UNIVERSITY OF BIRMINGHAM

Birmingham Status and Plans

GridPP Collaboration Meeting, 7th April, 2017

Mark Slater, Birmingham University

UNIVERSITY OF BIRMINGHAM

Current Status (1)

As of right now, the Birmingham Grid Site consists of:

- ~1500 Cores providing ~17K HS06
- 980TB Storage with another 200TB being prepared

Cluster management is done with Puppet/Foreman. Current running services include:

- Torque/CREAM batch system
- ALICE Storage on XRootD install
- All other storage on DPM
- Squid, BDII, APEL, ARGUS, VO Box

The divide between experiments is ALICE 60%, ATLAS 30%, LHCb 5%, Other 5%

UNIVERSITY OF BIRMINGHAM

Current Status (2)

The biggest change in recent months has obviously been the loss of Matt Williams

There were a few teething problems last term due to this, mostly because I was swamped with teaching

However, generally, the local issues I have to deal with don't stop me keeping the site running, just reduce my ability to make large scale improvements

With this in mind, I have been moving forward with plans for the future taking into account the reduced manpower:

- Making progress on server room rearrangement
- Integrating all monitoring into Grafana
- Switching to using VAC
- Switching to using ZFS for the storage

One significant area where I haven't had chance to cover what Matt was doing is with Ganga – I do still plan to return to this but it will be limited for the next few months

Switching to ZFS

After several discussions with people, I have decided to (gradually!) move all our storage from hardware RAID 6 to ZFS and not buy RAID6 cards for new storage

From my point of this has several benefits:

- Easy to monitor disk health across all systems
- Can use ~any disks in the RAID
- Cheaper to buy new hardware

I have currently moved 40TB of storage over to ZFS on Tier2 and have had no issues at present...

Making Monitoring Easier

Before Matt left, he installed Grafana and started setting up monitoring pages. I've been continuing this work to cover both T2 and T3 machines:

- Graphite/Carbon system is incredibly easy to setup
- Can monitor everything I want and easily add more
- Grafana makes setting up dashboards trivial

Moving Workers to VAC

The biggest ongoing change is that I'm switching all the workers from Torque/CREAM to VAC. Again, this has many benefits for us:

- Very easy to setup (after initial teething problems)
- Don't have to worry as much about OS updates, etc.
- Minimal ongoing administration required
- Don't have to run CREAM, Torque, APEL
- Reduces complexity of other services (Squid, BDII, Argus)
- Overall a *significant* reduction in manpower required

Drawbacks I've currently encountered:

- Initial setup did have problems (mostly because of me!)
- Much harder to overprovision due to HD and memory reqs being 'enforced'
- I found I needed a Squid per VM factory/Worker

Over the next few slides, I'll briefly cover the setup, (minor) issues encountered and our current status

Many Thanks to Andrew McNab for helping me through the setup!

Installing/Configuring VAC

Generally, the install and setup of VAC was very easy. I just followed the instructions on the web page:

https://www.gridpp.ac.uk/vac/admin-guide.html

Fundamentally though, after installing appropriate libvirt tools, it's just a case of installing a single RPM

The configuration is managed through a small handful of easy-to-understand config files.

The only issues/gotchas I encountered were:

Firewall:

As I use puppet to manage iptables, it took a few tries to get every rule put in correctly

HS06, GOCDB entry:

VAC is able to send accounting records directly, however you must remember to add an appropriate GOCDB entry and the HS06 values for each worker node

Squid:

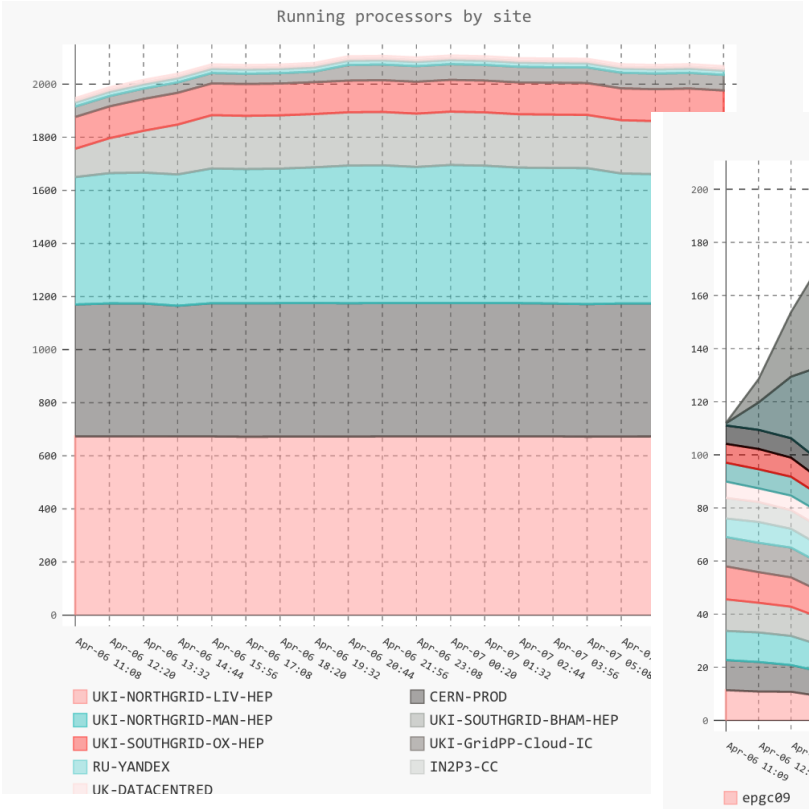
You will probably need multiple squids to cover the additional load because, as far as the squid is concerned, you will have a worker node per core. Our squid couldn't handle this and so I went to a squid per factory. Hopefully I can reduce this in the future.

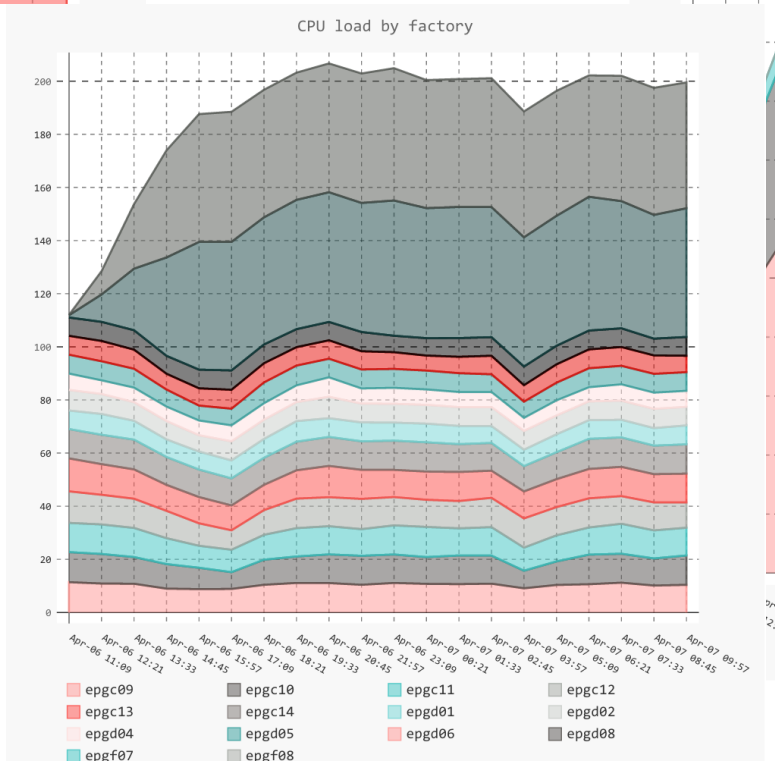
Monitoring VAC

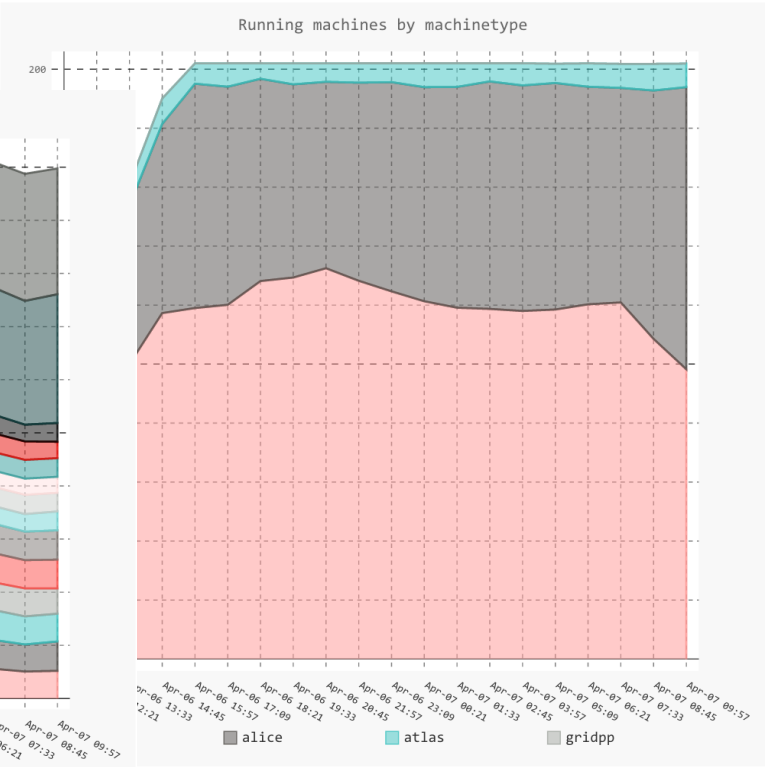
There is very good overall VAC monitoring available here:

http://vacmon.gridpp.ac.uk/1f4:15180::/

From this you can drill down to your site and individual workers







UNIVERSITY OF BIRMINGHAM

Current VAC Status at Bham

At time of writing we have ~200 cores devoted to VAC (~13% of the whole site).

The relatively slow movement to VAC is not due to problems with VAC or lack of time, it's that I have been taking this opportunity to improve the cabling and server layout when a worker is drained

At present I can see no reason not to keep this transfer going until the whole site has been converted. I could then decommission the CREAM CE, Torque server and APEL box

Timescale for this (including the server recabling/movement) is ~6 months.

After this, we should be in a very good and much more sustainable position to keep the site running with minimal man power.

UNIVERSITY OF BIRMINGHAM

Conclusions

Birmingham would very much like to continue being part of GridPP in the future if at all possible

I've put in place several plans to make this possible within the reduced funding scheme:

- Reorganised server room allowing for ease of installation and expansion
- Easy monitoring of all aspects of the site via Grafana
- Switching to ZFS over HW RAID 6
- Moving all workers to VAC

Of these, by far the most significant will be the switch to VAC as this will remove what can be a large drain on my time without the big hurdle of installing and maintaining ARC/Condor