Oxford Site Update
HEPiX

Sean Brisbane
Tier 3 Linux System Administrator

March 2015

Oxford Physics

Contents

The Oxford Physics department
Working environments for HEP
The Oxford Tier 2 and 3
Experience with Lustre 2.5
Experience with puppet
University upgrades

Oxford – March 2015

Oxford Particle Physics - Overview

- Oxford University has one of the largest Physics Departments in the UK
  - ~450 staff and 1100 students.
- Particle Physics sub-department is the largest in the UK
- Department now supports Windows, Ubuntu, Mac on the desktop across the department
  - Ubuntu desktops and departmental Linux infrastructure sharing some staff with HEP
  - Cobbler used to install base system, Cfengine2 used for package control and configuration.
- Two SL6 computational clusters for the HEP
  - Grid Cluster part of the SouthGrid Tier-2
  - Local Cluster (AKA Tier-3)
- Cobbler and Puppet for all SL6 machines
  - Additional SL6 desktops (HEP only) off the back of puppet system
  - Significant investment in a clean module, node and configuration hierarchy

Oxford Cluster's Hardware

- Most of the storage now Dell R510 / R720xd
  - TB 1,300 (T2) and 700 TB (T3)
  - Most recently used 12x SATA not 12x SAS this time. Unlikely to do this in the future.
  - SAS becoming the default and support for SATA costs extra.
  - Disk I/O on T3 benefits from SAS
  - Raid cards ?possibly? rejecting SATA
- Most of the compute is supermicro twin squared
  - HS06 16,800 T2 and 7,200 T3
- Most T3 infrastructure nodes VMWare and most T2 OVirt

Oxford Physics

Oxford's Grid Cluster

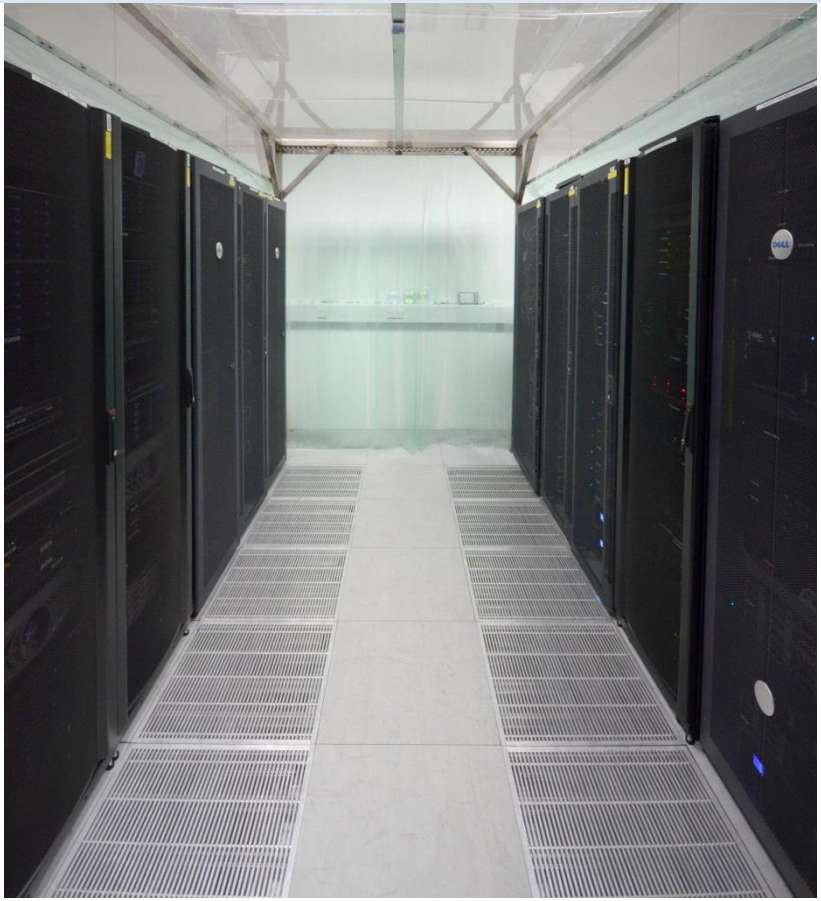
- Capacity:- 16,800 HS06, 1,300TB

Grid Cluster

- Two thirds HT Condor behind ARC CE.
  - Soft cgroup limit on CPU and memory
- Remaining third running legacy torque/maui driven by CREAM CE.
- Storage is DPM
- Small test beds of alternative worker technologies
  - VAC
  - Open Stack
- Oxford runs the Nagios based WLCG monitoring for the UK
  - These include the Nagios server itself, and support nodes for it, SE, MyProxy and WMS/LB
  - Multi VO Nagios Monitoring added two years ago.
- Take a leading role in IPv6 Testing
  - Our campus network is barely IPv6 capable but will be replaced over the coming year

Local Computer room showing PP cluster & cold aisle containment

- Very similar h/w to the Grid Cluster.
- Same Cobbler and Puppet management setup.
- Lustre used for larger groups
- Capacity:- 7192 HS06, 716TB



Oxford Physics

Local PP cluster

- Capacity:- 7200 HS06, 720TB
  - LHCb and Atlas use Lustre (~550TB)
  - NFS (cdf, t2k, sno+, dark matter, accelerator science, mars)
- SL6 + torque
  - Small <5% SL5 (or ubuntu) requirement met with a chroot on SL6
  - Need to translate maui configs blocking move to HT condor
- All remote nodes
  - Connect with ssh + X windows or xrdp
  - T3 disks available to department over SMB, NFSv4 (krb5) and webdav interfaces
- CVMFS stratum zero for local software
  - Allows external sharing to university and beyond
  - Top-level separation between software for generic Linux, Ubuntu and SL versions

Lustre 2.5.3

Our Lustre experience in summary

- Lustre 1.8 worked really well.
- Lustre 2.1 we had no problems but didn't try to decommission or move any storage nodes
- Lustre 2.5 has been more problematic than anticipated

Lustre 2.5

- Our lustre 2.5 life history
  - Start with a clean lustre 2.5 filesystem (ldiskfs)
  - Retired older storage server/targets using lfs_migrate on 2.5 file-system
  - Migrated metadata server (mds) from old to new hardware within the new 2.5 file system using tar/untar
    - This process has now disappeared from the l2.x manual
- Some number of annoying issues
  - Documentation outdated
  - Some operations no longer supported as they were in lustre 1.8
  - File system is overall slower
- Some number of more serious bugs resulting in corruption and loss of file-system parallelism

A selection of our lustre issues

The following are traced to a Lustre bug (we think)

1. lfs find misbehaves for some OST names if OST name rather than index used [1]
   - Just use find by index
2. Large directory corruption not fully categorized on mds [2]
   - ".." (dot dot) Entry in wrong place.
   - Note for those running lustre 2.5.1 there is a different related bug.
   - Htree conversion messed up for large directories
3. Cannot free space on deactivated OST
   - Just don't deactivate when freeing space

[1] possibly LU-1738 [2] LU5626, [3] LU-4825

Not understood

- OST round robin and space-aware placements stopped working (not reported)
  - Assume decommissioning OST triggers the edge cases?
  - Makes us want [3] in production - for manual rebalance
  - Accumulation of files on a few OSTs

All compounded by out of date documentation

Integrate with university cluster?

Becoming more attractive for local and grid use

- Recent upgrade to 6000 cores for university RHEL6 cluster – free at point of use
- Campus network upgrade (~6 months) may give us a 10G link in to the service

How to configure working environment to reduce barrier to entry for users?

- With cvmfs
  - Challenges getting HEP software installed (cvmfs + fuse) as we scale up or try to run grid jobs
  - Per-user self configuration with parrot works at low scale
- Data access
  - HPC facility panasas storage is cluster-local and limited to 5TB/group
  - Looking at Xrootd + Kerberos / X509 to our T3 disks / the grid
- Some challenges reconciling our change management policies/expectations

Current status

- Local HEP users offload bulk, standalone, low data rate jobs to this cluster
- Keep other jobs on T3

hysics

Summary, Questions?

- Challenges integrating multiple client OS
- Two solid computer rooms
- Involvement in many grid development projects
- Issues with Lustre 2.5
- Puppet is powerful, but needs an investment of time to get right
- Involving University HPC clusters

Oxford Physics

Additional Material

Puppet

- Separate cfengine2 each for SL<=5 & Ubuntu >=10.04
- SL6 straight to puppet
- A lot of effort has gone into our puppet design
- Puppet is a framework from which to build a configuration management system
- Collaboration on shared puppet modules is fantastic but not the whole story
- Categorize your site and build a model (hierarchy)
- When all this is done, one has a clean and flexible infrastructure
- Balancing flexibility, duplication (i.e. forgetting to change the oddball), complexity of the module lists and configs

Puppet Hierarchy