UNIVERSITY OF BIRMINGHAM

Birmingham Status

Alice T1/T2 Workshop, 14th May 2019

Mark Slater, Birmingham University

UNIVERSITY OF BIRMINGHAM

Birmingham Status

The Bham site currently consists over 1700 CPU cores and 1.4PB of storage of which ~60% are available to ALICE. This equates to ~1100 cores and just over 1PB of storage

These resources are presented using the VAC system for compute and EOS for Storage.

ALICE shares the compute resources but the EOS instance belongs exclusively to ALICE (looking into caching mechanisms for ATLAS)

This technology selection has been driven by the diminishing manpower available and have eased the ongoing administrative burden.

For those not familiar with VAC, runs jobs in VMs or containers that are created based on work available from central services and fairshare

This means that the usual CE/Batch system services are not needed but you can be a bit limited with who can use your systems.

For ALICE, at the moment, you also don't need a VOBox to run – all jobs are run under the 'Altaria' site. This is being looked at for future improvements though.

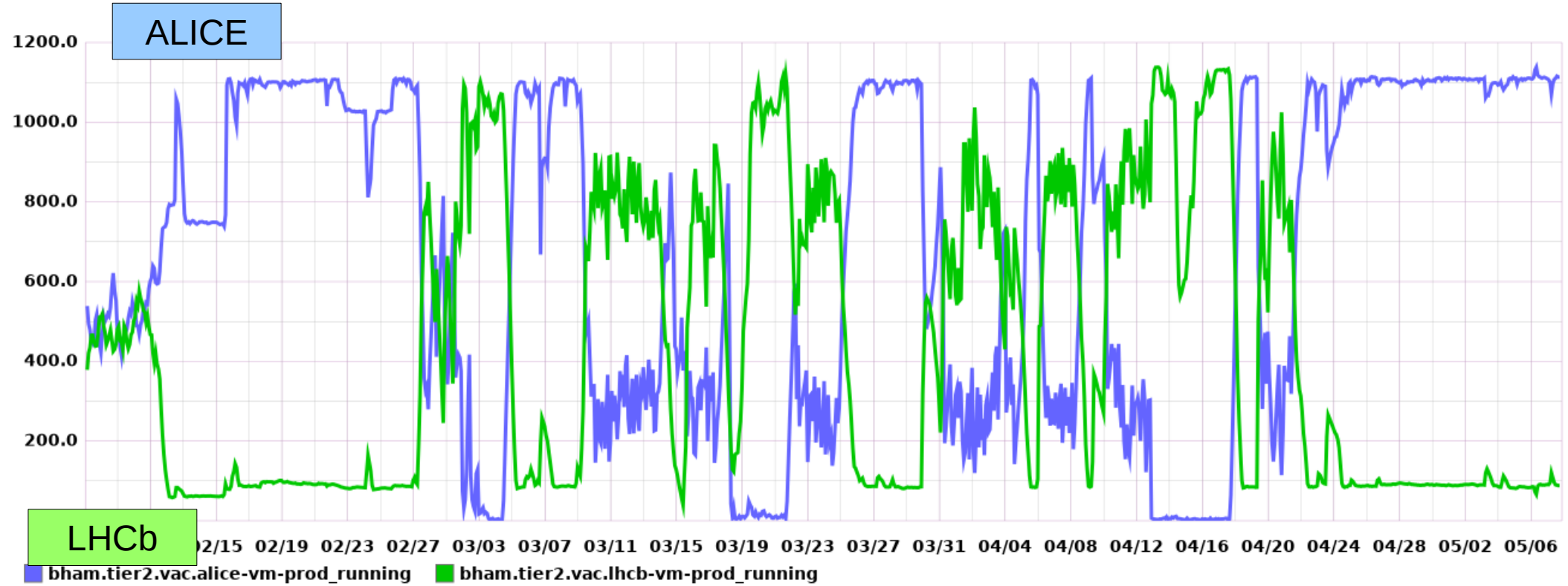
This technology choice means I no longer have a CE, batch system, APEL box or BDII to look after!

For more info on VAC, please see:

https://www.gridpp.ac.uk/vacproject/

Compute at Bham

We've been running at capacity and stably in this mode since last Autumn (showing from February):



The drops have generally been due to lack of work where LHCb have taken up the slack.

UNIVERSITY OF BIRMINGHAM

Storage at Bham

Early last year, by request from ALICE, I started moving the storage to EOS because it boasted many desirable features:

- **Low latency for operations due to in-memory namespace**
- **No need for hardware RAID due to replication or erasure encoding**
- **Unix user, GSI or Kerberos access**
- **Rich Access Control Lists**
- **User, Group and Project Quotas**
- **Server on Linux but clients available for Linux and OSX**
- **SAMBA and WebDAV access available**

It also has a lot of administrative tools available:

- **Comprehensive CLI tools to control all aspects of the system**
- **Automatic balancing between and within groups**
- **Draining of FSs and auto-repair from broken disks**
- **Highly configurable redundancy/striping on a per file/directory basis**

Unfortunately, I encountered several problems during the transition:

FST Connection Problems:

Initially found that FSTs would stay up for 1-2 hours under load and then connections would start backing up. These seemed to go away so maybe solved in later releases

Couldn't seem to get 'RAID-like' behaviour working:

This was the biggest problem – I couldn't get the RAID behaviour working across servers. Over August, lost many files when the server was put under load and the files were not copied properly

Documentation is a little lacking:

Though there is quite a bit of documentation available, it doesn't seem to spell everything out as plainly as I'd have liked. I spent sometime piecing together things from various sources.

The biggest of these problems has certainly been the erasure-encoding (or however the striping is done)

It was working mostly OK for 2-3 servers but after adding another node and being put under heavy load, a large number of the files had missing stripes and weren't recoverable

This may have been an issue with my configuration but I could not find any answers in the docs

I am now in the process of switching to a single replica but with the redundancy provided by ZFS

This seems to be working a lot better though it is taking some time to transfer the files from the 'emergency' 2 replica system I put in to the single replica one

I estimate a few months of draining/transferring yet before the system is how I want it

Current Status and Plans

Other than the continued conversion of the storage from 2-replica to 1-replica on ZFS, I have very few ALICE plans in the next 12 months

I have more storage to install (200TB => ~120TB more for ALICE) and in the summer (Brexit dependent!) plan to buy some more cores, all of which will be added to the existing systems.

I'm happy to help setup a VOBox to manage the (now 100%) VAC cluster but I believe that's the only development required.

Basically, everything seems to be working now so I am happy to keep it that way!