Diamond Light Source site report

Tina Friedrich

Diamond Light Source Ltd

03 May 2011

Outline

What is Diamond Light Source?

Science Computing

Setup / Installation / Management

- Facilities
- Science Server Room
- Science Network
- Science Data Storage
- Science computing Resources

Current work / Future plans / Outlook

Many thanks to my colleagues – Greg Matthews, Frederik Ferner, Max von Seibold and Nick Rees – for their contributions to this talk.

Diamond Light Source

Diamond Light Source is the UK’s national synchrotron facility. It is located at the Harwell Science and Innovation Campus in Oxfordshire.

- third generation light source
- 561.6 m circumference storage ring; energy 3GeV
- first users 2007
- three build out phases:
  - Phase I: 7 beamlines
  - Phase II: 15 beamlines
  - Phase III: 10 beamlines

Science Computing

Science Computing provides computing infrastructure for beamlines (data storage, compute clusters, local hardware, OS installation and configuration). We also provides standard services like

- DNS and DHCP
- LDAP directory services and Active Directory integration
- version control and issue tracking
- central home file system(s), software and package repositories
- remote access and remote beamline control
- print server, monitoring, . . .

We currently look after ~250 servers and ~250 workstations.

Setup / Installation / Management

We rely heavily on central provisioning and management. Servers and workstations are considered dispensable.

- operating system is Red Hat Enterprise (currently version 5)
- central home directories
- no 'local' modifications on servers or workstations
- all machines installed from single network boot (no manual intervention)
- all machines configured from central configuration control (cfengine, currently version 2)
- using kickstart for installation, but all configuration is done via cfengine
- all changes tracked in version control system

Changes to systems, as well as software upgrades, are rolled out sequentially after a test and approval phase.

Science Computing facilities

Diamond originally had no provision for central science computing (all computing to be local to the beamlines). We started to develop it in 2007-2008, with a major development in 2008 consisting of:

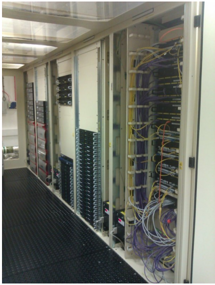
- a resilient high density computer room
- a resilient network
- a central, general use compute cluster
- a central file system (mainly data)

Science Server Room

We built a computer room for science computing in 2008, focus on resiliency and high density

- up to ~20 kW/rack.
- two separate feeds from separate sub-stations, one of which is UPS and generator backed up
- up to 320 kW redundant power total
- up to 320 kW cooling water
- primary cooling from site chilled water
- 220 kW standby chiller (with fast automatic switchover)
- initial build was to 10 racks (~50% capacity)
- currently expanding to 22 racks

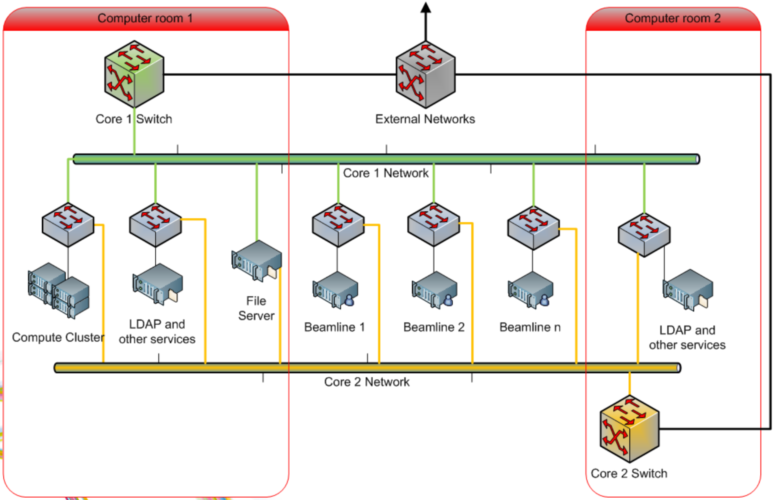
System proved its worth a number of times.



Science Network

- two core networks, using core switches from different vendors, located in different computer rooms
- each beamline is a separate subnet connected to both core networks
- 1Gbit or 10Gbit Ethernet uplinks to core — i.e. beamlines have 2 Gbit or 20 Gbit bandwidth available
- routing and resiliency managed by OSPF and ECMP
- one subnet per computing rack, similar to beamlines
- some beamlines now have some 10 Gbit clients

Science Network Diagram



Science Data Storage

In the original design, beamlines had local storage systems.

- not scalable for performance or maintenance
- still in use on some legacy / low data rate beamlines

For newer beamlines, we have a central Lustre file system. Some original beamlines moved to this (increasing storage demands).

- mix of DDN (OSTs) and Dell (servers, MDT, MGT) hardware
- now 400TB raw (~300TB usable); >50% full
- connected to core network via 10Gbit Ethernet (multiple links)
- aggregate write speed ~3.5GB/s
- good read/write performance for Lustre clients; some issues when re-exporting (NFS/CIFS)

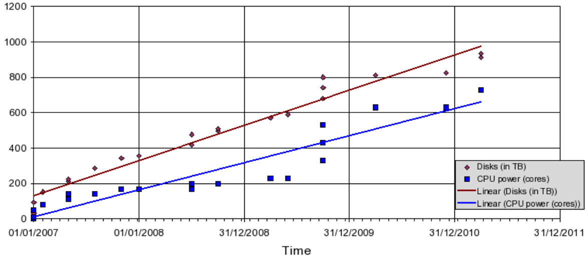
Science Computing Resources

We provide a 'general purpose' compute cluster.

- available to all users
- also used for automated fast data processing on beamlines
- currently operational cluster:
  - 80 Viglen/Supermicro nodes, 640 cores (X7DWT & X8DTT-IBX boards, Intel Xeon E5420 & E5520 CPUs)
  - ~27 IBM x3455 nodes with AMD Opteron CPUs (~108 cores)
  - 4 Tesla GPU 1U units (16 GPUs, 240 GPU cores each)
- scheduler is Sun Grid Engine (SGE 6.2u4)
- simple setup (currently no sharing policy, three queues, handling prioritisation through queue suspension)
- very little requirement for low latency interconnect

Computing and Storage requirements

Diamond storage and computing requirements



- file system usage growing about 1% every 3 days during a run
- we have started to implement a data management procedure

Outlook

- recently purchased
  - an additional 600 TB (raw) DDN SFA 10000 system
  - another 40 compute nodes — Viglen HX425T^{2}i Quad HPC nodes (Supermicro X8DTT-F boards), Intel Xeon X5650 CPUs
- upgrade of core network proposed in 2012/13
  - new core switches, with some 40/100 Gbit beamline uplinks
- investigate and purchase new storage facilities
  - commodity, low data rate beamlines, . . .
- GPU cluster upgrade planned for this year
- investigate and implement replacement monitoring solution
- upgrade to Red Hat 6
- upgrade to Cfengine 3

and

- now it is approved, need to provision for Phase III

diamond

Thank You!