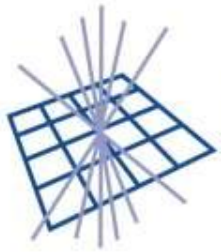
UK Status and Plans

Catalin Condurache

ALICE T1/T2 Workshop, CC-IN2P3, 4 June 2013

Science & Technology Facilities Council
Rutherford Appleton Laboratory



GridPP
UK Computing for Particle Physics

Content

- UK GridPP Collaboration
- Tier2s Status and Plans
- RAL Tier-1 Centre
  - Status
  - Plans
  - Storage – next generation

UK Grid Collaboration

- GridPP is a collaboration of 19 UK universities + RAL + CERN with the primary goal of providing computing resources to LHC particle physics experiments

GridPP resources

Site	CPU kSI2k	Storage TB
RAL	22,193	11,800 (+10,645)
Edinburgh	9,015	355
Glasgow	7,912	1,313
Queen Mary, London	7,513	1,697
Imperial College, London	6,973	2.004
Lancaster	6,434	970
Manchester	5,728	882
Sheffield	4,936	360
RALPP	4,914	1,587
Royal Holloway, London	3,704	728
Brunel, London	3,491	593
Oxford	2,948	669
Liverpool	2,782	544
Birmingham	1,572	315
Cambridge	646	278
Sussex	599	54
Bristol	562	122
University College, London	502	160
EFDA Jet	332	2
Durham		53
TOTAL	92,778	24,562 (+10,645)

from GStat

GridPP resources

Site	CPU kSI2k	Storage TB
RAL	22,193	11,800 (+10,645)
Edinburgh	9,015	355
Glasgow	7,912	1,313
Queen Mary, London	7,513	1,697
Imperial College, London	6,973	2.004
Lancaster	6,434	970
Manchester	5,728	882
Sheffield	4,936	360
RALPP	4,914	1,587
Royal Holloway, London	3,704	728
Brunel, London	3,491	593
Oxford	2,948	669
Liverpool	2,782	544
Birmingham	1,572	315
Cambridge	646	278
Sussex	599	54
Bristol	562	122
University College, London	502	160
EFDA Jet	332	2
Durham		53
TOTAL	92,778	24,562 (+10,645)

from GStat

Tier-2s Status and Plans

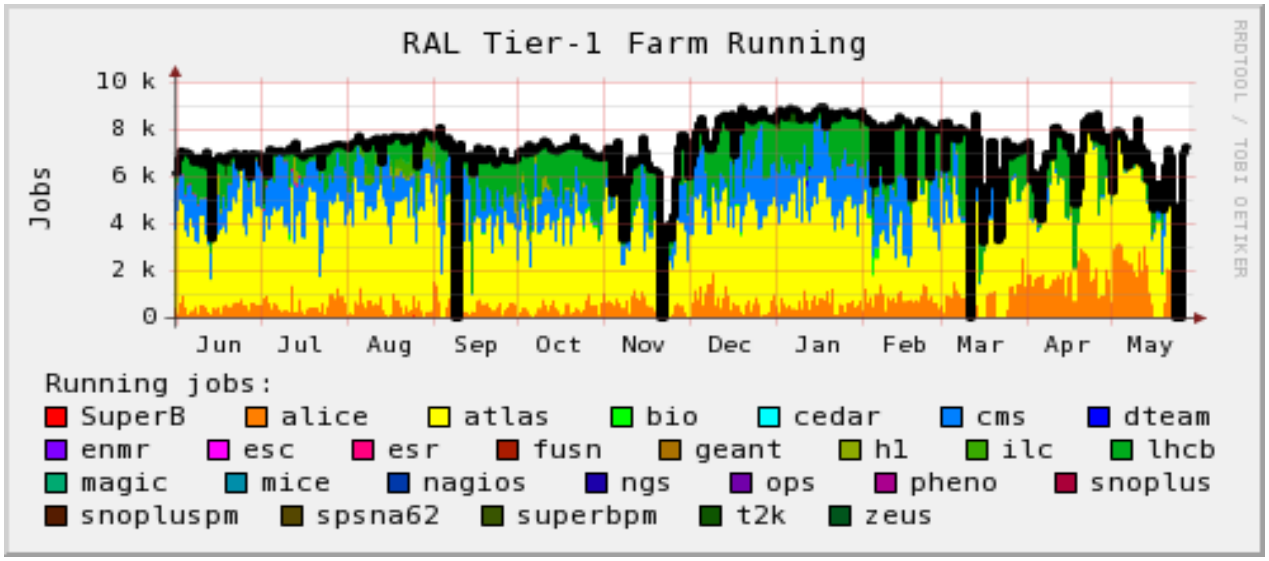
- *'jobs done'* ratio last year
  - 72% Birmingham, 28% Oxford
- Birmingham
  - 110TB storage available – 89% used
  - "...jobs running through well without any problems..."
  - VOBOX – gLite – to be migrated soon to WLCG flavour
  - 60% overall fairshare (740 job slots out of 816)
    - ATLAS 30%, LHCb 5%, others 5%

Tier-2s Status and Plans

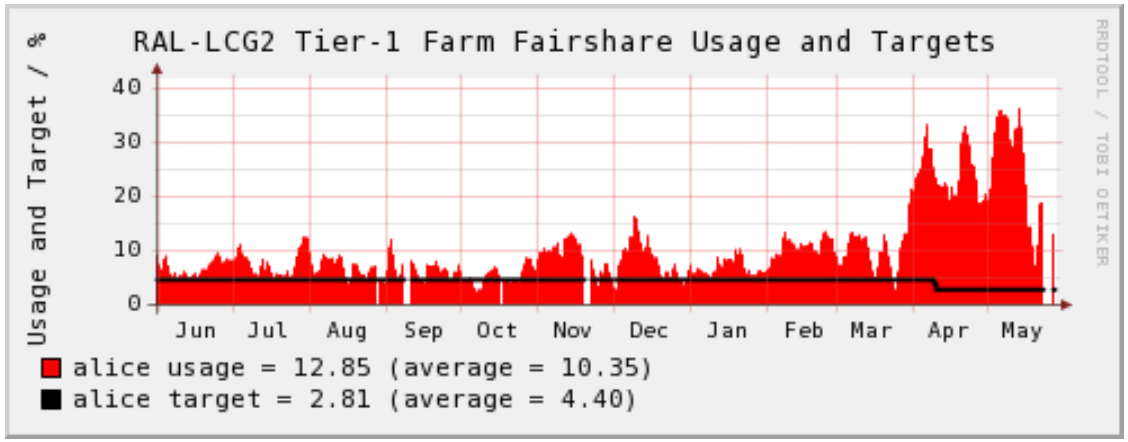
- Oxford
  - "...Alice jobs run very smoothly..."
  - limited to 150 job slots
    - out of 1392
    - manual increase possible
  - VOBOX – gLite
    - farm to be moved to SL6 in few months
    - will be migrated to WLCG flavour at that time
  - no storage provided

RAL Tier-1 Centre

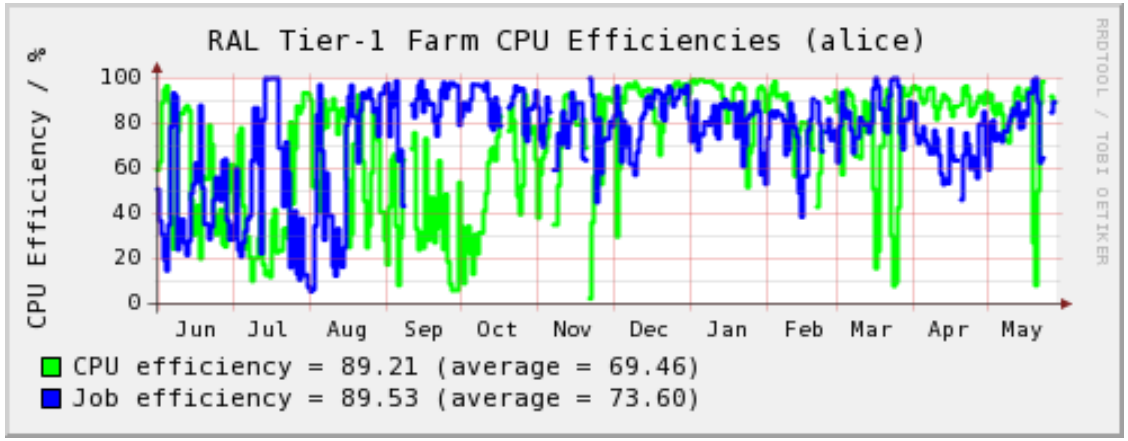
- Currently 9472 job slots, 11PB disk, 10PB tape
- Targets for availability, occupancy, efficiency
- ALICE can use spare cycles - up to ~2500 jobs



RAL Tier-1 Centre



- CPU fairshare – 2.81%
- max 3000 jobs
- but 4500 when farm empty



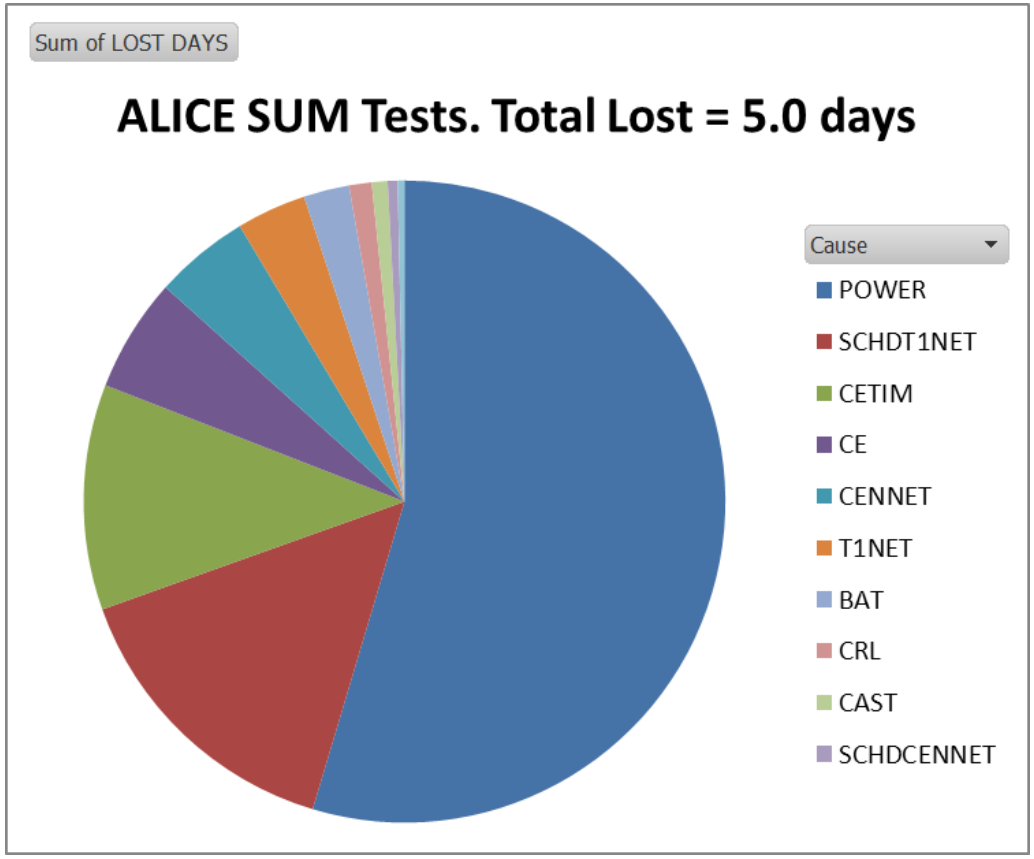
Large number of ALICE jobs killed because exceeded resources (i.e. 3GB+ RAM)

RAL Tier-1 Centre

- Two major power incidents last 12 months
- 7 November 2012
  - RAL Site power outage (circuit breaker trip)
  - Diesel generator did not start (was tested, but not with load!)
  - 26 ½ hours downtime
- 20 November 2012
  - UPS Room Over-voltage (smoke, fire alarm…)
  - Major damage at Tier-1 (tape robots, switches, PDUs, no cooling)
  - 28 ½ hours downtime

RAL Tier-1 Centre

Availability data from September 2012 to March 2013 (212 days)



RAL Tier-1 Plans

- "…we are happy with ALICE…"
- WLCG VOBOX in production
- ALICE resource allocations – until April 2014
  - 284 TB disk
  - 420 TB tape
  - 2400 HEPSPEC06 CPU
- Can ALICE submit to ARC-CE?
- Can ALICE use *'generic queue'?*
  - Specific memory requirements, number of cores

RAL Tier-1 Plans

- Infrastructure
  - Electrical safety check – requires significant downtime (2 days)
  - Will accommodate also intervention on "Essential Power Board" and remedial work on 3 (out of 4) transformers
- Networking
  - Single link to UKLight Router to be restored as paired link
    - 2x10 Gbit/s
  - New mesh network core and routing layer
    - 160 Gb/s core network, 40 Gb/s link to SJ5/LHCOPN, 10 Gb/s to site
    - Resilience and extensibility
    - Testing phase on-going, infrastructure being laid in

RAL Tier-1 Plans

- Fabric
  - Tier-1 resilience centre – ATLAS building
    - Resilience for services via virtualization
    - Complete move of Castor DB standby systems
  - Complete DNS deployment and Quattor 'conversion'
  - Regular new batches of storage and CPU hardware
  - Fill in any gaps in resilience
    - Multiple links to individual systems

RAL Tier-1 Plans

- Grid Services
  - Complete SL6 migration (incl WNs)
  - Move to replace batch system (MAUI -> Condor, SLURM)
    - Possible look into ARC-CE (can ALICE submit to ARC-CE?)
  - Track EMI/UMD version of grid services
  - Virtualization
    - Shared storage coming online
    - Replication between buildings
    - Investigations to make full use of that
  - Distribute services
    - BDIIs, CEs, FTS spread between locations

RAL Tier-1 Plans

- Cloud
  - Departmental cloud well proven - ~300 cores, 90-95% use
  - Storage – small CEPH cluster to be deployed
  - Active use cases – internal development & testbeds
  - Developing use cases – STFC RAL (ISIS, RAL Space), also EGI, GridPP, WLCG cloud work
  - What would take to consider it solid enough for services now on Hyper-V?

RAL Tier-1 Plans

- Databases
  - HW refresh for non-Castor systems (LFC, FTS, 3D)
    - Requirements established, working on implementation plan
  - Plan for Castor DB to 2019
  - Plan for increased use of MySQL (LFC, FTS)

- Castor
  - v2.1.13 testing completed – Tier-1 upgrade – June/July
  - T10KD support
  - IPv6 support via xrootd (next Castor version will support IPv6)
  - High Oracle costs alleviated by new Oracle/CERN pricing model

RAL Tier-1 Storage – Next Generation

Why?

- Will CASTOR be able to sustain transactions rates during Run 2 (especially load for disk-only analysis)?
  - Unknown but there is no indication that it won't
- How will CASTOR cope as spindle count and disk server count decrease?
  - Need to look at this; impacted by hardware and firmware technology changes
- Should we continue to rely on a product that is not widely used outside HEP?
  - But don't do something just because everyone else is doing it

RAL Tier-1 Storage – Next Generation
Evaluation

- If need it tomorrow
  - dCache is probably only viable solution
    - Shared risk since widely deployed
    - Increase in staff effort (storage & database) though
- More strategic view
  - Other technologies are maturing
  - CEPH, LUSTRE – most promising
    - HDFS not sure it meets needs for analysis
    - "largest CEPH deployment (4PB) takes 0.5 FTE to run"
    - DPM, dCache suffer from domain specificity (HEP)

RAL Tier-1 Storage – Next Generation
Planning for the future

- No immediate need to move from Castor
  - But need to prepare to do so
- Set up a dedicated Castor-TAPE instance
  - Gives better disk/tape separation
  - Less over-provisioning of disk-cache in front of tape
  - Possibly can run on mostly virtual infrastructure
- Focus on most promising solutions and gaining experience
  - CEPH – used for SCD cloud storage
  - LUSTRE to be installed at STFC
  - Revisit EOS? - feedback from ALICE?
- When confident, deploy 1PB, ask VOs to evaluate

Any (other) Questions?

Thank You!