

UK Status and Plans

Catalin Condurache - STFC RAL

ALICE Tier-1/Tier-2 Workshop

Bergen University College, 18 April 2016



Science & Technology Facilities Council
Rutherford Appleton Laboratory



GridPP

UK Computing for Particle Physics



Content

- UK GridPP Collaboration
- Tier-2s Status and Plans
 - Birmingham
 - Oxford
- RAL Tier-1 Centre
 - Components status and plans
 - ALICE highlights
- Latest on storage at RAL





GridPP UK

- The GridPP Collaboration is a community of particle physicists and computer scientists based in the United Kingdom and at CERN
- It consistently delivers world-class computing in support of all LHC experiments and many more user communities in a wide variety of fields

GridPP UK

- ~10% of WLCG
- Collaborating Institutes
- ScotGrid
- NorthGrid
- SouthGrid
- LondonGrid

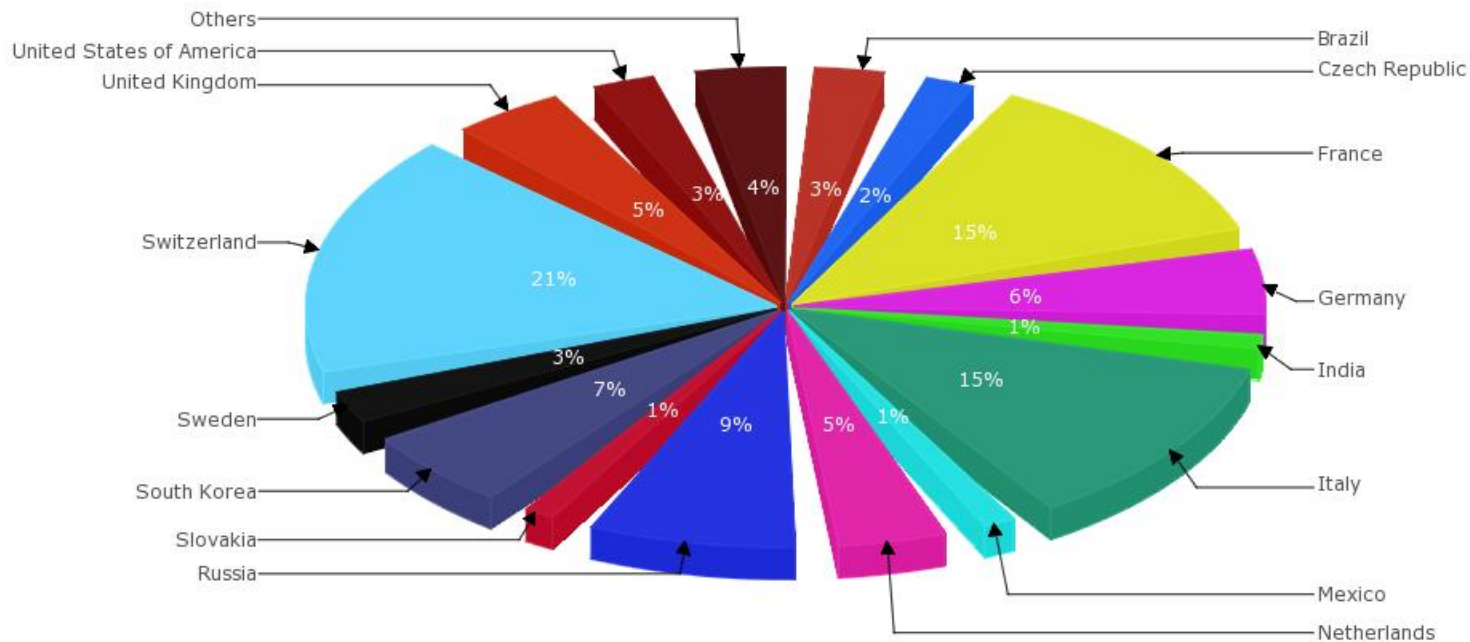


ALICE - CPU Accounting Last 12 Months Worldwide

Developed by CESGA 'EGI View': / normcpu / 2015:5-2016:4 / COUNTRY-VO / custom (x) / GRBAR-LIN / 1

2016-04-10 02:06

Normalised CPU time (kSI2K) per COUNTRY

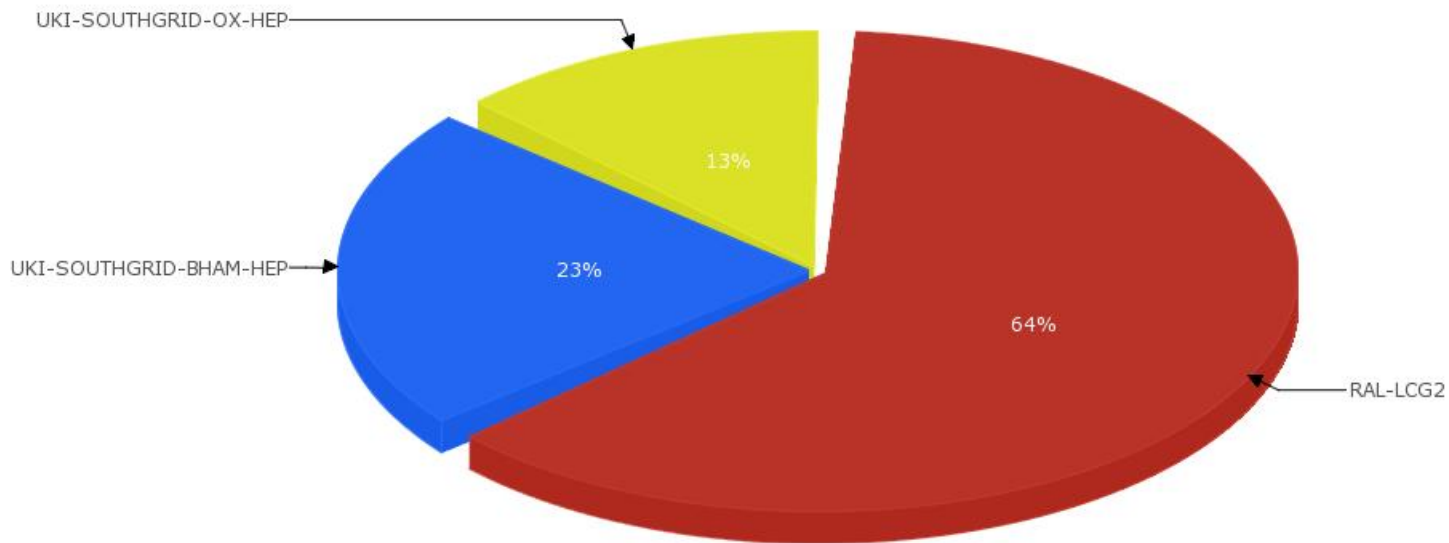


ALICE - CPU Accounting Last 12 Months UK

Developed by CESGA 'EGI View': / normcpu / 2015:5-2016:4 / SITE-VO / custom (x) / GRBAR-LIN / 1

2016-04-11 02:06

Normalised CPU time (kSI2K) per SITE





Tier-2s Status and Plans

- Birmingham

- Disk storage

- 522TB pledge for ALICE (from 280TB)
- to cover 2016, 2017, early 2018
- native XRootD

- CPU

- ~60% of UK T2 ALICE CPU allocation
- from current 1216 cores (12489 HS06) to 1408 cores (soon)
- 60% - fairshare for ALICE - *“very good at filling it”*



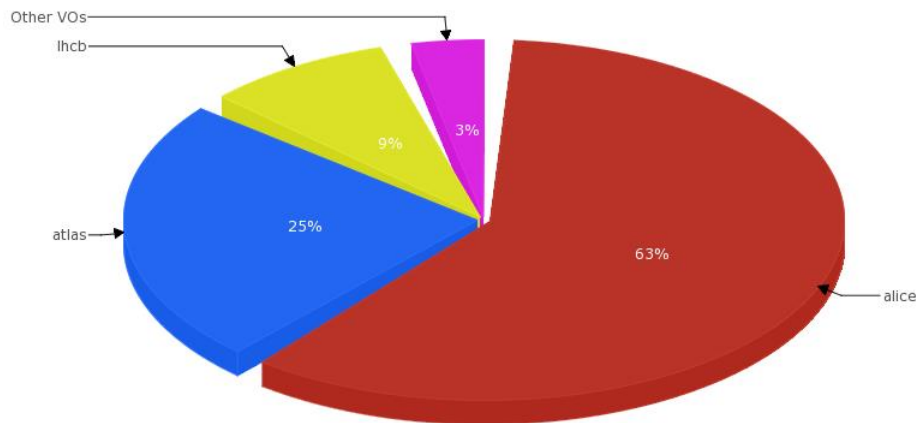
Tier-2s Status and Plans

- Birmingham
 - 63% overall CPU usage for ALICE
 - ATLAS 25%, LHCb 9%, others 3%

Developed by CESGA 'EGI View': / normcpu / 2015:5-2016:4 / SITE-VO / lhc (1) / GRBAR-LIN / 1

2016-04-07 02:06

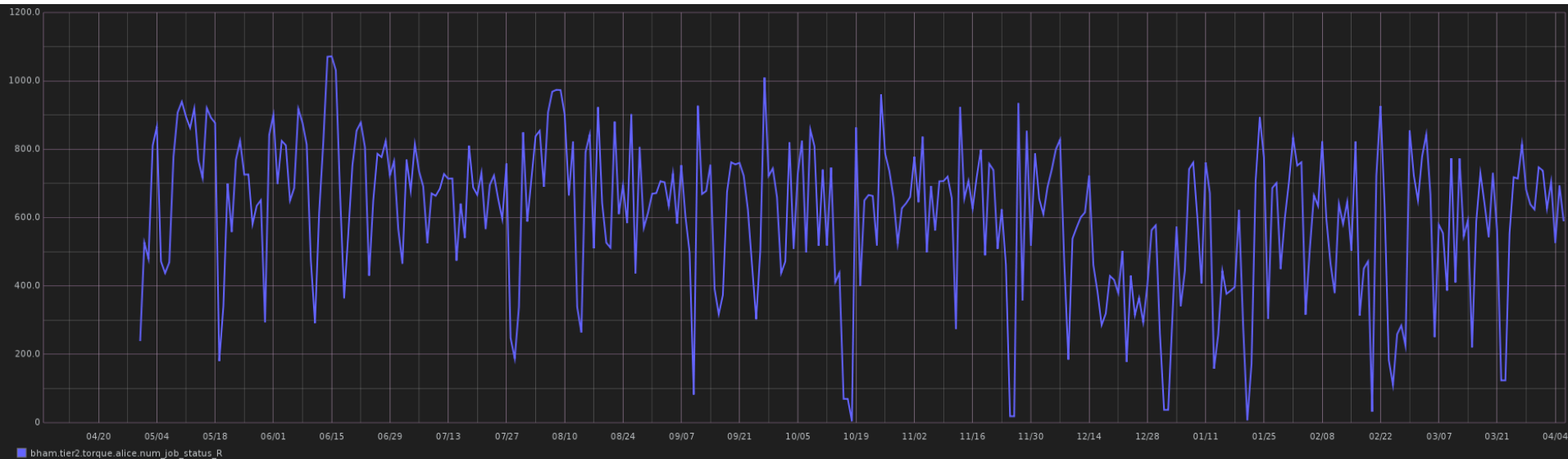
UKI-SOUTHGRID-BHAM-HEP Normalised CPU time (kSI2K) per VO





Tier-2s Status and Plans

- Birmingham
 - Running jobs since May 2015





Tier-2s Status and Plans

- **Birmingham**

- Currently still CREAM, ready for ARC in the next few months
- IPv6 - not yet, need addresses out of the University
 - maybe some progress by end Summer 2016



Tier-2s Status and Plans

- Oxford
 - Need to supplement the support given by Birmingham
 - ~40% of UK T2 ALICE CPU allocation
 - The Grid Cluster now runs HT Condor behind ARC-CE
 - Some problems with limiting jobs by number in Condor, so control by job priorities (150-200 ALICE jobs)
 - No storage provided

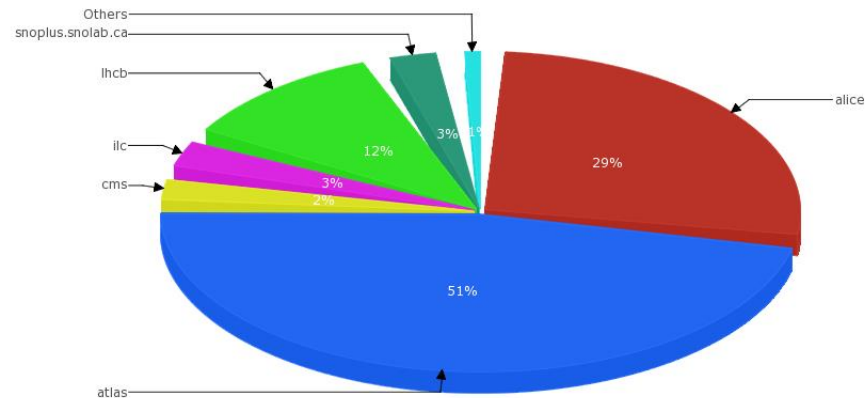
Tier-2s Status and Plans

- Oxford
 - Very efficient to fill the empty job slots (“*when ATLAS have problems!*”)
 - 29% CPU resource usage for ALICE

Developed by CESSA 'EGI View': / normcpu / 2015:5-2016:4 / SITE-VO / all (x) / GRBAR-LIN / I

2016-04-07 02:06

UKI-SOUTHGRID-OX-HEP Normalised CPU time (kSI2K) per VO





Rutherford Appleton Laboratory (RAL)

- 15 miles south of Oxford on Harwell Campus
- Run by STFC
- Multi-discipline centre supporting university and industrial research in big facilities:
Neutron Science, Lasers, Space Science, Computing
- Hosts UK LHC Tier-1 Facility (RAL Tier-1, RAL-LCG2)





RAL Tier-1 Centre

- **Hardware**

- CPU: ~140k HS06 (~14.8k cores) - from 10.6k cores
 - FY 15/16: additional ~106k HS06 in test
 - ~250 kHS06 in use in July 2016
- Storage: ~16.5PB disk - from 14PB
 - FY 15/16: additional ~13.3PB raw - CEPH specs
- Tape: 10k slot SL8500
 - 44PB - T10K C/D
 - migrations to D-only started (estimated 1 month/PB)



RAL Tier-1 Centre


- **Services**

- Migration to ARC + HTCondor
- Last CREAM-CEs stopped in August 2015
- Batch system
 - developed a new method for draining WNs for multi-core jobs, enabling to run pre-emptable jobs on the cores which would otherwise be idle
 - in production since late last year
- Mesos - project to investigate management of services

RAL Tier-1 Centre

- CernVM-FS
 - Stratum-0 for EGI
 - Soon larger backend storage for Stratum-1 service (WLCG, EGI, OSG etc)

CernVM File system CernVM[-FS] Workshop, RAL, 6th-8th June!



<https://indico.cern.ch/event/469775>

New developments, discussion, hands-on, and technology outlook

Invited speakers

- Josh Simons, VMware's Office of the CTO
- Martin Stadler, Director, Linaro Enterprise Group ← ARM
- Oliver Oberst, IT Architect at IBM's HPC Division
- George Lestaris, Software Engineer at Pivotal
- Artem Harutyunyan, Director of Engineering at Mesosphere

1 / 1

← Registration still open!



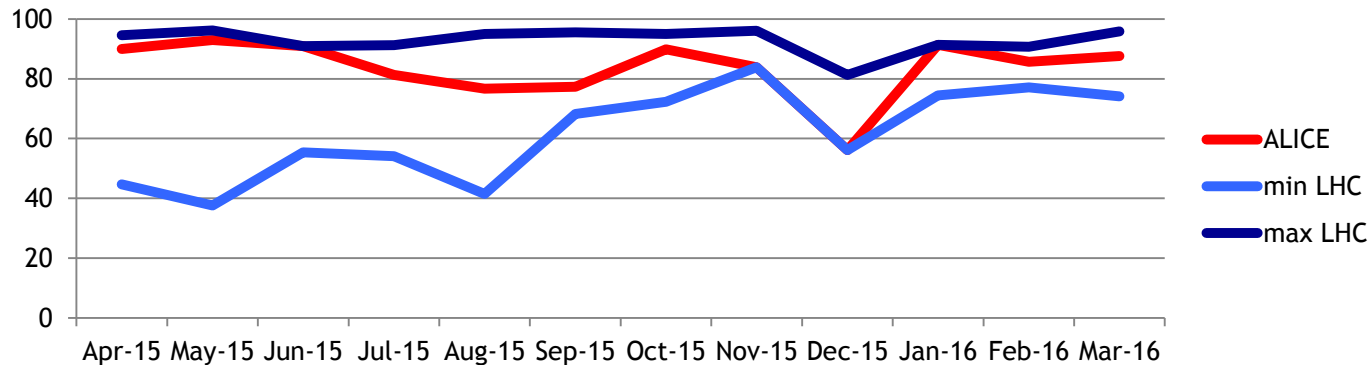
RAL Tier-1 Centre

- **CASTOR, Cloud**
 - **Castor - v2.1.14**
 - stable running for the start of run 2
 - major improvements in data throughput from disk thanks to scheduling optimisation
 - OS SL6 and Oracle version upgrades for entire system
 - no plans yet to upgrade to 2.1.16
 - **Cloud**
 - production service using OpenNebula
 - department and wider use in STFC
- **Also CEPH (few slides later...)**



RAL Tier-1 - More on ALICE

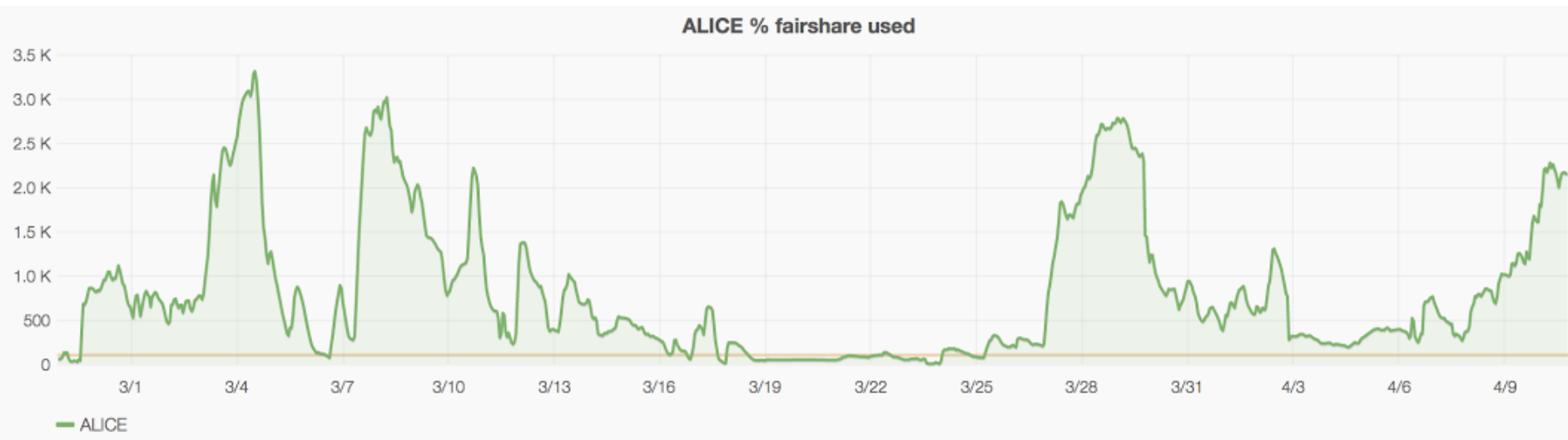
- CPU fairshare - 1.865% (2400 HS06) in 2015
- NO limit on opportunistic use of spare cycles jobs
 - February 2015 - intermediate 6000 limit (from 3500, following discussions at ALICE T1/T2 Torino)
 - Capping removed - no more limits!! - June 2015
- CPU efficiencies - >80% average for ALICE





RAL Tier-1 - More on ALICE

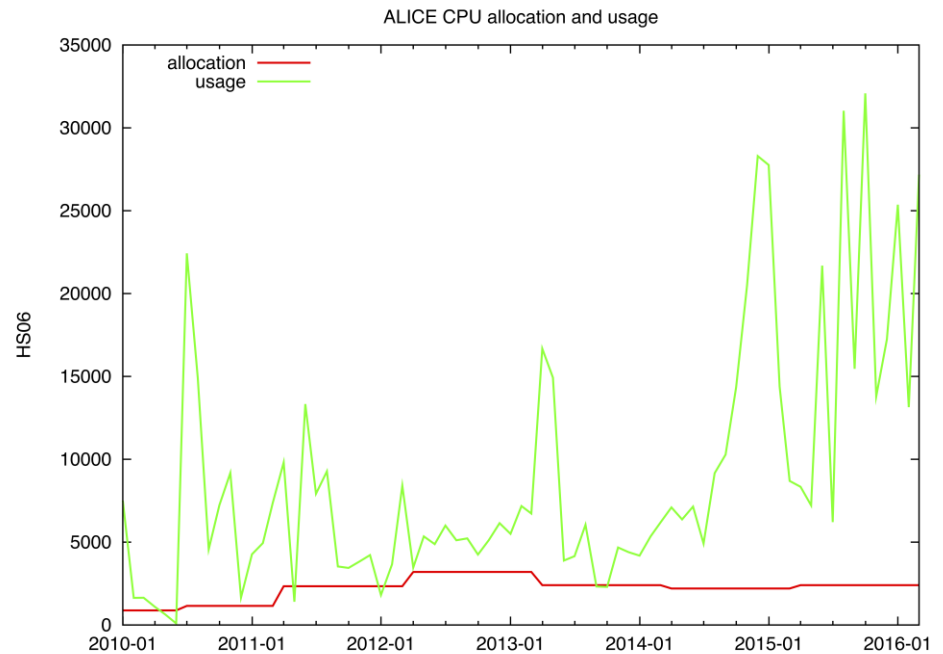
- Good use of a significant amount of opportunistic CPU
- In the graph below (March 2016), the expectation is 100
- Average for ALICE - at least 10 times the fairshare
 - Significantly much higher peaks





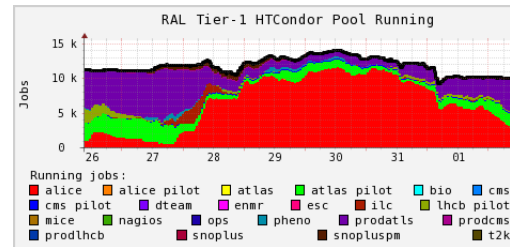
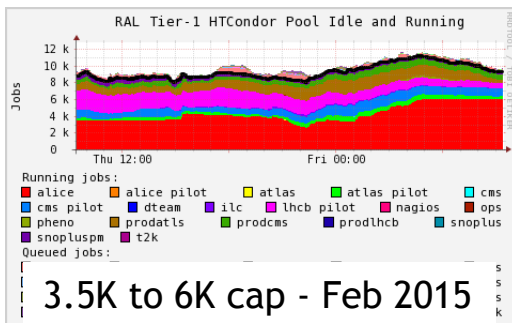
RAL Tier-1 - More on ALICE

- Monthly allocation and usage for ALICE since 2010
- Usage is consistently high
- If this was a non-LHC VO, we would probably revise their allocation

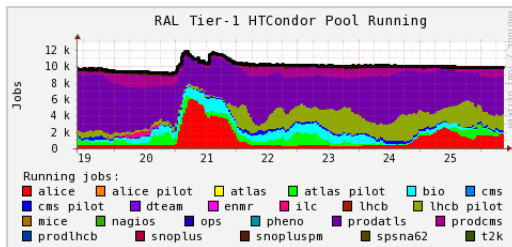
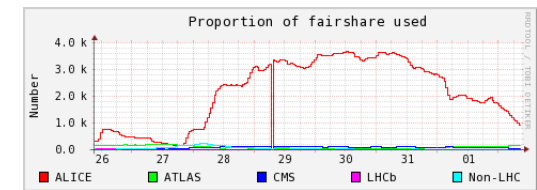


RAL Tier-1 - More on ALICE

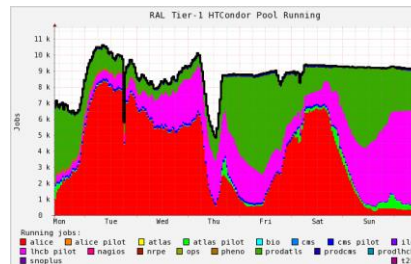
- Few more nice CPU graphs



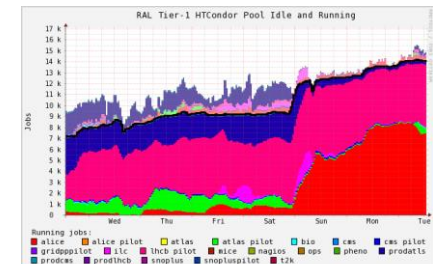
30 August 2015



Castor d/time - May 2015



Xmas 2015



ATLAS, CMS issues - 29 Mar 2016



RAL Tier-1 - More on ALICE

Disk storage

- 356TB disk allocated (395TB deployed) in 2015

Tape storage

- 420TB allocated in 2015
- 365TB used in April, then 500TB in August, 736TB in September and *capped* at 870TB in October



RAL Tier-1 - Immediate Future

- CPU for ALICE
 - 2016Q2 - 3140 HS06 (~340 job slots)
 - from 2400 HS06
- Disk storage for ALICE
 - Pledge increased to 420TB in April 2016
- Tape storage for ALICE
 - 2016Q2 pledge is 312TB...
 - ...but *gentlemen agreement* to keep it at 870TB (deployed)



RAL Tier-1 - Disk Storage for ALICE

- Funding (GridPP project 2016 - 2020)
 - 10% reduction in staff at Tier-1 since 2014
 - Further 20% reduction planned in April 2018
- To continue to meet WLCG commitments - need to reduce costs:
 - Make further efficiency savings (do same for less)
 - Share costs with other communities (common technologies)
 - Maximise convergence of LHC services (do less different things)



RAL Tier-1 - Disk Storage for ALICE

- Need to streamline storage services and simplify
- Plan to deliver LHC disk only storage on CEPH, sharing costs with other projects
 - ATLAS, CMS, LHCb ready to move to CEPH
 - Intend to phase out CASTOR D1T0 and D1T1 by March 2018 (start this year)
 - In 2017 consolidate four CASTOR D0T1 tape instances down to one shared instance



CEPH at RAL Tier-1

- Production level service underpinning Cloud infrastructure
- CEPH Hammer 0.94.4
 - 4000TB raw storage space (~42 nodes)
 - 950 OSDs (2-3GB RAM per Object Storage Daemon)
 - 2x10GbE networking (one for public, one for cluster)
- 3 physical monitors, 3 physical gateways
- Each gateway to provide three interfaces
 - S3/Swift, GridFTP (for FTS transfers), XrootD (WNs to access the object store) - last two not ready yet
 - Can provide access credentials for any interested developers



CEPH at RAL Tier-1

- XrootD/GridFTP interface is built directly on to object store
 - You can call your object “alice/foo/bar/myfile.root”, but there is no actual directory structure
 - Basic set of operations - Read, Write, Delete
 - RAL is developing an authorization plugin - very simple
 - DNs will be mapped to (a small number of) users
 - Users will be given R, RW or no access to a pool
 - 1-3 pools per VO
 - This is sufficient for ATLAS/CMS/LHCb
- RAL does not have the effort or knowledge to develop anything separate for ALICE
 - But can provide access for development/testing



CEPH at RAL Tier-1

- Monitoring with InfluxDB and Grafana
- Plans for SL7 and CEPH Jewel
- Working towards deployment for large scale science data storage



What Does This Mean for ALICE?

- ALICE D1T0 disk on CASTOR is now at retirement age
 - Would prefer to deliver new HW for ALICE in CEPH through either RAL's gridftp/xrootd interfaces or S3/Swift
- If not feasible to use CEPH, RAL will guarantee to deliver disk to ALICE in CASTOR until 30 September 2017
 - Cannot go beyond this date owing to planned re-organisation of CASTOR instances in 2017. Service will terminate promptly
 - Will attempt to continue to run existing aging disk servers. If not feasible will deploy alternative HW for CASTOR
 - Cannot guarantee I/O rates beyond load generated by CPU MoU commitment. Will not provision bandwidth for opportunistic use



What Does This Mean for ALICE?

- Situation to be reviewed in March 2017
 - Alternatives may emerge
 - Cannot afford to deploy ALICE specific solution. Nor EOS
 - UK Tier-2s cannot fill the gap either



**But ready to take (or forward)
questions!**

Thank You!



...and it's not always like today in Bergen!