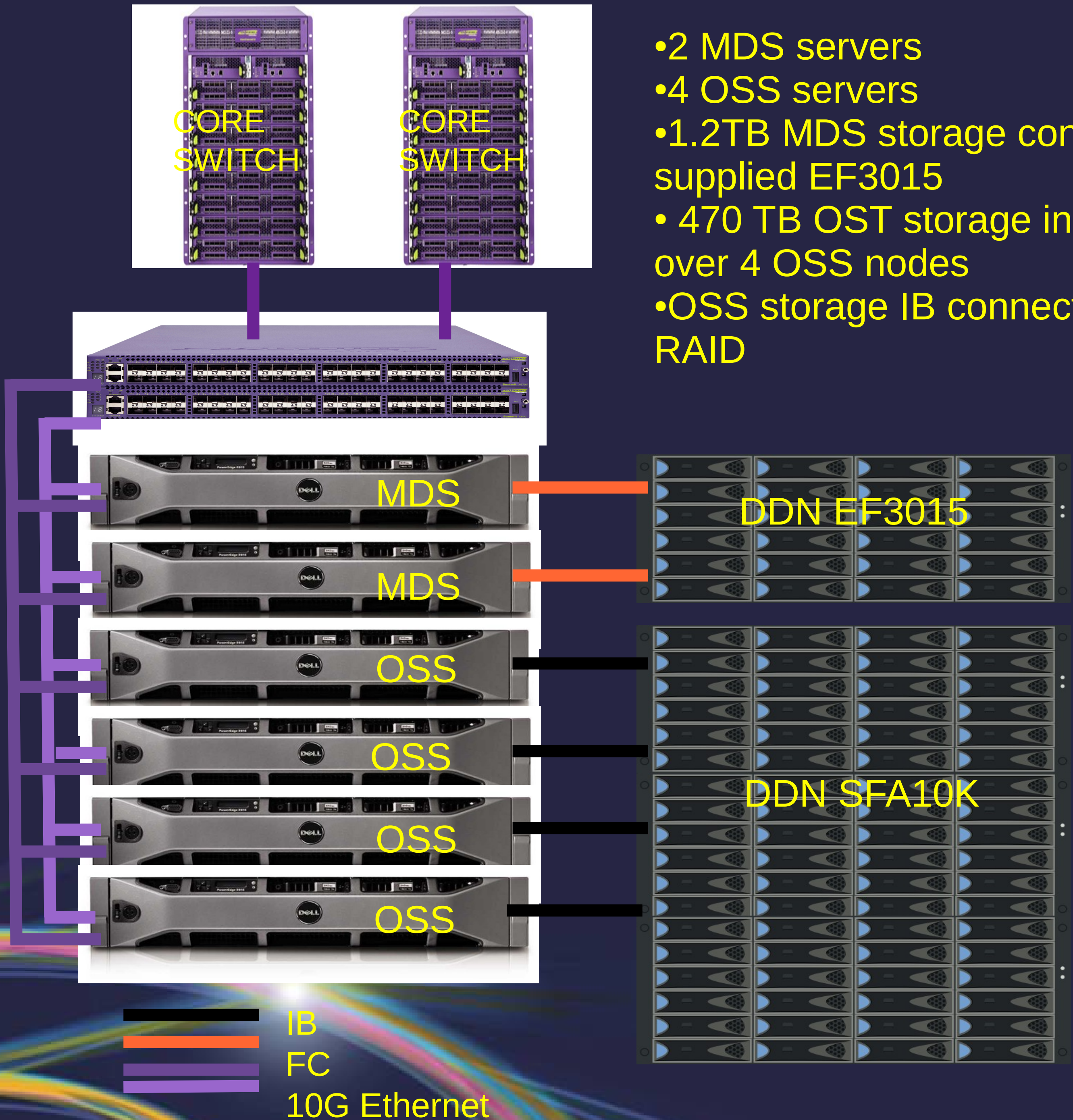
Structure of the Histamine H1 receptor

diamond

What lustre setup do we have?



- 2 MDS servers
- 4 OSS servers
- 1.2TB MDS storage connected via FC. DDN supplied EF3015
- 470 TB OST storage in 30 volumes spread over 4 OSS nodes
- OSS storage IB connected DDN SFA10K 8+2 RAID



How Lustre at Diamond has evolved....

- The first production systems were Lustre 1.6
  - 10G Ethernet connected
  - DDN 9900
- The second production file system was Lustre 1.8
  - 10G Ethernet connected
  - DDN SFA10K
- Both file systems have been upgraded to at least a couple of versions of Lustre 1.8
- Most recently upgraded to Lustre 2.5 with the second of the two production file systems now is IB capable

What other filesystems do we use?

XFS Filesystem

1.1PB of available storage used for low data rate beam lines and on-line archiving

NetApp™

150TB of available storage used for data not collected at Diamond, home areas, databases, virtual machines

IBM® GPFS

877TB of available storage used for high data rate beam lines

lustre®

814TB of available storage used for high data rate beam lines

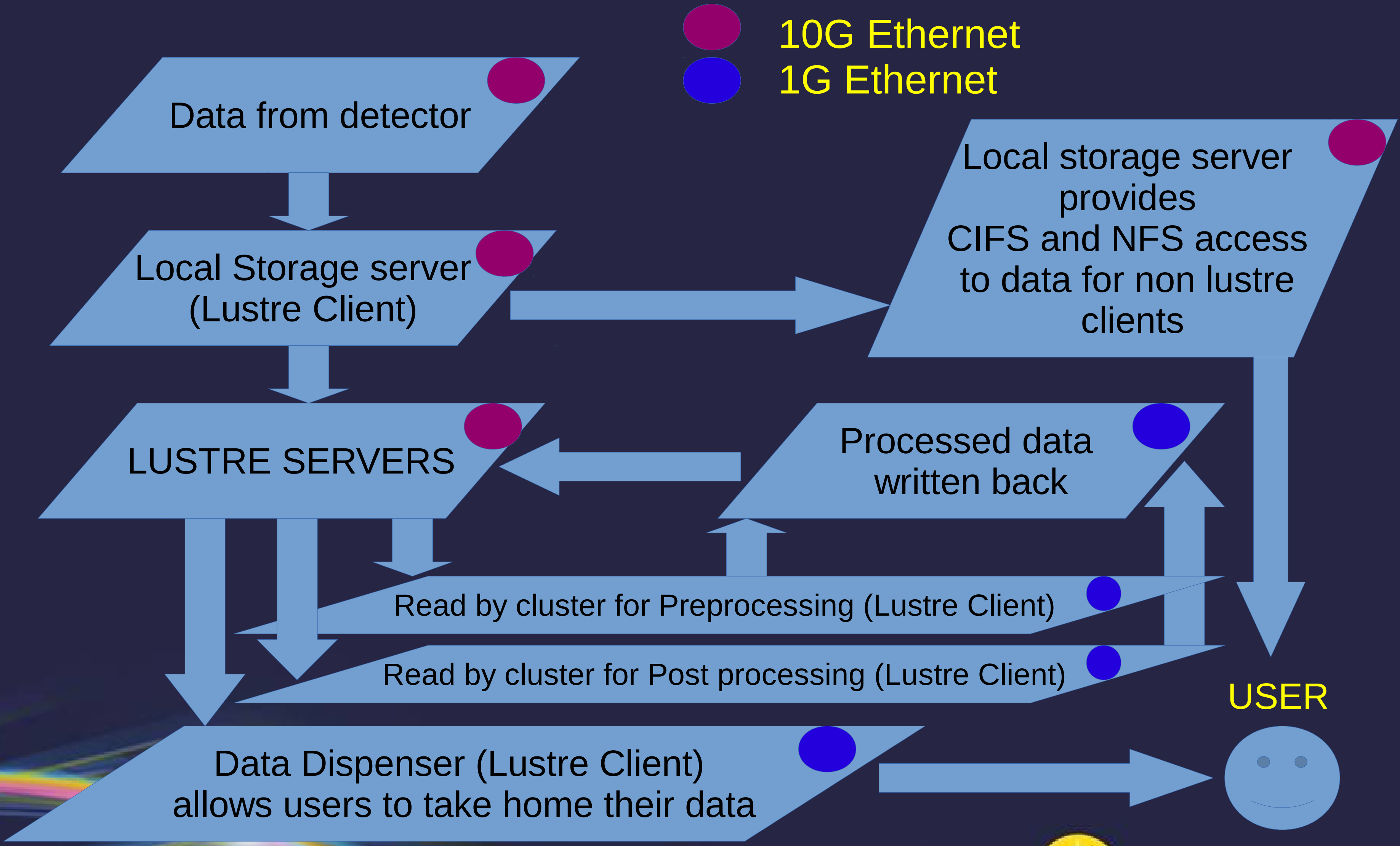
diamond

Why do we have GPFS as well as Lustre?

- Single stream performance is better for GPFS, we were aiming for 900Mbs sequential read and write. This was a hard requirement for two beam lines.

- Though we achieved this we have had issues since.
  - Mixing 1G and 10G crippled performance. When writing over 10G and reading over 1G. There was large performance penalties.
  - Upgrading to IB on the NSD servers and communicating to the cluster over IB resolved this
  - Native CIFS access is difficult to manage and deploy, we had regular issues where it would drop out of the GPFS cluster.
    - We are currently using SAMBA for CIFS access and RSYNC as a data mover.
    - In GPFS 3.5 the CIFS issues are believed to have been improved. We are awaiting the DDN release.

Typical Data flow to HPC file systems



diamond

Upgrade to 2.5 and what we hope to gain?

We currently have strict archiving rules as we aim to only store 6 months of data on disk, before access is only from tape or for some data archive disk.

HSM -> Change Logs

This will aid us with robin hood and the ability to gather archiving lists easier.
HSM may in the future automate moving files to low data rate storage.

diamond

Upgrade to IB and what we hope to gain?

IB

After a recent upgrade to IB on our GPFS system this was the logical next step to upgrade Lustre to provide high bandwidth low latency connectivity between the storage and the cluster

6036 Spine Switch 6025 Leaf Switch

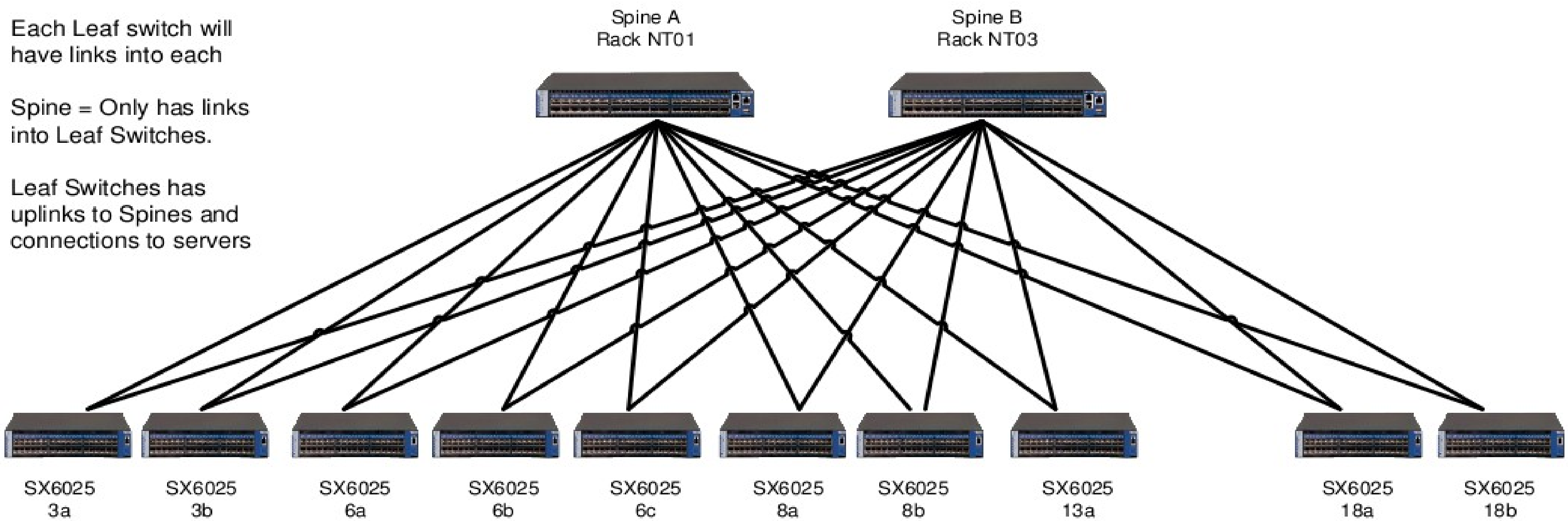
Current Diamond Infiniband layout - 09/09/2014

Information:
Spines are Mellanox SX6036
Leafs are Mellanox SX6025

Each Leaf switch will have links into each

Spine = Only has links into Leaf Switches.

Leaf Switches has uplinks to Spines and connections to servers



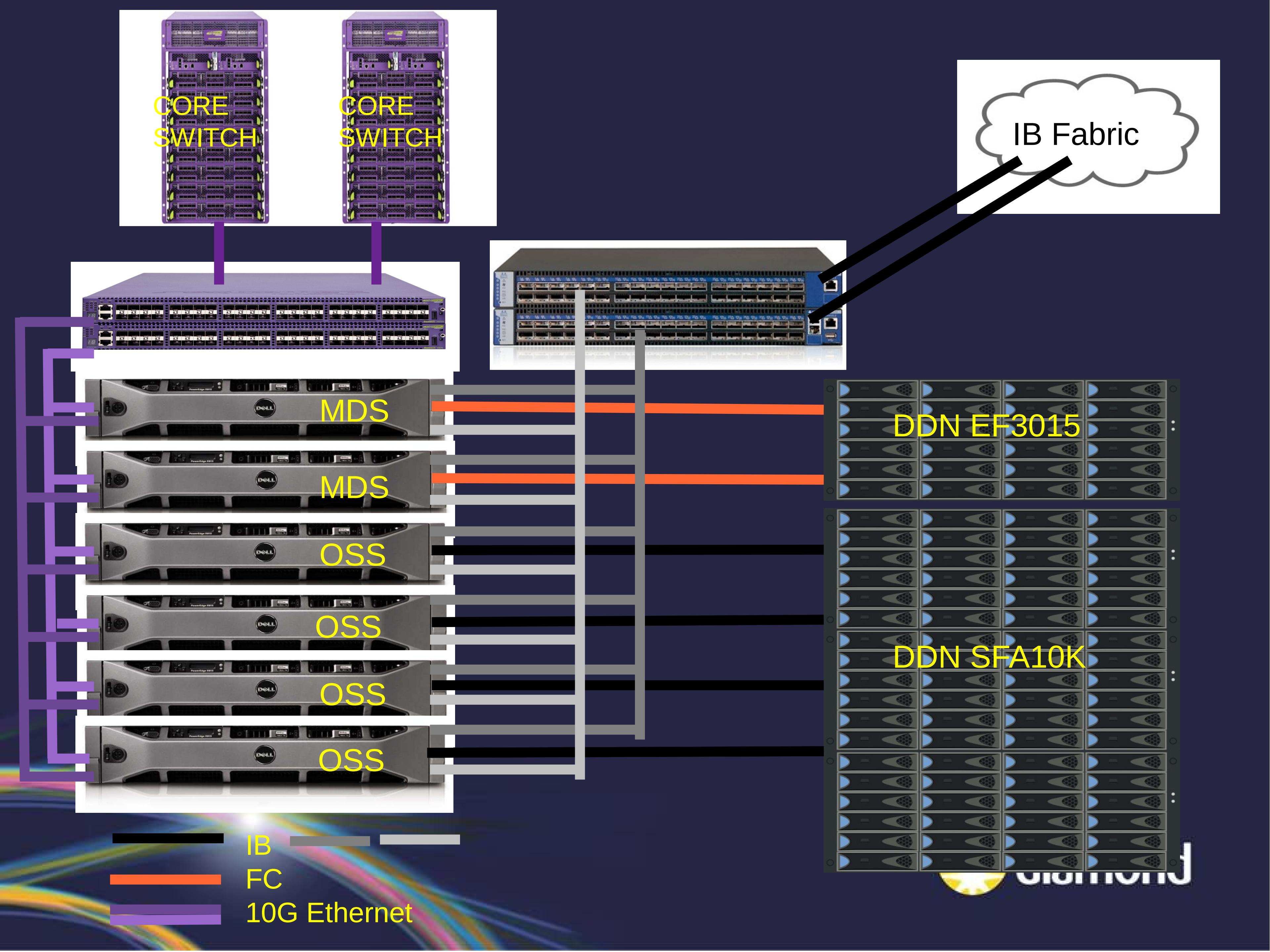
Rack 3
Current Setup is 2 Switches
3a has 5x 56gbps to each spine switch.
3b has 5x 56gbps to each spine switch.

Rack 6
Current Setup is 3 Switches
6a has 2 x 56gbps uplinks to each Spine switch.
6b has 2 x 56gbps uplinks to each Spine switch.
6c has 3 x 56gbps uplinks to one Spine and 4 to the other spine switch. (There are also 6 spine connection waiting to be connected after server cable moves)

Rack 8
Current setup is 2 Switches.
8a has 3 x 56gbps uplinks to each Spine switch.
8b has 3 x 56gbps uplinks to each Spine switch.

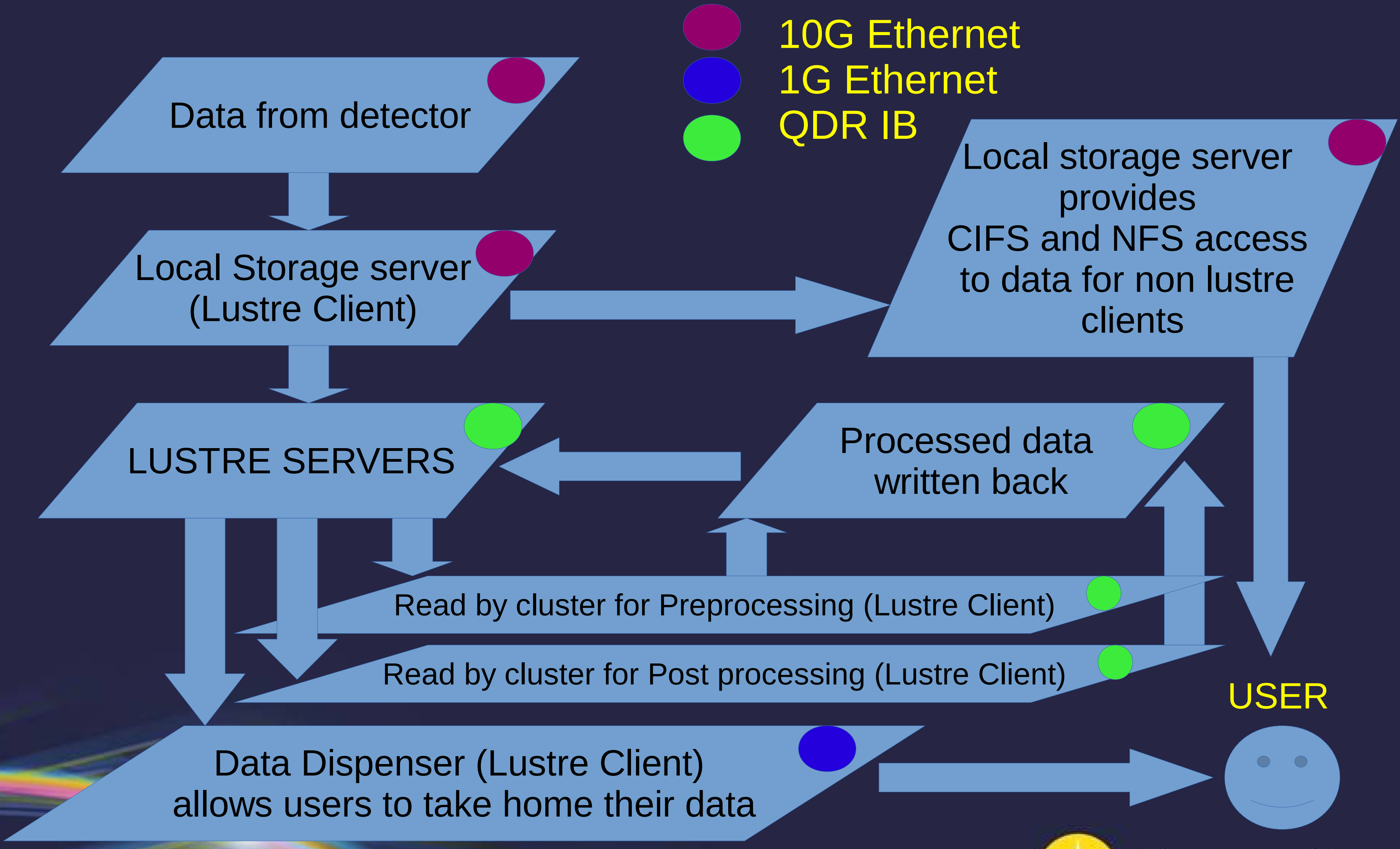
Rack 13
Current setup is 1 switch
13a has 4 x 56gbps uplinks to each Spine switch.
Rack 13 will end up with a second leaf switch.

Rack 18
Current setup is 2 switches
18a has 3 x 56gbps uplinks to each Spine switch.
18b has 2 x 56gbps uplinks to one spine switch and 3 to the other spine switch.



CORE SWITCH
CORE SWITCH
IB Fabric
MDS
MDS
OSS
OSS
OSS
OSS
DDN EF3015
DDN SFA10K
IB
FC
10G Ethernet

Typical Data flow to HPC file systems

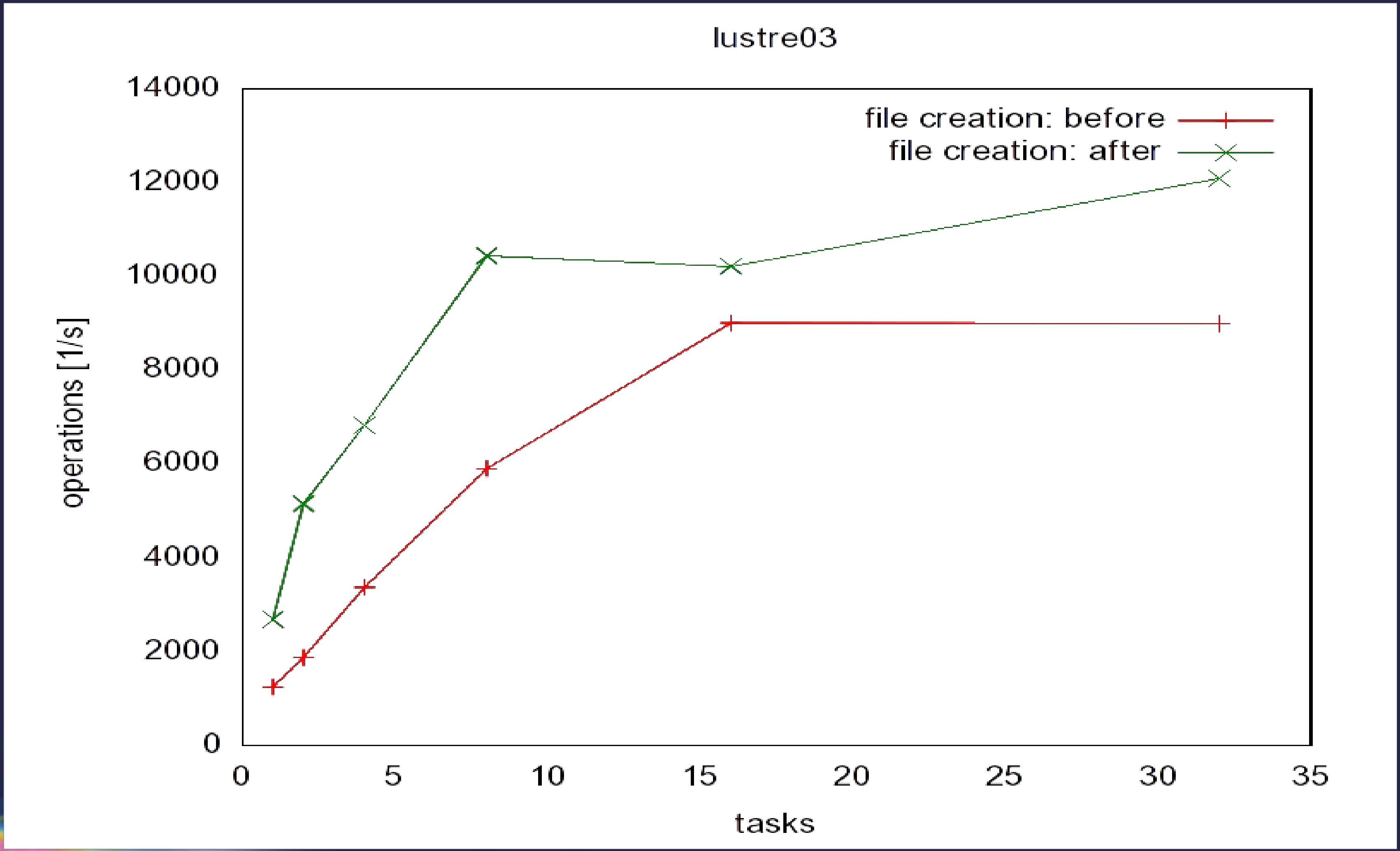


diamond

How we went about testing...

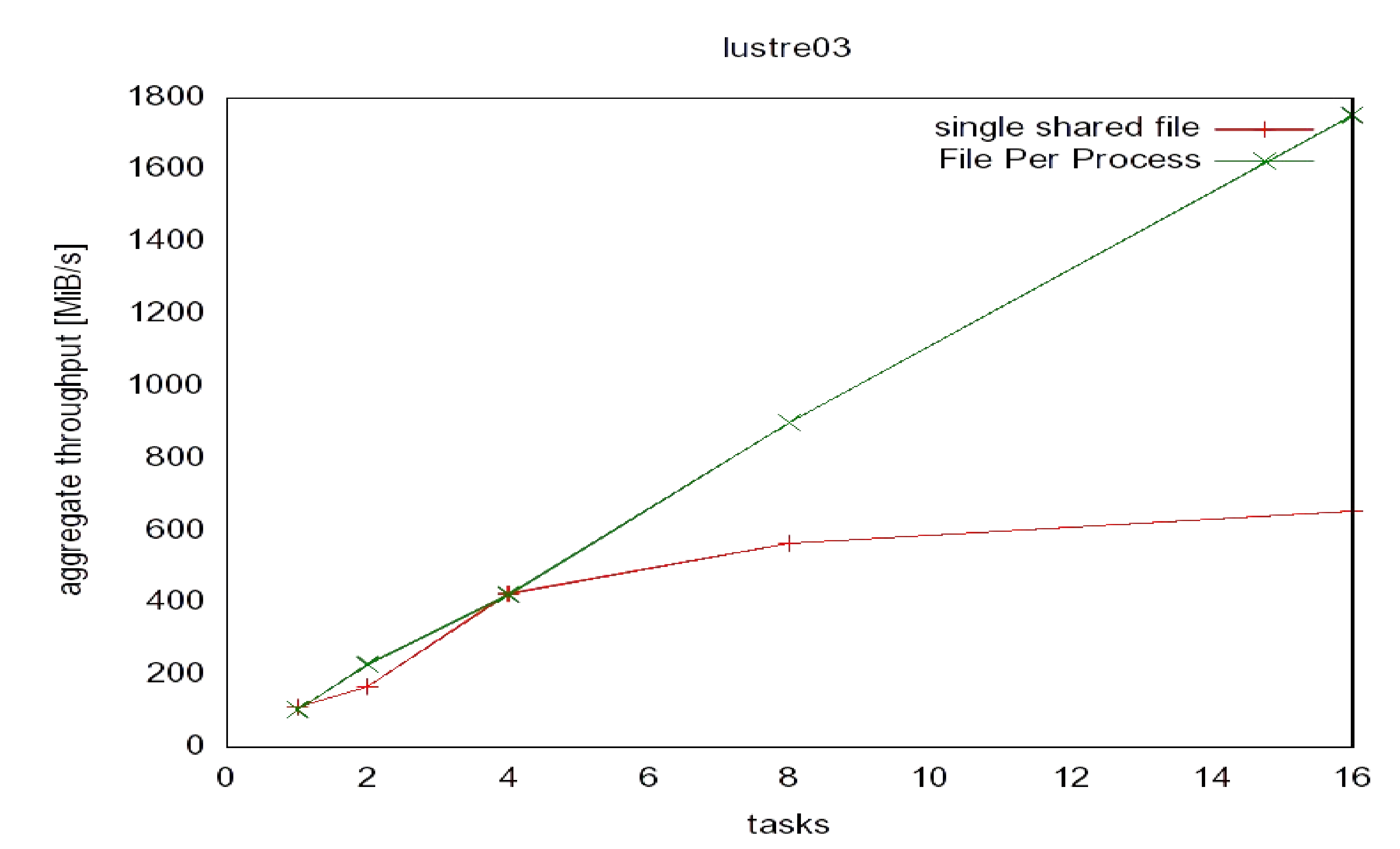
- MDTEST
- ${MPI}/bin/mpirun -mca btl self,tcp,sm --hostfile ${UNIQHOSTS} -np ${jobs} ~bnh65367/code/mdtest/mdtest -I 10 -z 5 -b 5 -i 5 -u -d ${TESTDIR}
  - Run with our cluster one process per host
  - -I 10 -> 10 items per directory
  - -z 5 -> Tree depth of 5
  - -b 5 -> branching factor of 5
  - -u -> create unique working directory for each task
  - -d -> working directory
- IOR
  $MPI/bin/mpirun -mca btl self,tcp,sm --hostfile ${HOSTFILE} -np ${jobs} /home/bnh65367/code/ior/src/ior -o ${TESTDIR}/ior_dat -w -r -k -t1m -b ${BLOCKSIZE} -i 5 -e -a POSIX
  - -w -> write file
  - -r -> read existing file
  - -k -> keep the file do not remove it
  - -t -> maximum transfer size 1m
  - -b -> block size
  - -i -> repititions
  - -e preform fssync
  - -a -> use posix
- IOZONE
  iozone.x86_64 -i 0 -i 1 -s $2 -r 4k -t $3 -+m ~/dls-science-user-area/iozone-host-list

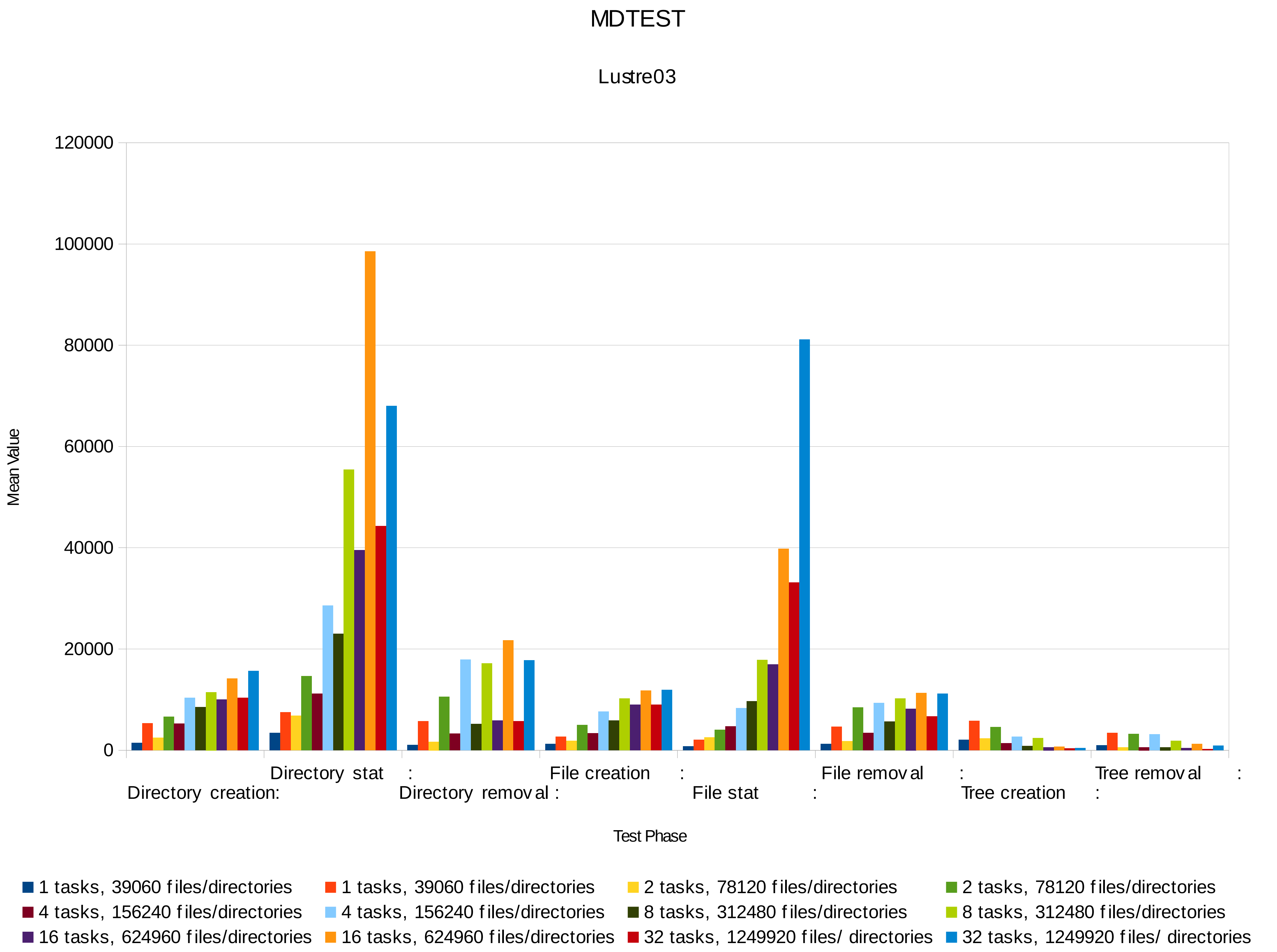
MD Test



lustre03
14000
12000
10000
8000
6000
4000
2000
0
0
5
10
15
20
25
30
35
operations [1/s]
tasks
file creation: before
file creation: after

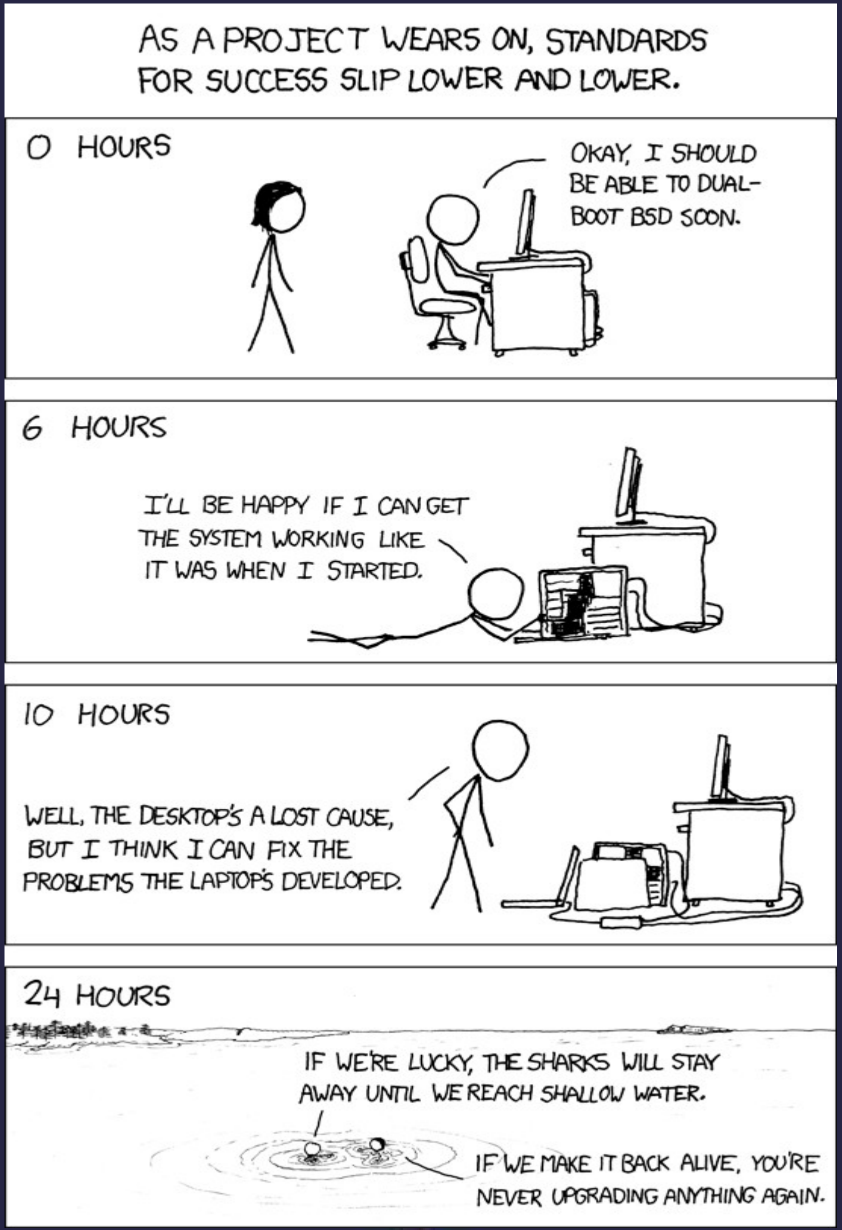
IOR After code upgrade




MDTEST
Lustre03
120000
100000
80000
60000
40000
20000
0
Mean Value
Directory creation:
Directory stat :
Directory removal :
File creation :
File stat :
File removal :
Tree creation :
Tree removal :
Test Phase
1 tasks, 39060 files/directories
1 tasks, 39060 files/directories
2 tasks, 78120 files/directories
2 tasks, 78120 files/directories
4 tasks, 156240 files/directories
4 tasks, 156240 files/directories
8 tasks, 312480 files/directories
8 tasks, 312480 files/directories
16 tasks, 624960 files/directories
16 tasks, 624960 files/directories
32 tasks, 1249920 files/ directories
32 tasks, 1249920 files/ directories

This work only has been completed in the last few weeks, testing is still ongoing.

The first users of the upgraded Lustre system will be using it today.



Credit xkcd.com

diamond

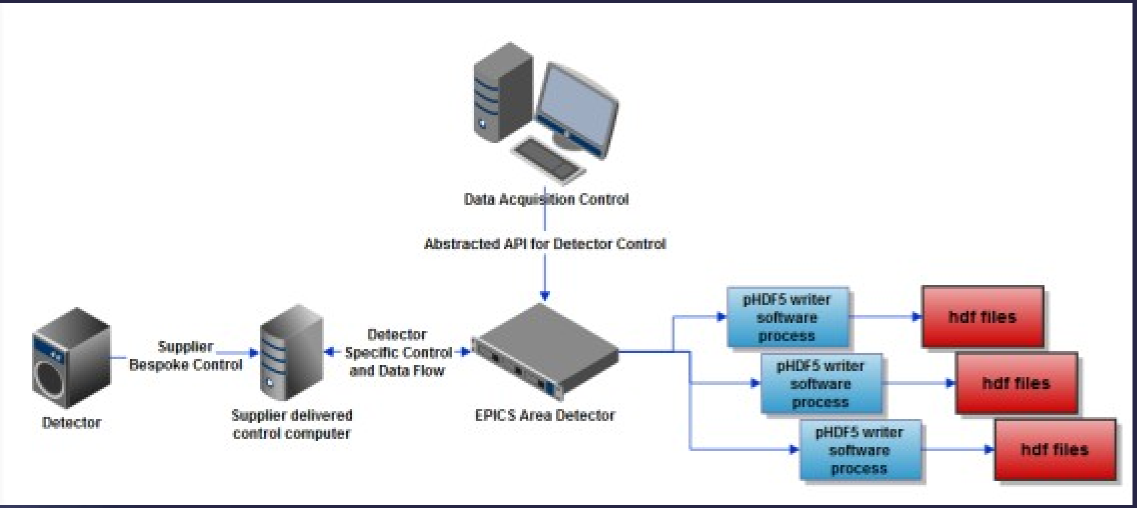
Future HPC filesystems at Diamond?

With new high data rate detectors already being tested and the new national electron microscopy facility being built to hold 4 electron microscopes

The planning for the next HPC file system has begun!

Examples of detectors at DLS

Specification	PCO 4000 PCO 4000 and PCO Edge are high speed cameras developed and supplied by PCO AG,Donaupark 11, 93309 Kelheim, Germany	PCO-Edge	Pilatus 6M Pilatus 6M is a very high capability detector developed and supplied by DECTRIS Ltd. Neuenhoferstrasse 107, 5400 Baden, Switzerland	Excalibur Excalibur is a detector development collaboration between the Science and Technology Facilities Council and Diamond Light Source - Journal of Physics: Conference Series Volume 425 Part 6 J Marchal *et al* 2013 *J. Phys.: Conf. Ser.* **425** 062003 doi:10.1088/1742-6596/425/6/062003	Percival PERCIVAL (Pixelated Energy Resolving CMOS Imager, Versatile and Large) is an Ongoing development project between DESY and RAL / STFC
Frame	2D	2D	2D	2 - 3D	2D
Scan Size	1D	1D	1 - 3D	1 - 2D	1D
Frame rate	5Hz	100Hz	100Hz	100Hz	120Hz
Data Rate	100MB/s	700MB/s	~640MB/s	~600MB/s	~5-6 GB/sec
Status	Complete	In development	Complete	Commissioning	In development



At the moment the contenders are GPFS and Lustre
Lustre is preferred because of the seemingly high sensitivity of GPFS for issues in a distributed environment

Requirements

- 6 machines writing into the same file at 10GB/s, just for one beam line
- Saturating 10GB/s with a single stream or an IB connection .... even better
- A nice to have would be a native lustre client for Windows

PERCIVAL

diamond

diamond