GridPP perfSONAR refresh

Duncan Rand
on behalf of the Group

- Notes made from meeting yesterday
- Ideas for discussion please interrupt me as I proceed

- Introduction in Tim's talk at GridPP43 in Ambleside

  https://indico.cern.ch/event/828577/contributions/3488933/

Tier-2 evolution: networking implications

- In the future mode of operation of Tier-2s will diverge
- Five large Tier-2s to become storage-heavy sites
  - exchanging significant amounts of data with other WLCG sites
  - many sites overseas, possibly long distances away
  - anticipating increasingly high bandwidth connections (up to 100G over next few years)
    - RAL and Imperial already 100G
    - QMUL possibly upgrade around end of 2020
    - Others in discussion
- Bulk data exchanged via FTS – mainly Gridftp now, but increasingly http and XrootD
- Also supplying local compute-heavy Tier-2s with data – XrootD
- Network will remain important for successful operations
- Need ability to monitor network and debug issues

Compute-heavy Tier-2s

- Storage will decrease at these sites and they will run as compute farms
  - possibly with local disk caches

- Exchanging data with nearby storage-heavy Tier-2s (and Tier-1)
  - uploading Monte Carlo output
  - reading data (via XrootD)
- Significant network usage
  - mainly local, not long distances – perhaps only over Janet
  - local data caching e.g. Xcache?
  - But the network will become critical - any issues may directly affect site's operation

Plan

- In the first instance refresh just the hardware
- In the past we installed two hosts, one throughput, one latency and loss
- Now only need one perfSONAR host with three interfaces - throughput, loss/latency and management
  - already being done at Jisc, RALPPD and Bristol
- Tier-2s do not need to worry about funding for this host
  - will either be receiving some extra capital this year, in which case you will be expected to buy your perfSonar box out of this grant or
  - RAL will provide it for you
- Who might need a host?
  - RALPPD and Bristol have already updated hosts themselves – but do we want them to be the same specification as others?
  - Leaves RHUL, Oxford, Liverpool, Edinburgh, Brunel, Sussex, Cambridge, Birmingham, Durham, Sheffield, UCL

Rough Timescale

- Agree a spec ~2 weeks
- Delivery in ~6-8 weeks
- Sort out loan agreements
- Transport around UK ~2 weeks

- So by January 2020 sites should have hardware to run

Specification

- Needs to start of being capable of 100G
  - over ~5 year timescale that this hardware will be in service it is likely that sites may have upgraded to 100G
  - don’t want to have to upgrade them midway through lifetime
- All same specification
- Redundant features such as PSU, hot swap RAID etc
- See https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind1201&L=TB-SUPPORT&O=D&P=5928 for the discussion in 2012 on a GridPP-funded deployment of perfSONAR at all sites, and of the desired specification and features

Example: the new 100G CERN specification

- Supermicro X10DRi
- Intel(R) Xeon(R) CPU E5-2620 v4 (2 CPUs/16 cores, Intel S2600WT family)
- 128GB of DDR4 Micron memory (MTA18ASF2G72PDZ-2G3B1)
- 1x hardware RAID controller providing 8x SFF-8643 ports
- 1x 960GB Micron 5200 Pro SSD
- 2x 10GBase-T on-board Intel
- 1x two port 40GBase-T Mellanox Connect X-4 (MT27700)
- 1x two port 100GBase-T Mellanox Connect X-5 (MT27800)

Another possible spec (Martin and Alastair)

- A machine that we estimate will cost around £4.6k we could get:
  - CPU: AMD EPYC 7261 8C/16T
  - Memory: 32GB
  - Disk: 2 x 480GB SSD in a hardware RAID1 configuration
  - Network:
  - 1x Mellanox Connect X-4 Dual port 10/25GbE
  - 1x Mellanox Connect X-5 Dual port 40/100GbE
  - 6 year warranty.

- a thread on the perfsonar-user list a few months ago: https://lists.internet2.edu/sympa/arc/perfsonar-user/2019-04/msg00090.html

Deployment

- Recommended to use ISO image or stock OS install and rpms on bare metal
- My preference would be that each larger site looks after its own host
- For example the Jisc hosts with modern hardware and auto-updating software require little time to administer
- Perhaps for smaller sites with fewer system administrator personnel on site we might be able to relax requirements a little
- Ideas?
  - containers, ansible, remote management of the box via ssh?

Configuration, Archiving and Monitoring

- Continue to use existing WLCG infrastructure to configure and monitor the perfSONAR hosts
- Data is archived centrally by OSG
- MaDDash dashboard: https://psmad.opensciencegrid.org/maddash-webui/index.cgi?dashboard=UK%20Mesh%20Config
- Grafana pages
- Existing WLCG check_mk monitoring works well
- https://psetf.opensciencegrid.org/etf/check_mk/index.py?start_url=%2Fetf%2Fcheck_mk%2Fview.py%3Fhostgroup%3DUK%26opthost_group%3DUK%26view_name%3Dhostgroup
- Possible to set up email alerts to issues – contact me if you are interested

Documentation and Testing

- WLCG/OSG documentation available at https://opensciencegrid.org/networking/
- Write a FAQ for GridPP specific stuff?
- Testing: might consider funding two boxes at certain sites if they agree to do some testing of things like containers