Lancaster: Almost Dry again

EPPSYSMAN SITE REPORT

Matt Doidge, Robin Long, Peter Love

22nd June 2016

And the rain kept coming down.

As you may of heard we had a spot of weather last December...

Figure : source:bbc.co.uk

Lancaster.

A tale of too many bulletpoints.

Seriously, I’m really sorry for all these bullet points.

- Shared Site.
  - 50% stake in the "HEC" cluster.
- Atlas T2D, large local neutrino presence.
- Decent wadge of storage.
  - 2 Petabytes behind a DPM.
- Great relationship with our central services.
  - We’re forging a better one with the users too.

What's the HEC?

The University's HPC cluster, which we have root privilages (and responsibilities for). Purchasing is overseen by Robin, and Roger sits on the HEC steering committee. Our counterpart from ISS, Mike Pacey, proper knows what he's on about.

- 4800 cores, all in various generations of twin^2 units (bar a few GPUs).
  - Last purchases were dual E5-2640v3s with 64GB RAM and 1TB HDDs, packing 279 hepspec06 a node.
  - We use 10GbE to our nodes, for bandwidth and improved MPI performance.
- We're running S(on of) Grid Engine - beginning to show its age.
- Installation and Management is done using Cobbler - a lot of "bespoke" scripts and snippets.
  - Ansible will provide a better way (see Robin's talk).
- Local storage is provided using a 60TB Panasus shelf.

HEC III: Son of the Son of HEC.

T’other year we moved our SGE frontend nodes and CEs to (one of the) University’s VMWare instances, running them as VMs on the Uni’s Enterprise hardware (for free).

- For a while we tried to run on our own instance whilst the networking was being sorted out - this didn’t work out so well due to the slower performance of the commodity disk backend.
- Running wonderfully smoothly - not a blip. Actually more resiliant then we can make use of!
- Robin just moved our squid and BDII there- using ansible.
- So smooth that we’re thinking of moving our DPM headnode to the VMWare cluster - despite the "do not virtualise" warnings.

Diagram Break-VMWare Hardware.

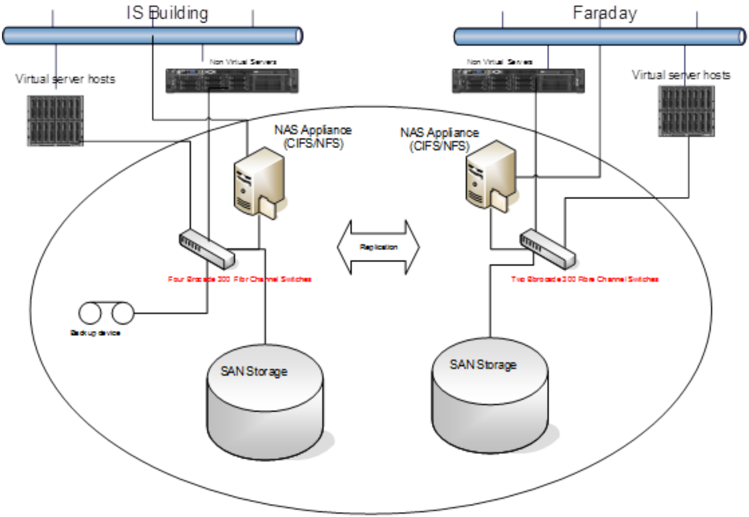

Figure : source:Matt Storey

Networking.

Like many, we’ve monopolised the University’s "backup" link to JANET for an (almost) dedicated 10Gb link.

- But no IPv6 yet - router work needed. August downtime should get that sorted.
- We’ve often had disappointing perfsonar results, despite excellent performance. Not sure what’s going on there - picture improved on its own last summer.
- WNs are NATed, and likely always will be (due the sharedness). Currently using the spare perfsonar box for this - may need a better solution.
- Our lovely network guys show Willingness to upgrade our link if someone ponies up the cash - or at least show it’s needed.

HEC 4.0

(Son of) Grid Engine is showing it’s age, and so is the SL6 kernel.

- Currently we’re running a cycle which sees a full "reincarnation" of the HEC service every 2 years.
  - The next reincarnation is due in 2017, probably around Easter.
- There will be a strong push from local users to move to an OS ending in a "7".
- There will be a strong push from the VOs and Robin and myself for a more feature rich (cgroups please!) batch system.

We’re going to have to start seriously thinking about this soon, and suggestions are welcome.

DPM dips.

We’ve been running DPM for 8 years, largely without error - until recently.

- Beginning to creak - rate of jobs/test timeouts increasing.
- Memory Usage and number of connections creeping up.
- Headnode reboots seem to cheer things up - for a bit.

The Hardware is at the end of its life, but shows no errors. This instance is 5 years old - maybe the database is like my basement and jammed with rubbish that should be thrown away? There’s an opportunity for an intervention at the start of August-but how big?

- For reference, our recent storage purchases have been 36-bay supermicros with 4TB drives, dual 10GbE interfaces delivering 115TB of usable space each.

Last Few Purchases.

- Our kit is getting old and has started to need to be "retired".
  - The big challenge/hassle is retiring the storage.
- Our Current Stategy is keep at 2PB, buying enough storage to maintain this number.
- The rest of the cash goes on compute.
- Of course things are easier for us, as a lot of this compute money is coming from the University via refresh budgets.
- Luckily the University forks out for our Water Cooled Rittal Racks (which we currently can’t fill anyway due to power maximums).

EPP, our local users

It’s not just the Grid Site that makes use of the offerings of our Central Services.

- Much use of the VMware cluster for service nodes and providing raw compute.
  - Purchased a blade to go in the VMWare cluster to provide most of this umph.
  - Many of these VMs plugged into the HTCondor local batch system.
- Mounted the University "SILO" on EPP cluster nodes.
  - Provides a high availabity, very large volume.
  - *But* not very high performance - meant for archiving rather then analysis.
  - ...we had a bit of an "expectation management" crisis when this was first announced, where a single user almost broke it.

Bringing in the Strays

You might have noticed something on the previous slide...

- There's a bit of a disconnect formed between the Grid and non-Atlas EPP members.
- Trying to bring in these lost shEEP with carrots - like offering T2K more storage.
- Also reaching out to the newer users groups - including the Astros as well as the new (and future) EPP experiments.
  - Some success straight away with microboone.
- Also also, there's (possibly) call for our "expertise" from other fields - particularly when it comes to storing and moving data. Nothing beyond an informal chat - so far.

Miscellanea.

Some random plugging

. Other bits we do are:

- Robin engages in a lot of user education, particularly in Linux use. How do you guys train yours?
- Robin is also a Master of Monitoring - we use icinga, ganglia and a few local scripts somewhat nicely presented in dashing.
- Mike has a number of exciting homegrown cluster management tools, including an excellent over-temperature protection system.
- Matt is the keeper of the tarball WN and UI software
    - I'm always interested to hear views or wishes.
    - Both clients are in cvmfs (/cvmfs/grid.cern.ch)
    - (Poor) documentation is at https://www.gridpp.ac.uk/wiki/EMITarball
    - I've been working on CentOS 7 versions - a basic UI is available.

Questions?

Matt Doidge, Robin Long, Peter Love
22nd June 2016