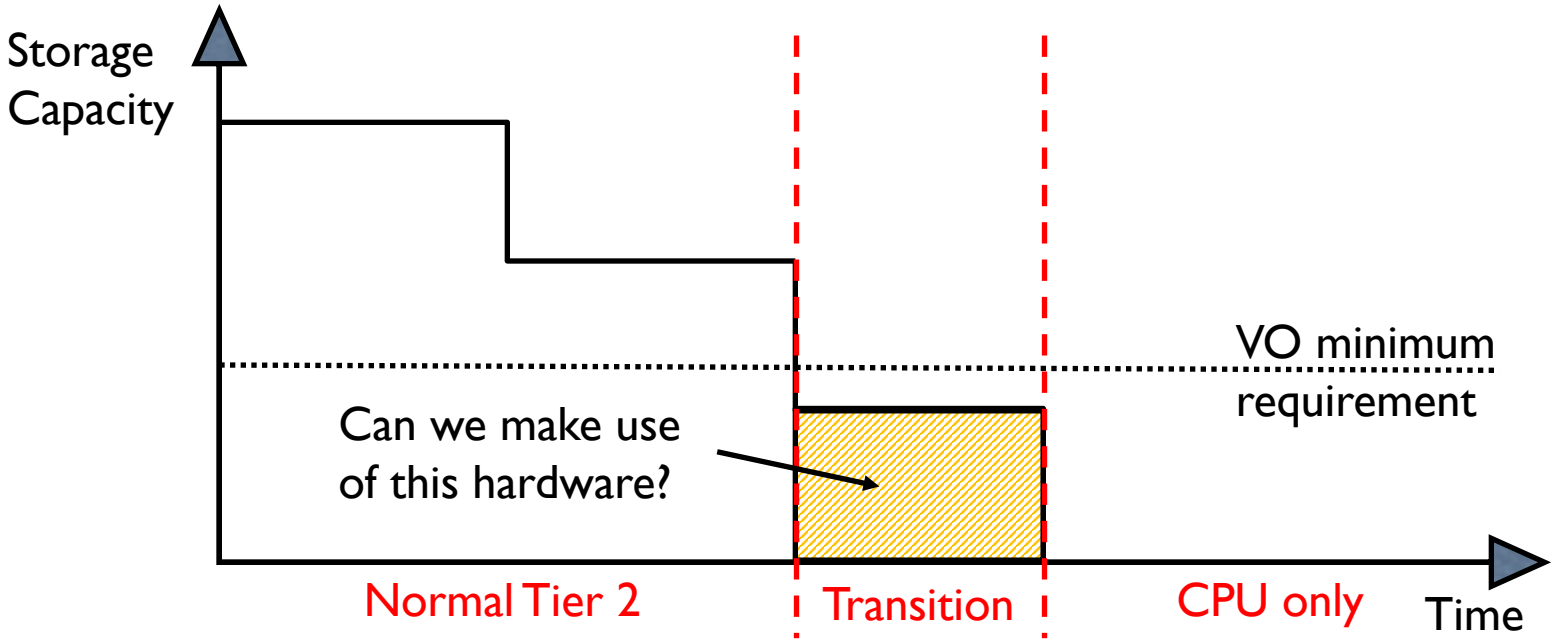
UK Tier-2 site evolution for ATLAS

Alastair Dewhurst

Introduction

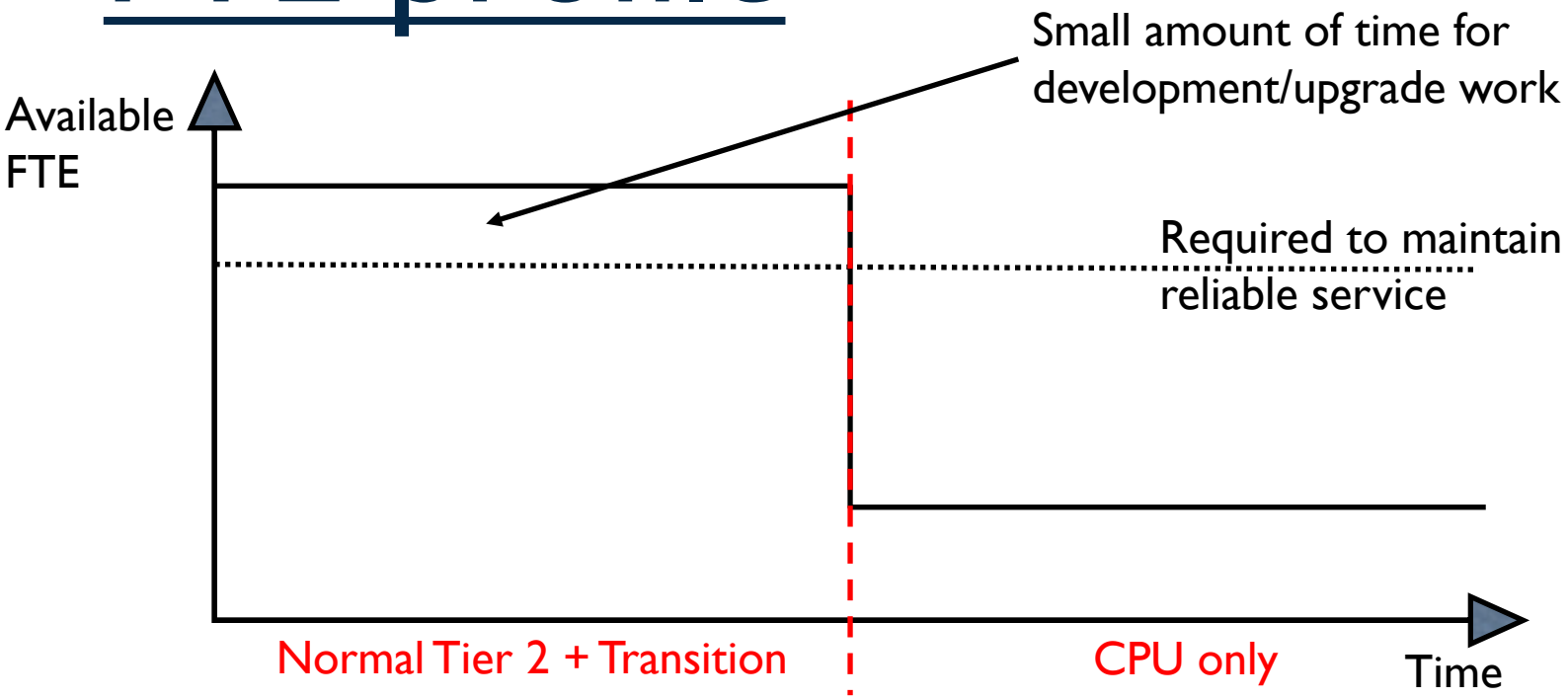
- My understanding is that GridPP funding is only part of the story when it comes to paying for a Tier 2 site.
  - Each site is unique.
- Aim to describe various options and let sites decide what is best for them.
- Where possible, I have tried to site my sources.
  - If not, then assume it is my opinion!

Hardware profile



- The UK may well be unique in having sites with significant storage capacity that gets decommissioned within the next 3 - 5 years.

FTE profile



- Solutions need to be deployed before the manpower drops.

CPU Only - UCL Model

- UCL was migrated to CPU only a while ago.
- More recently the French Cloud migrated some (non-French) sites to CPU only[1].
- UCL Panda Queues (PQs) points to QMUL DDM endpoints.
  - For all reads and writes.
- Two problems with extending this approach:
  1. Extra CPU could overload the other sites storage.
  2. If other site is down, then your resources are wasted.

[1] https://indico.cern.ch/event/609911/contributions/2605033/attachments/1479532/2293620/WLCG_2017_DISKLESS_EV_final.pdf

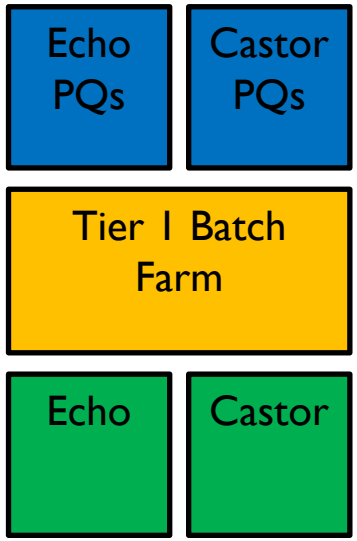
Lessons from Echo

- Adding a new storage endpoint at RAL, taught me a lot about the limitations/assumptions of ATLAS systems:
  - A Panda Queue (PQ) can only talk to one storage endpoint.
    - Hence, we have two sites in AGIS.
  - You don't need much to run a site.

Two separate sites in AGIS:
- RAL-LCG2
- RAL-LCG2-ECHO

The same batch farm is defined exactly the same way for both sites.
Both storage endpoints can take full load of Batch farm.

When Echo first started working for ATLAS, we had the backend storage, GridFTP access and a .json file which I manually edited for storage accounting.



ATLAS Vision

- ADC Technical Interchange Meeting (TIM) 20th - 22nd September [1].
  - Torre outlined his vision for his term as Computing Coordinator[2].
- [A]ES - ATLAS Event Service. Jobs can process as little as one event. Makes use of opportunistic resources such as HPCs.
- ESS - Event Streaming Service. Extension to AES, uses a prefetcher to pull in data a little at time for the job.
- CDN - Content Delivery Network. (Lets copy Netflix?!)
- Data Lake - A bit like a federation of storage endpoints…

[1] https://indico.cern.ch/event/654433/
[2] https://indico.cern.ch/event/570497/contributions/2726102/attachments/1524794/2384292/CompThemesWenaus.pdf

ATLAS Development

- There is a general consensus that we will need to have fewer replicas of data in future.
  - More remote access for jobs will be necessary.
- The data management team gave a talk on long term planning[1].
  - Primarily focused on improving what we have via better monitoring etc.
- The pilot team gave a talk on future developments[2].
  - Lots of stuff being re-written.
  - Alternative stage out not high on priority list. (ie being able to write output to a different site)

[1] https://indico.cern.ch/event/654433/contributions/2716593/attachments/1528356/2390751/2017-09-22_-_Data_management_-_Long_term_planning.pdf
[2] https://indico.cern.ch/event/654433/contributions/2712321/attachments/1528329/2390701/Pilot_2_CERN_TIM_22_September_2017.pdf

‘Closeness’

- ATLAS define ‘Closeness’ as:
  - The maximum throughput over one hour in a one month period. Convert this throughput in TB/s and take -log(throughput in TB/s).

Closeness of 0 = Same site
Closeness of 1 = 100 GB/s
Closeness of 2 = 10 GB/s
…
Closeness of 10 = 0.1 kB/s
Closeness of 11 = No data

		To RAL Castor	To RAL-echo	To B'ham	To rhul	To qmul	To glasgow	To lancaster	To edinburgh	To Mancs	To liverpool	To sheffield	To ralpp	To SUSX	To Cambridge	To Oxford	To durham
From	RAL Castor		3.00	3.00	4.00	3.00	3.00	3.00	4.00	3.00	4.00	4.00	4.00	5.00	4.00	4.00	4.00
From	RAL-echo	4.00		5.00	5.00	4.00	4.00	4.00	5.00	4.00	5.00	6.00	6.00	6.00	4.00	5.00	6.00
From	B'ham	5.00	6.00		5.00	5.00	5.00	5.00	5.00	5.00	5.00	6.00	6.00	5.00	4.00	6.00	11.00
From	rhul	3.00	5.00	2.00		4.00	5.00	4.00	3.00	4.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00
From	qmul	3.00	3.00	4.00	2.00		4.00	4.00	3.00	4.00	4.00	5.00	4.00	4.00	5.00	4.00	5.00
From	glasgow	3.00	4.00	3.00	2.00	4.00		4.00	4.00	4.00	4.00	4.00	4.00	5.00	5.00	5.00	6.00
From	lancaster	3.00	3.00	2.00	4.00	4.00	4.00		5.00	4.00	3.00	5.00	4.00	5.00	5.00	4.00	5.00
From	edinburgh	4.00	4.00	5.00	5.00	4.00	4.00	5.00		5.00	5.00	6.00	5.00	5.00	5.00	6.00	11.00
From	Mancs	3.00	4.00	4.00	2.00	4.00	4.00	4.00	4.00		4.00	4.00	4.00	5.00	5.00	4.00	6.00
From	liverpool	4.00	4.00	5.00	5.00	4.00	5.00	5.00	5.00	5.00		5.00	5.00	5.00	5.00	6.00	6.00
From	sheffield	4.00	5.00	3.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00		5.00	5.00	5.00	6.00	6.00
From	ralpp	4.00	5.00	5.00	5.00	4.00	5.00	5.00	5.00	4.00	5.00	5.00		5.00	5.00	5.00	6.00
From	SUSX	5.00	6.00	6.00	6.00	5.00	5.00	5.00	6.00	5.00	5.00	6.00	5.00		5.00	6.00	11.00
From	Cambridge	4.00	5.00	5.00	3.00	4.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00	5.00		6.00	11.00
From	Oxford	5.00	5.00	5.00	5.00	4.00	5.00	5.00	4.00	4.00	5.00	5.00	6.00	11.00	5.00		6.00
From	durham	5.00	6.00	11.00	6.00	5.00	6.00	6.00	11.00	8.00	6.00	6.00	11.00	11.00	11.00	11.00	

Federations (DynaFed)

- FAX was killed off by ATLAS.
  - Insufficient manpower (return on investment)
- I do not believe an http federation will be any more successful.
  - Webdav support is currently less well deployed than XRootD.
- What about DynaFed?
  - It could provide a solution for future experiments, that integrate it from the start.
  - RAL (Echo) uses it just as an authentication frontend for the S3.
- I don't think ATLAS need a federation, they have the information to target jobs to the best available site.

XRootD Proxy Cache

- XCache is the new name for XRootD Proxy Cache!
- Old disk servers could be configured as an XRootD Proxy Cache.
- We have some experience already from using them for CMS AAA.
  - Echo is copying RALPP's setup.
- ATLAS are trialing XCache at BNL and some US Tier 2, for use with ESS.
  - I am slightly concern it may become a 'requirement'.
- Note ATLAS and CMS use case is slightly different.
  - WN running CMS jobs anywhere connect to an XCache dedicated to a site.
  - WN running ATLAS jobs at a site can connect to any other site via XCache.

[1] http://xrootd.org/doc/dev42/pss_config.htm
[2] https://indico.cern.ch/event/330212/contributions/1718785/attachments/642383/883832/PFC-Details-Status.pdf

ARC Cache

- If the site has a shared file system (between CE and WN), then ARC Cache could be a useful choice to improve job performance.
- At a recent Ceph meeting[1], SiGNET gave a presentation on their CephFS setup which is used as an ARC cache as well as for local users.
- The UK does have plenty of experience with ARC CEs, but not with ARC Cache.

[1] https://indico.cern.ch/event/669931/contributions/2740166/attachments/1533388/2401265/Luminous_on_SiGNET.pdf

Analysis Facility

- We know that users tend to prefer their own site if using LOCALGROUPDISK.
- Analysis facility has a large LOCALGROUPDISK (a SCRATCHDISK but no DATADISK)
- Tried before with Sussex but wasn't used.
  - No existing user base.
- If there is a site with an active analysis base, who wanted this, then there is no reason it can't work.

Flexible batch farm

- In the long term, if a site is only going to have CPU resources for the Grid then maybe it should just focus on having a really flexible batch farm?
- Brian Bockelman gave a presentation recently on fancy new features in HTCondor.
  - https://indico.cern.ch/event/654433/contributions/2724530/attachments/1526468/2386983/HTCondorCE-ATLAS-TIM-2017.pdf

Conclusions

- If we can keep the storage reliable, I suspect ATLAS will allow us to bend the rules as the amount of storage on some sites shrink.
- I think we should investigate [further] XCache.
  - My concern is that this could become another expected service at all sites rather than something useful for just CPU only sites.