UK Status and Plans

Catalin Condurache – STFC RAL

ALICE Tier-1/Tier-2 Workshop

Strasbourg, 3 May 2017

Science & Technology Facilities Council
10 Years of Impact and Inspiration

GridPP
UK Computing for Particle Physics

Content

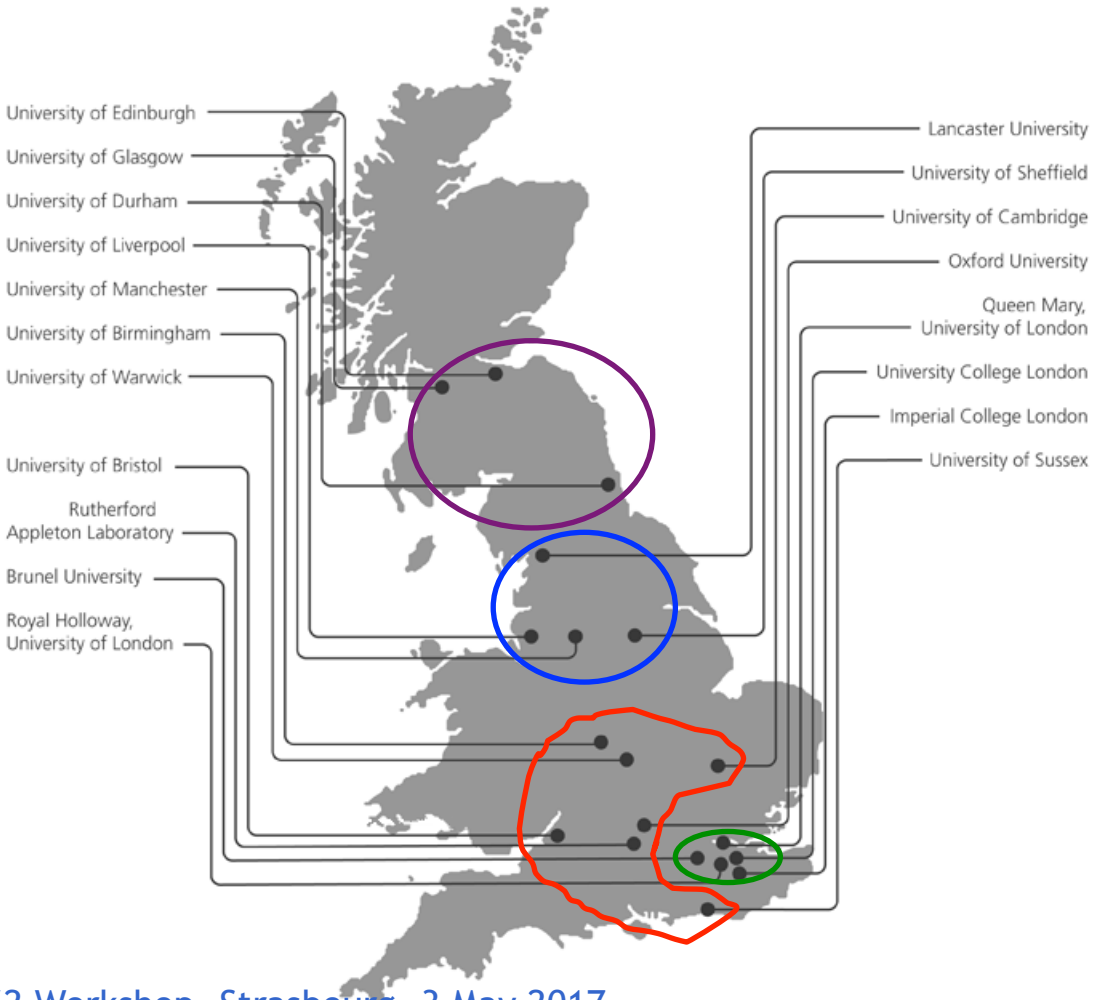
- UK GridPP Collaboration
- Since last report
- Tier-2s status and plans
  - Birmingham
  - Oxford
- RAL Tier-1 Centre
  - Components status and plans
  - ALICE highlights
- Computing centres federation
- UK funding status
- More on storage for ALICE at RAL

GridPP UK

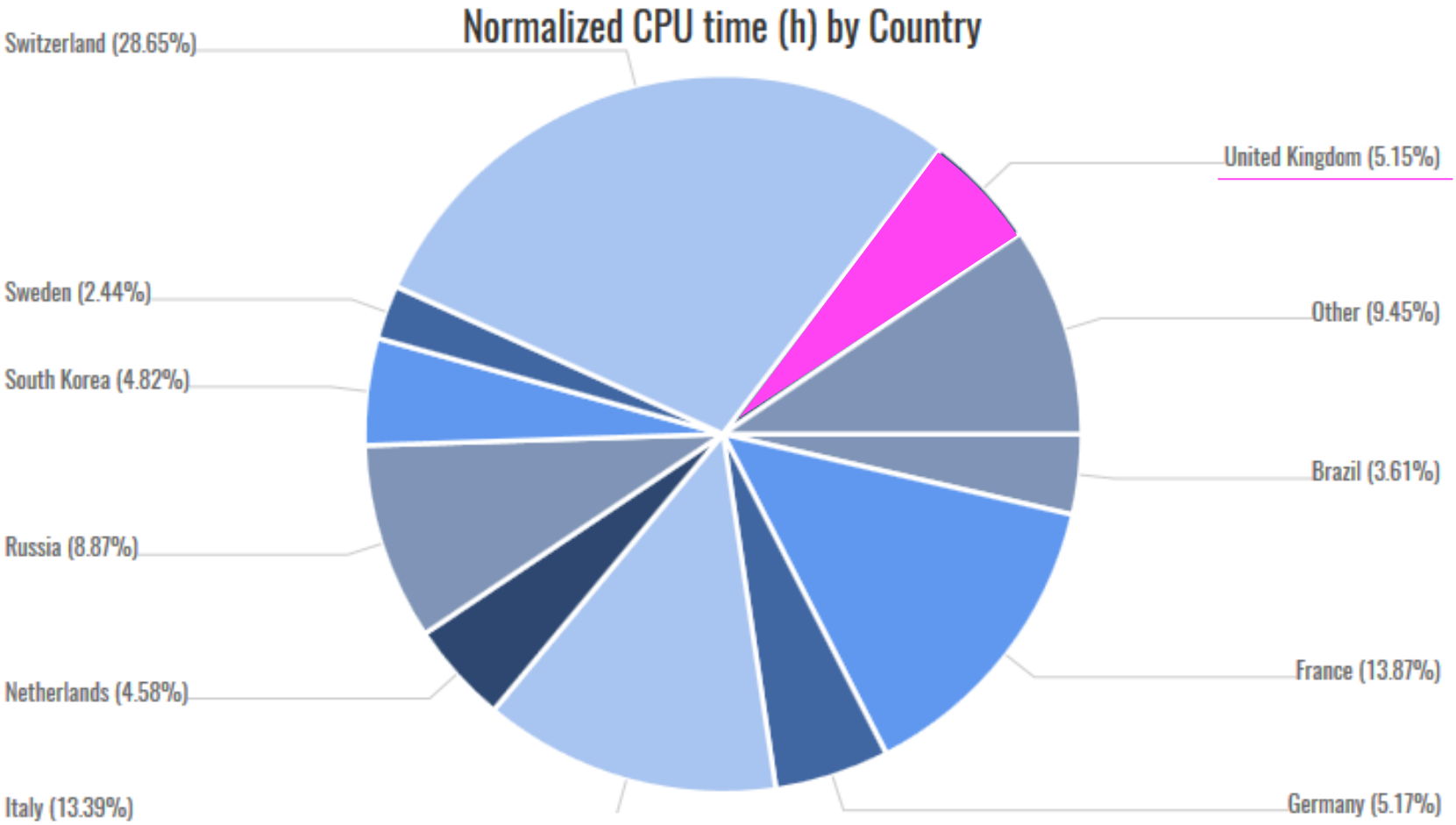
- The GridPP Collaboration is a community of particle physicists and computer scientists based in the United Kingdom and at CERN
- It consistently delivers world-class computing in support of ***all*** LHC experiments and many more user communities in a wide variety of fields

GridPP UK

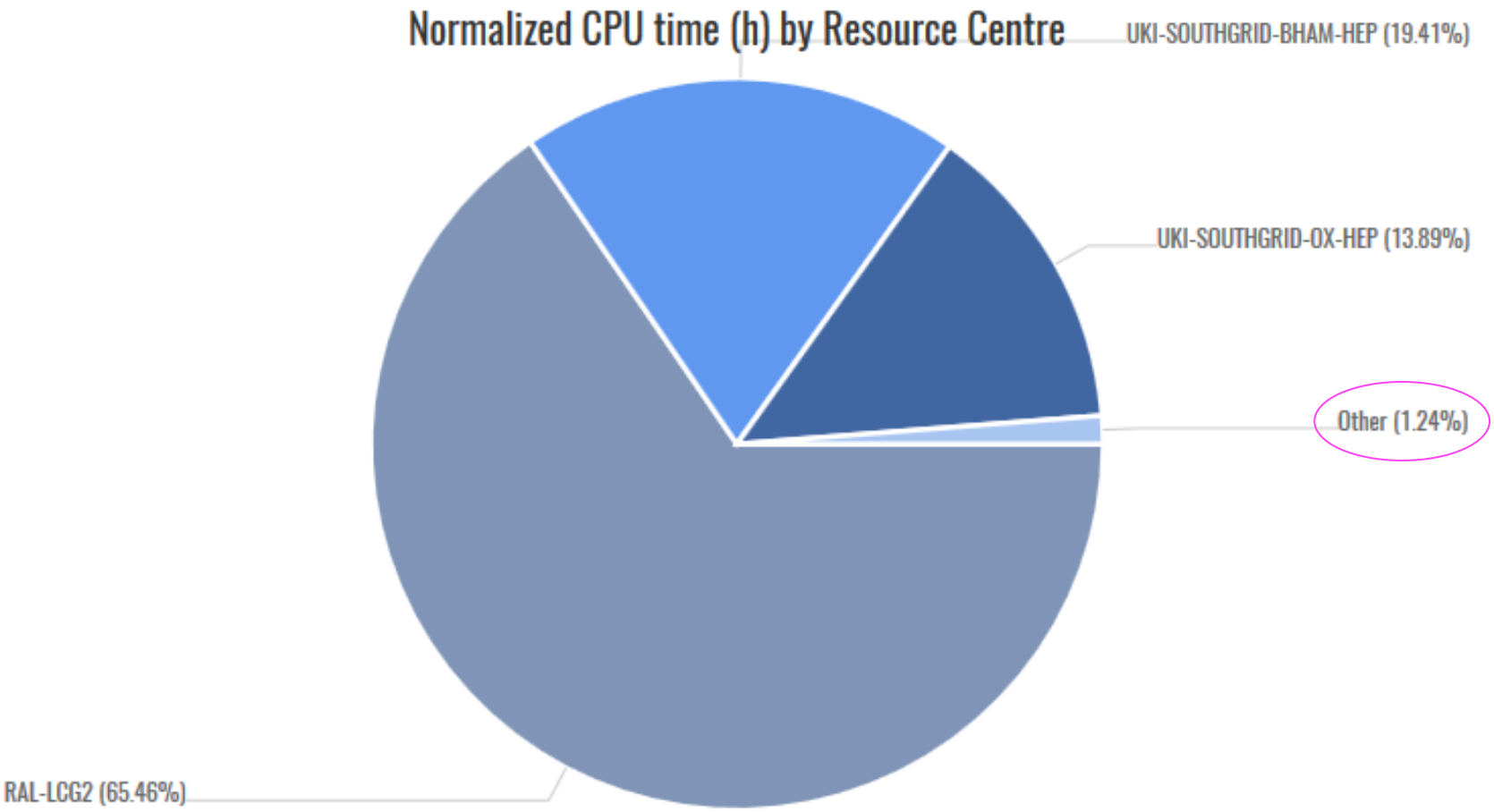
- ~10% of WLCG
- Collaborating Institutes
- ScotGrid
- NorthGrid
- SouthGrid
- LondonGrid



ALICE - CPU Accounting
Last 12 Months Worldwide



ALICE - CPU Accounting
Last 12 Months UK

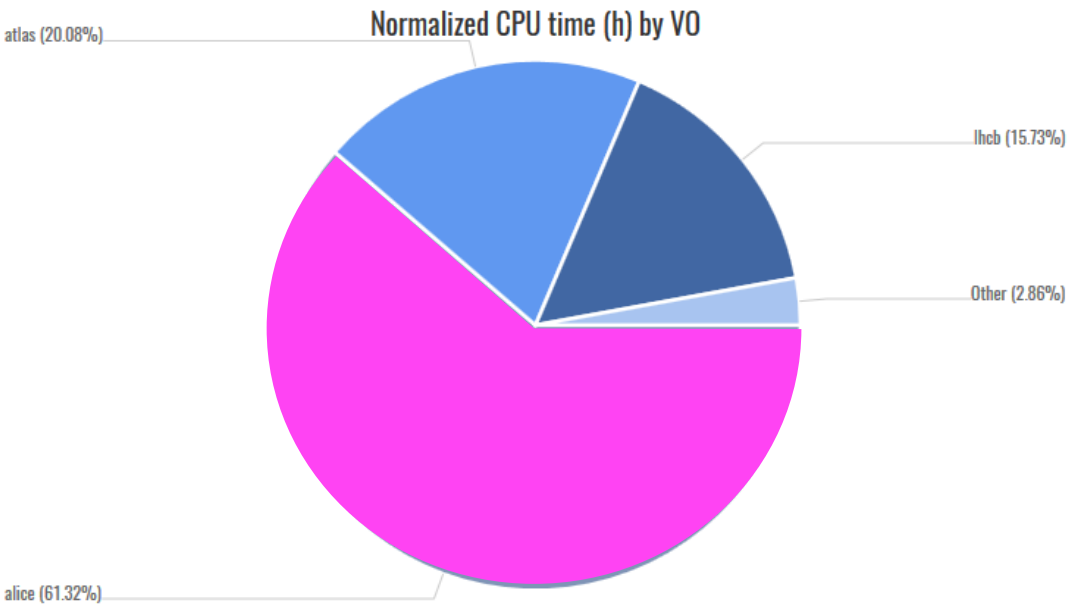


Tier-2s Status and Plans - Birmingham

- *UKI-SOUTHGRID-BHAM-HEP*
- Disk storage
  - Native XRootD for ALICE, DPM for others
  - Total 980 TB with another 200 TB being prepared
  - For ALICE
    - 418 TB (+ 200 from above)
    - 838 TB by Apr 2018 – achievable with some re-arrangements
    - looking into EOS (at Latchezar's request)
    - if successful then migration in 12 months

Tier-2s Status and Plans - Birmingham

- CPU
  - ~1500 cores providing ~17k HS06
  - Formal divide between experiments is
    - ALICE 60%
    - ATLAS 30%
    - LHCb 5%
    - Other 5%



Tier-2s Status and Plans - Birmingham

- CPU
  - ~60% of UK T2 ALICE CPU allocation

Tier-2s Status and Plans - Birmingham

- Trying to do efficiency savings
  - No migration to ARC/Condor, but…
  - **Moving workers to VAC** – ongoing change
  - Don't have to run CREAM, Torque, APEL
  - Reduces complexity of other services (Squid, BDII, Argus)
  - Overall a significant reduction in manpower required
  - Currently ~200 cores devoted to VAC (~13% of the whole site).
  - Once whole site converted => decommission CREAM, Torque, APEL (timescale ~6 months)

Tier-2s Status and Plans - Birmingham

- Also…
  - Current bandwitdh 10Gb/s
    - a second 10Gb/s line to be added soon
  - IPv6 on hold
    - lack of manpower

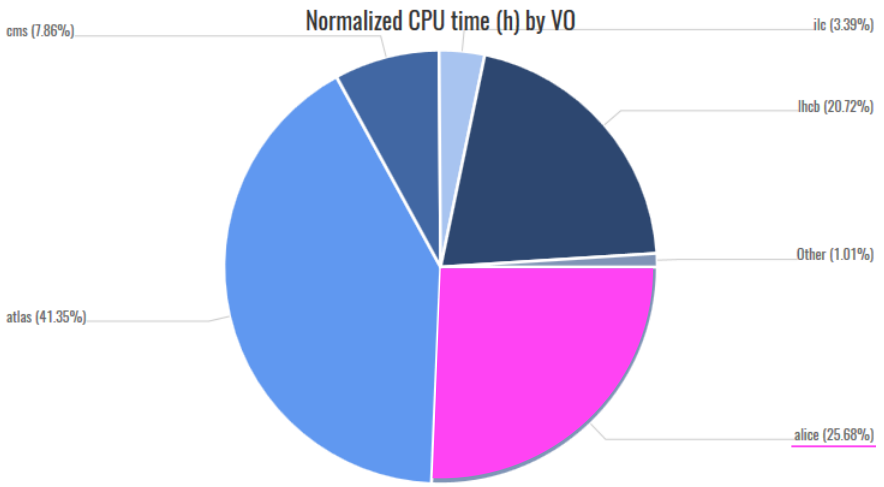
- Many thanks to Mark Slater!

Tier-2s Status and Plans - Oxford

- *UKI-SOUTHGRID-BHAM-HEP*
- NO disk storage!
- Pledged to deliver 40% CPU of UK Tier-2 ALICE requirement

Tier-2s Status and Plans - Oxford

- CPU
  - Total 2,600 CPU cores
    - ALICE frequently uses up to 800 cores
    - opportunistic use of spare cycles when primary customer (ATLAS) has not used its full quota
  - 37 Mil HEPSPEC06 Wall Clock Hours used by ALICE
  - More 2x Oxford ALICE pledge to WLCG



Tier-2s Status and Plans - Oxford

- Also…
  - The Grid Cluster runs HT Condor behind ARC-CE
  - Some ALICE jobs keeps running for 7-8 days and have to kill those jobs manually
  - “Otherwise we haven't seen any issues and ALICE operations team is very prompt in dealing with any issues at our site.”

Tier-2s Status and Plans - Oxford

- Networking
  - 10Gbs connection for Tier-2
  - Small IPv6 testbed and successfully run some jobs
  - Also IPv6 only UI to some external users for testing purpose
  - New edge switches (IT Services)
    - fully support for IPv6 in the next 6 months

- Many thanks to Kashif Mohammad!

Tier-2s Status and Plans

- Other – ALTARIA
  - Recent finding
    - GridPP38 meeting – April 2017 – "Vac, Vcycle, VMs status and plans" - Andrew McNab

    https://indico.cern.ch/event/601969/contributions/2473738/attachments/1441700/2220016/20170407-mcnab-vac-vcycle.pdf
    - ALICE Offline week – March 2017 – "Grid status" – Maarten Litmaath

    https://indico.cern.ch/event/624025/contributions/2524245/attachments/1436392/2210011/ALICE-grid-170330-v11.pdf
  - A virtual site to drive cloud resources

Tier-2s Status and Plans

Altaria

- A virtual site to drive cloud resources
- Currently being used in a proof of concept for sites that want to provide their resources via cloud instead of traditional grid mechanisms
- In particular the UK T2 sites are moving to that model
  - Manchester (first ALICE jobs since 2008)
  - Liverpool (first ALICE jobs ever?)
  - Birmingham
  - Oxford
- Cloud VMs are configured such that they connect to an HTCondor pool at CERN to which Altaria submits its jobs
  - For monitoring and accounting it may be desirable to have an HTCondor pool per site, hosted on its own VOBOX
- The VMs are managed by the sites, AliEn just sees resources appear as if they were WN job slots
  - Managed e.g. through Vac or Vcycle

CERN

7

Tier-2s Status and Plans

NGI_UK — Normalized CPU time (h) by Resource Centre and Month (Custom VOs) ← ALICE

Resource Centre	Nov 2016	Dec 2016	Jan 2017	Feb 2017	Mar 2017	Total	Percent
RAL-LCG2	11,504,894	5,782,593	18,433,587	10,178,165	13,305,329	176,051,002	65.46%
UKI-NORTHGRID-LIV-HEP	0	6	721,208	440,283	787,977	1,949,474	0.72%
UKI-NORTHGRID-MAN-HEP	0	627,998	374,539	218,465	168,616	1,389,617	0.52%
UKI-SOUTHGRID-BHAM-HEP	4,807,589	5,783,706	4,957,707	3,889,977	3,207,369	52,207,618	19.41%
UKI-SOUTHGRID-OX-HEP	810,889	2,862,695	3,678,986	3,225,300	4,530,195	37,359,685	13.89%
Total	17,123,372	15,056,998	28,166,027	17,952,189	21,999,486	268,957,396	
Percent	6.37%	5.60%	10.47%	6.67%	8.18%		

1 - 5 of 5 results ‹ 1 › Number of rows per page 30

Deployment by site and experiment

		ATLAS	ALICE	LHCb	GridPP DIRAC
Vac	Birmingham	✔	✔	✔	
	Liverpool	✔	✔	✔	✔
	Manchester	✔	✔	✔	✔
	Oxford	✔	✔	✔	✔

UK LHC Tier-1

- Hosted and run by STFC Rutherford Appleton Laboratory
- 15 miles south of Oxford on Harwell Campus

RAL Tier-1 Centre

- *RAL-LCG2*
- Hardware
  - CPU: ~240k HS06 (~24k cores) – from 14.8k cores
    - FY16/17: additional ~19.6k HS06, 1920 cores
  - Storage:
    - ~16.5PB disk useable in Castor
    - 13.3PB raw for Ceph
    - FY16/17: additional 6720TB raw (~4.9PB configured) for Ceph
  - Tape: 10k slot SL8500
    - 50PB – T10KD

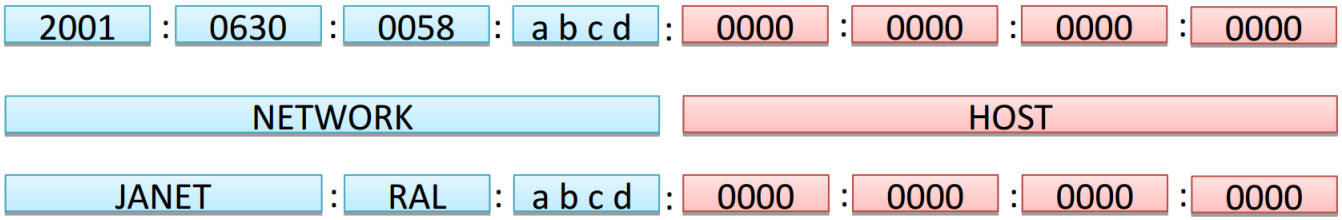
RAL Tier-1 Centre

- Networking
  - OPN link to be increased to 30Gb/s
  - No intention at this time to join LHCONE
    - JANET is providing the service required for GridPP project
    - some recent successful tests at Imperial
    - lack of manpower (and funds) for progressing tests at RAL

RAL Tier-1 Centre

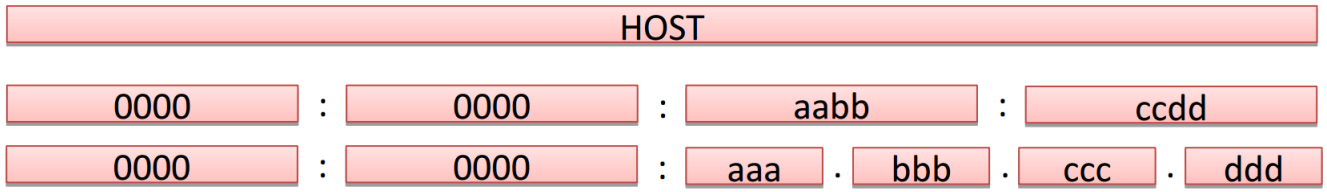
- IPv6
  - Now available on Tier-1 network
    - significant effort recently at RAL and Tier-1 level

    https://indico.cern.ch/event/595396/contributions/2558578/attachments/1448031/2231673/jrha-hepix-2017-ipv6.pdf
  - STFC addressing scheme agreed
    - each STFC site allocated an IPv6 /48
    - each project allocated one or more IPv6 /64



RAL Tier-1 Centre

- IPv6
  - Tier1 addressing scheme
    - all hosts will be dual-stack
    - map all existing IPv4 addresses (RFC2374 style)
      - allocate addresses automatically with Quattor



In hex notation
::82F6:B43C
Or mixed notation
::130.246.180.60

RAL Tier-1 Centre

- IPv6 Plans - Services
  - By Sep'17
    - FTS (extensively tested at other sites)
    - CVMFS Stratum 1 (tested at other sites)
  - By Dec'17
    - Squids (tested internally)
    - Frontier (ATLAS testing now)
    - GOCDB (ran test instance on previous test-bed)
    - SCD Private Cloud (in-use by power users now)
  - By Apr'18
    - All hosts dual-stack by default

RAL Tier-1 Centre

- IPv6 Plans - Storage
  - CASTOR
    - Will not implement IPv6
    - Disk storage is migrating to Echo
  - Echo (Ceph)
    - Production endpoint currently IPv4 only
      - focused on achieving production service for Echo
    - Currently testing dual stack gateways
      - aim for full production IPv6 access by June 17
    - This will meet the April 2018 Tier-1 storage requirement

RAL Tier-1 Centre

- Batch farm
  - ~24000 cores
  - Around 50% of the farm migrated to SL7
  - Using HTCondor Docker universe to run jobs in containers
- Load balancers
  - Using a pair of VMs running HAProxy and Keepalived as a highly-available load balancer
  - Has been used in front of FTS3 for over a year now
  - Other services now using them include Top BDII, Site BDII, Dynafed, Argus

RAL Tier-1 Centre

- Containers
  - Investigating Kubernetes as a means of providing portability between on-premises resources and multiple public clouds
- Monitoring
  - Ganglia still exists, but usage is slowly fading away
  - Instead using Telegraf (metrics collector), InfluxDB (time series database), Grafana (visualisation)
  - Used by many grid services, batch system, Ceph, Windows HyperV

RAL Tier-1 Centre - Storage

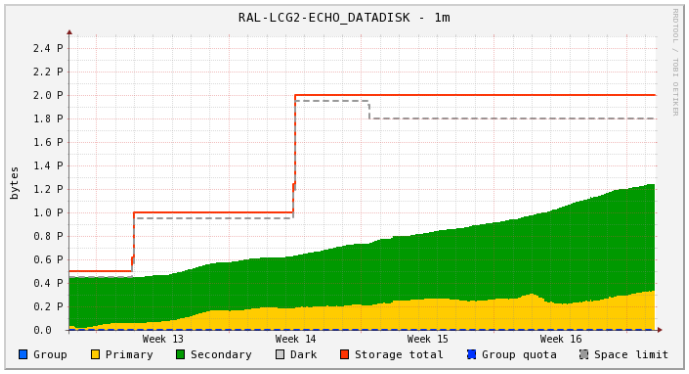
- Castor
  - Updated to v2.1.15-20 in January
  - Continuing to take production data from LHC
    - ~130TB/day
  - Update to 2.1.16-13 shortly
  - SRM upgraded to 2.1.16-10, but rolled back for LHCb
    - Performance problems
- Castor Tape
  - To be discontinued by CERN ~mid 2019
  - Looking at CTA, HPSS…

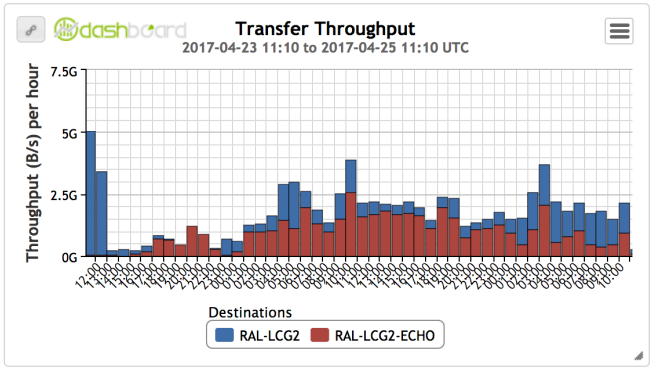
RAL Tier-1 Centre - Storage

- Echo
  - RAL's new Ceph based storage
    - entered production in March 2017 (Kraken release)
  - Accepting production data from LHC VOs
    - GridFTP and XRootD supported as production I/O protocols
    - also S3/Swift API access for all users
  - 7.1PB of WLCG pledge to be provided by Echo this year

RAL Tier-1 Centre - Storage

- Echo
  - ATLAS is main user so far
    - already migrated 1.2PB of data for ATLAS
    - average throughput 500 MB/s
    - primarily use GridFTP plugin
  - CMS and LHCb also significant progress on testing





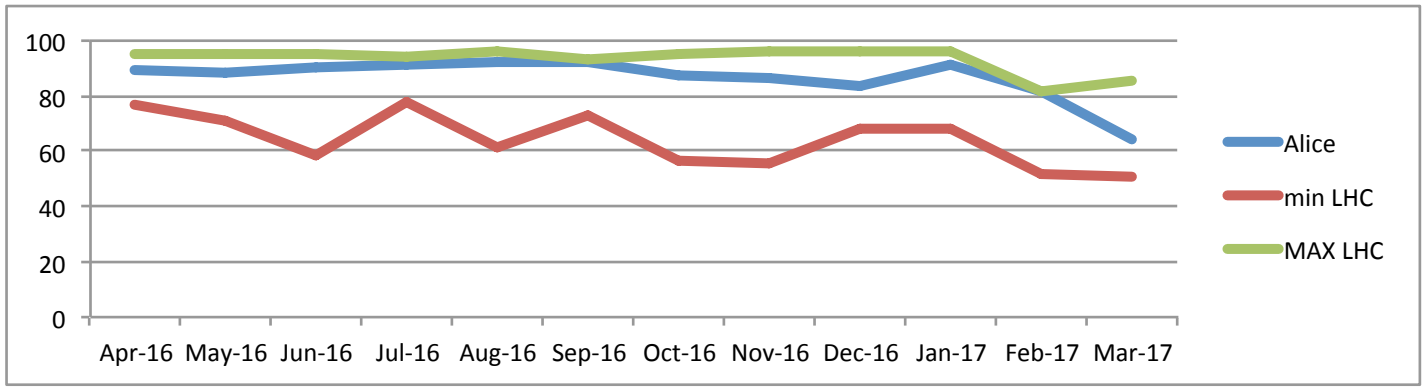
RAL Tier-1 Centre – Misc.

- CernVM-FS
  - 'secure' repositories possible - based on X.509
- Plan to move from Hyper-V to VMware
  - Consolidation of resources after rearrangement of divisions
- New chillers for the machine room – Mar/Apr 2017
  - Replacement under STFC spend-to-save initiative
  - PUE reduced from >1.64 to ~1.35

- Many thanks to Martin Bly, James Adams, Alastair Dewhurst + others!

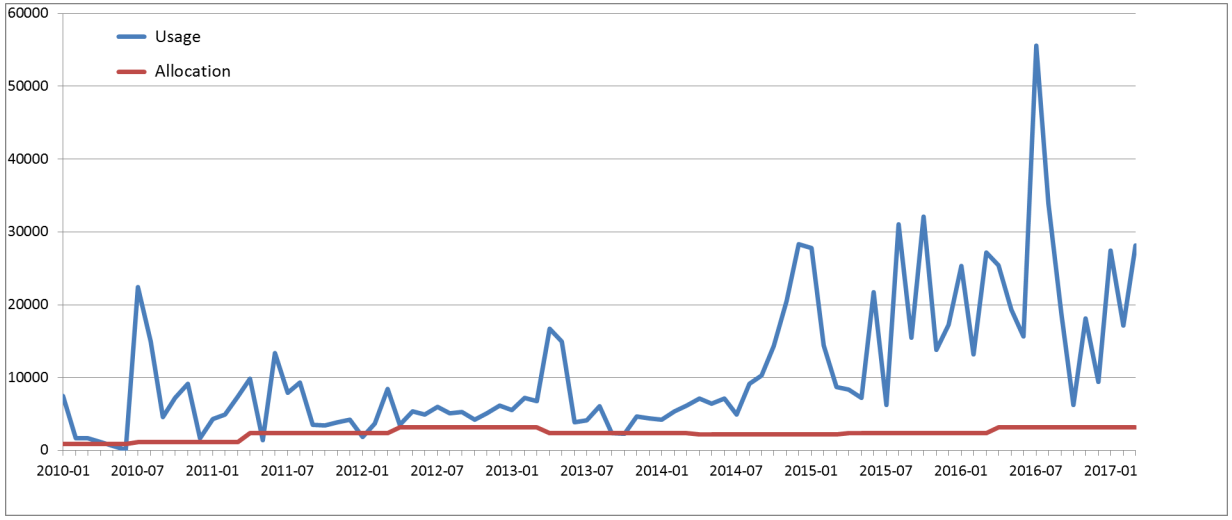
RAL Tier-1 – More on ALICE

- CPU fairshare – 1.96% (4840 HS06) in 2017
  - 1.865% (2400 HS06) in 2015
  - 1.27% (3140 HS06) in 2016
- NO limit on opportunistic use of spare cycles jobs
  - Since June 2015
- CPU efficiencies – >80% average for ALICE



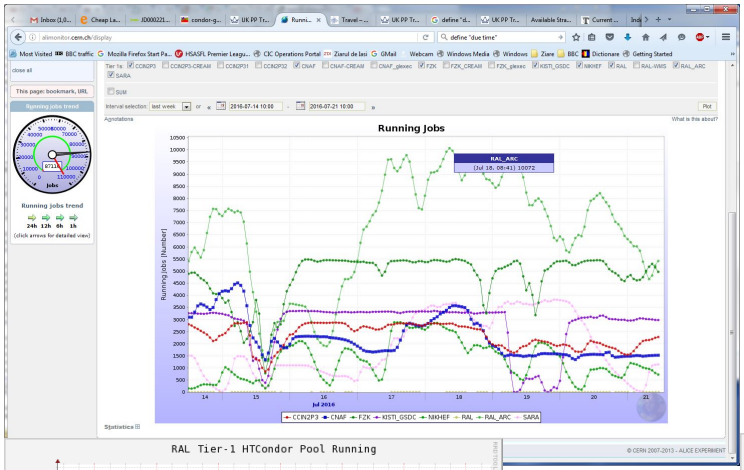
RAL Tier-1 - More on ALICE

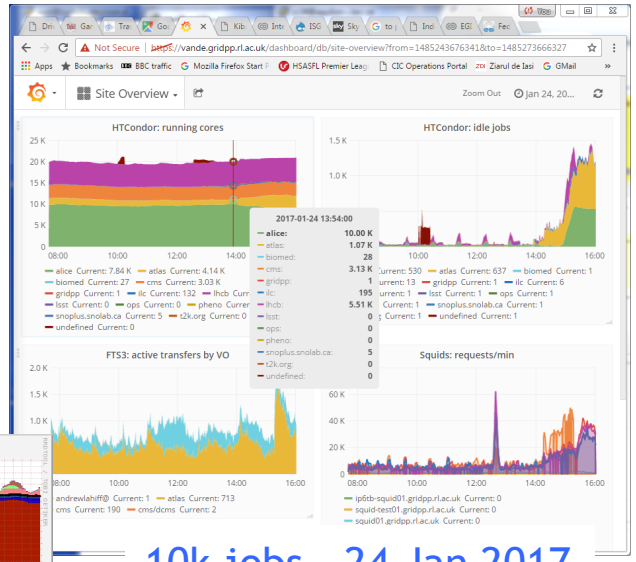
- Monthly CPU allocation and usage (HS06) for ALICE since 2010
- Usage is consistently high ~800% pledge
- 29K HS06 CPU used in April 2017



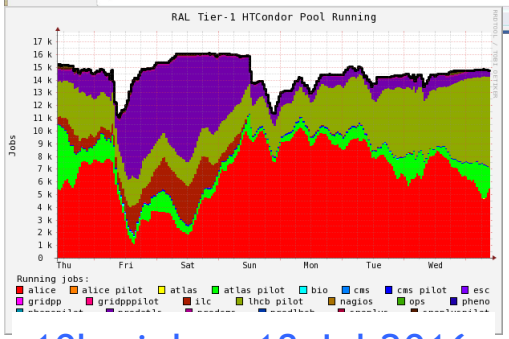
RAL Tier-1 – More on ALICE

- Max of 12367 running jobs on 11 April 2017

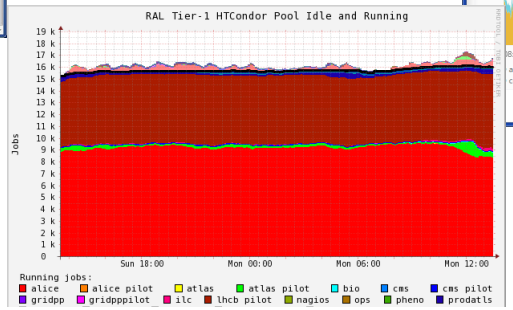




10k jobs – 24 Jan 2017



10k+ jobs – 18 Jul 2016



9k jobs constant – 15 Aug 2016

RAL Tier-1 – More on ALICE

- Disk storage
  - Currently 505TB disk allocated
  - 544.5TB deployed (incl disk tape buffer)
  - Only 199TB utilised
  - It cannot be increased
- Tape storage
  - 870TB allocated – over initial pledge
  - 882TB used in April

Computing Centres Federation

- NO plans to federate Birmigham and RAL storage
- Some UK Tier2s are looking at pooling their resources
  - But very little discussion on how this will happen
- DynaFed for Echo in our attention
  - Possibility of federating the storage with other UK sites
  - Access via https
  - How does it compare with ALICE XRootD data model?

UK Funding Status

- The current funding grant for LHC computing in the UK awarded ~95% of flat cash
- The UK pledged to meet the original 2017 request and 60% of the **additional** LHC requests for 2017
- Tier-1 – no resource growth foreseen in 2017
- Tier-2s trying to make efficiency savings
  https://indico.cern.ch/event/601969/contributions/2473732/attachments/1441962/2220372/GridPP_Meeting_Apr2017.pdf

More on Storage for ALICE at RAL

XRootD plugin

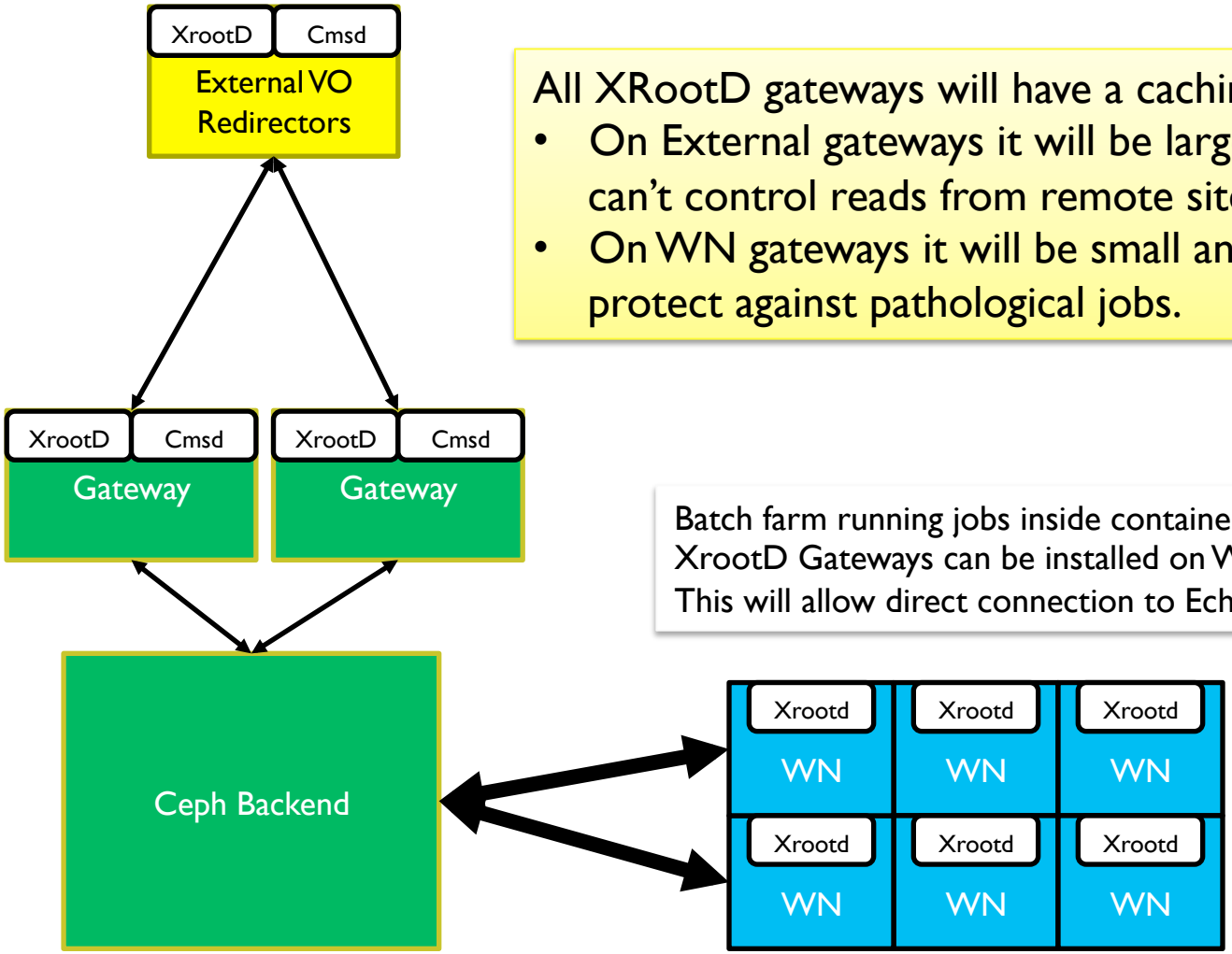
- XRootD plugin was developed by CERN
  - They have a very specific use case.
  - We have encountered (and fixed) many issues.
- Many problems found in components we assumed would be entirely independent from XrdCeph:
  - Proxy Cache both memory and disk didn't work.
  - Redirection didn't work.
  - N2N component didn't work.
- Currently working with CMS to fix bugs and optimize XRootD performance.
- Would greatly appreciate assistance in testing ALICE XRootD functionality against Ceph.

echo

Erasure Coding

- Echo use Erasure Coding to store data.
  - Data is split into 64MB stripes.
  - Each 64MB chunk is split into 8 x 8MB chunks from which a further 3 x 8MB parity chunks are calculated.
  - Each of the 11 x 8MB chunks is written to a different storage node.
- Reading even 1 byte of data requires the reconstruction of (at least) one 64MB stripe.
- Streaming data to jobs works most efficiently if we can get job to requests data in an integer multiple of 64MB.
  - XRootD proxy caches are being setup to protect against jobs requesting lots of small amounts of data.

XRootD Architecture



All XRootD gateways will have a caching proxy:

- On External gateways it will be large as we can't control reads from remote sites.
- On WN gateways it will be small and used to protect against pathological jobs.

Batch farm running jobs inside containers.
XrootD Gateways can be installed on WN.
This will allow direct connection to Echo.

Echo Timeline for ALICE

- September 2017 : Start testing ALICE XRootD on Echo.
  - Assuming work with CMS completed by this point.
  - RAL person comes to CERN for a couple of days to work with ALICE storage expert?
- April – September 2018: Migrate ALICE to Echo.
  - Not discussed any data migration possibilities yet although ~500TB is not a huge amount.
- April 2019: ALICE hardware in Castor reaches end of life.
  - Tier 1 cannot afford to run separate storage service for ALICE.

echo

Summary on Storage for ALICE at RAL

- We (RAL) think it should be possible to get XRootD for ALICE working with Echo
  - But we believe there are bugs that need fixes, so we **WILL** need your (ALICE) help
- Castor for disk will disappear and if Echo is not working by then => ALICE will just be without a disk endpoint
- We do not want this to happen
  - But we do not have a plan B!
- Also RAL is happy to provide S3 access is ALICE want

Merçi!

Des questions?