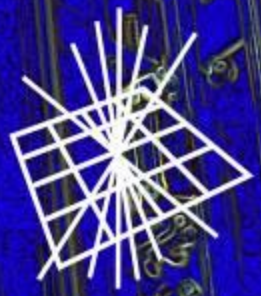

GridPP

UK Computing for Particle Physics

RAL Site Report

HEPiX Fall 2015 – BNL

12 – 16 October 2015

Martin Bly, STFC-RAL


Welcome
Rutherford Appleton Laboratory
SECURITY

Overview

- Hardware
- Network & IPv6
- CVMFS
- CEPH
- Cloud
- General

Tier1 Hardware

- CPU: ~140k HS06 (~14.8k cores)
- Storage: ~18PB disk
  - Latest batch for CEPH, in commissioning
- Tape: 10k slot SL8500 (one of two in system)

- Procurement for 2015/16 in progress
  - ~100k HS06
  - ~10PB (for CEPH)

Networking

- Tier1 LAN
  - Problem with Tier1 routers was delaying progress - fixed
    - RIP fix in firmware, long story…
  - Phase 2: 40Gb/s redundant link Tier1 to RAL Site Core - done
  - Phase 3: move the firewall bypass and OPN links to Tier1 routers
    - Will provide 40Gb/s pipe to border
    - Router configuration plan now finalised
    - Likely implementation early in new year
  - Some low level packet loss issues

- OPN
  - Still low level packet loss resisting attempts to isolate cause

IPv6

- IPv6 to RAL stripped off at RAL boarder and sent to isolated networks as required
  - IPv6 completely isolated from production IPv4 traffic on site
- IPv6 addressing plan requires approval from Network Technical Design Authority
  - Tier1 will be allocated 3 x /60 blocks for OPN routes, non-OPN routes and Facilities
- ip6tables firewall will be enabled on all Tier1 nodes before IPv6 is enabled on the production network
  - Defaults to 'DROP' on all traffic
  - Preparing a testing plan before submitting change management request
- Validating internal tools for provisioning, IP address management, monitoring…
- IPv6 will be enabled to perfSONAR nodes after OPN links are migrated to the Tier1 routers

IPv6 2

- Tier1 IPv6 Testbed
  - 3 in use, total 8 physical nodes
  - /64 address block for the testbed
  - IPv6 switch connects directly into the site border router
  - First NIC connected to a Tier1 IPv4 switch
  - Second NIC is connected to the IPv6 switch
  - Dual stacked UI box installed and working, with no off-site access.
  - Dual stacked Ceph S3 gateway installed and working, with off-site connectivity via IPv6 only. This is connected to our development Ceph cluster.
  - Currently investigating FTS3 support for S3 endpoints and GridFTP Ceph plugin
  - Working on setting up a perfSONAR node

CVMFS

- Stratum-0 for EGI infrastructure
- Stratum-1 for:
  - WLCG (cernvmfs.gridpp.rl.ac.uk)
  - EGI (2-node HA cluster - cvmfs-egi.gridpp.rl.ac.uk)

- Currently testing various options to move to larger storage (S3 CEPH gateway, RBD CEPH partitions), especially for Stratum-1

- RAL Tier1 is the main site outside CERN 'certifying' production releases before official announcement
  - Both for clients (on batch farm) and on server side (EGI infrastructure)
  - Always running the latest production releases, and from time to time the 'to-be-certified' releases

Grid Ceph @ RAL 1

- All hardware now deployed
  - Storage nodes: 21x120TB, 26x100TB
  - 1286 OSDs up+in, 5196TB raw space
  - 3 physical monitors, 3 physical gateways
  - 40 logical cores, 128GB RAM, 4x10GbE / server
  - No separate “cluster” (backend) network yet
- Ceph *Hammer* (0.94.1)
- RADOS gateways using civetweb for S3
- Xrootd 4 and GridFTP with Ceph support being deployed
- Ongoing work on monitoring
  - Custom Python scripts for stats collection
  - InfluxDB for time-series data
  - Grafana dashboards

Grid Ceph @ RAL 2

- Erasure coding tests
  - 3 nodes, 1 x 10GbE outgoing link
    - 16+3 EC pool
    - 32 nodes x 32 threads
    - 24hour test
  - 2.37GB/sec
- Further work
  - Logging to ElasticSearch
  - More EC testing
  - Extensive fault tolerance and recovery tests, gateway testing
  - Cluster networking

- Talk by George Vasilakakos

Cloud

- Now offering SL6, SL7 and Ubuntu images through a custom web-portal to staff for development and testing work (over 80 users)
- Using spare capacity up with virtual worker nodes
- Developing use cases with other communities, including some STFC facilities
- Underlying Ceph Storage proving very stable
  - Recently replaced the Ceph Monitors
  - Ongoing problems as a result of how OpenNebula behaves, but no problems on the Ceph side
- Moving towards a production service

Virtualisation

- Most T1 services systems run on virtual platform
- Hypervisors running MS 2008 Server / HyperV
  - Many small local-storage clusters
- New infrastructure in test
  - MS Server 2012 R2 / HyperV
  - Tiered shared storage on EqualLogic
  - At least 3 x 5- or 6-node clusters in two data centres
    - Expandable
  - Development clusters from remaining local storage clusters

General

- Batch system (HTCondor)
  - Now using CVMFS for grid middleware on worker nodes
    - except glexec & CASTOR clients
  - Continuous opportunistic use of private Open Nebula cloud
    - Generally ~300 cores from the cloud used in the batch system
- Evolution of grid services infrastructure
  - Investigating migration of services from VMs to containers managed by Mesos
    - Talk by Andrew Lahiff
- Monitoring
  - Investigating InfluxDB and Grafana as a replacement for Ganglia
    - Test dashboards setup for Ceph, HTCondor and FTS3

- Questions…?