LHCb in GridPP5

Andrew McNab
University of Manchester

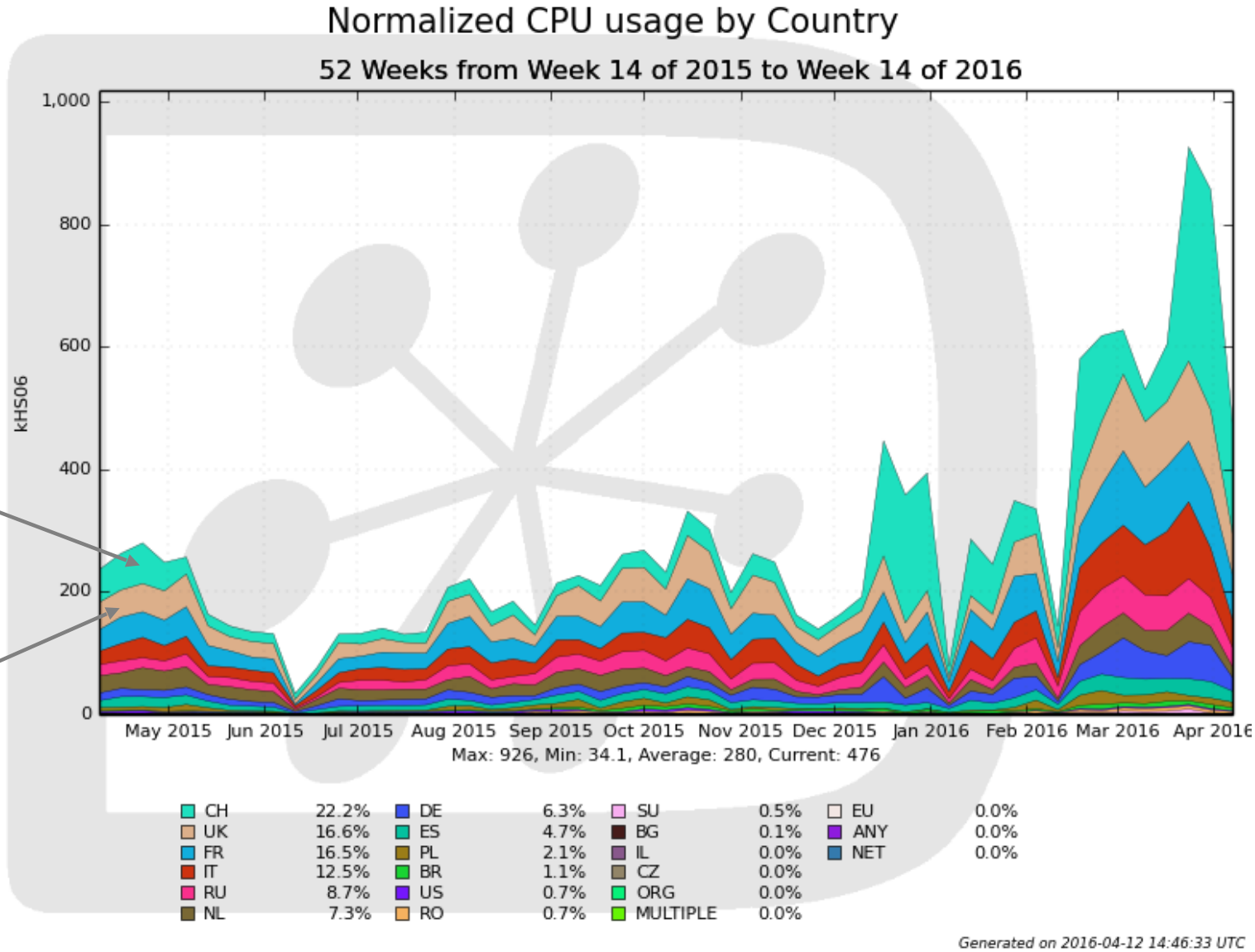
Overview

- Current state
- Data storage strategy for Run 2
- T2-D -> Tier-2A + SEs
- Compute and storage protocols
- MJF, IS evolution, multiprocessor jobs
- Remote data access
- GPUs
- Computing upgrade

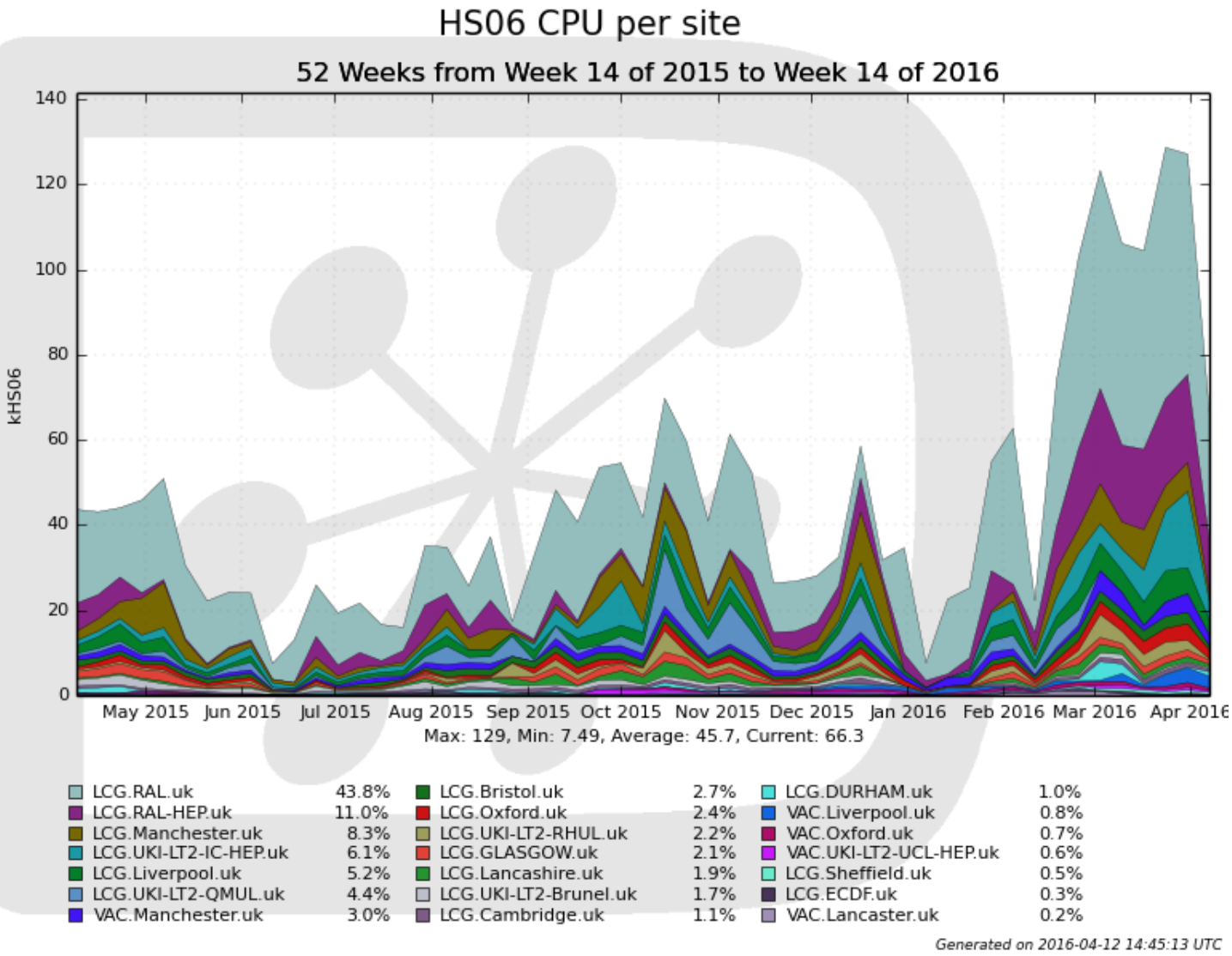
CPU power for LHCb jobs by country

.ch is mostly CERN Tier-0 and the LHCb HLT farm, which is enabled for offline work whenever possible

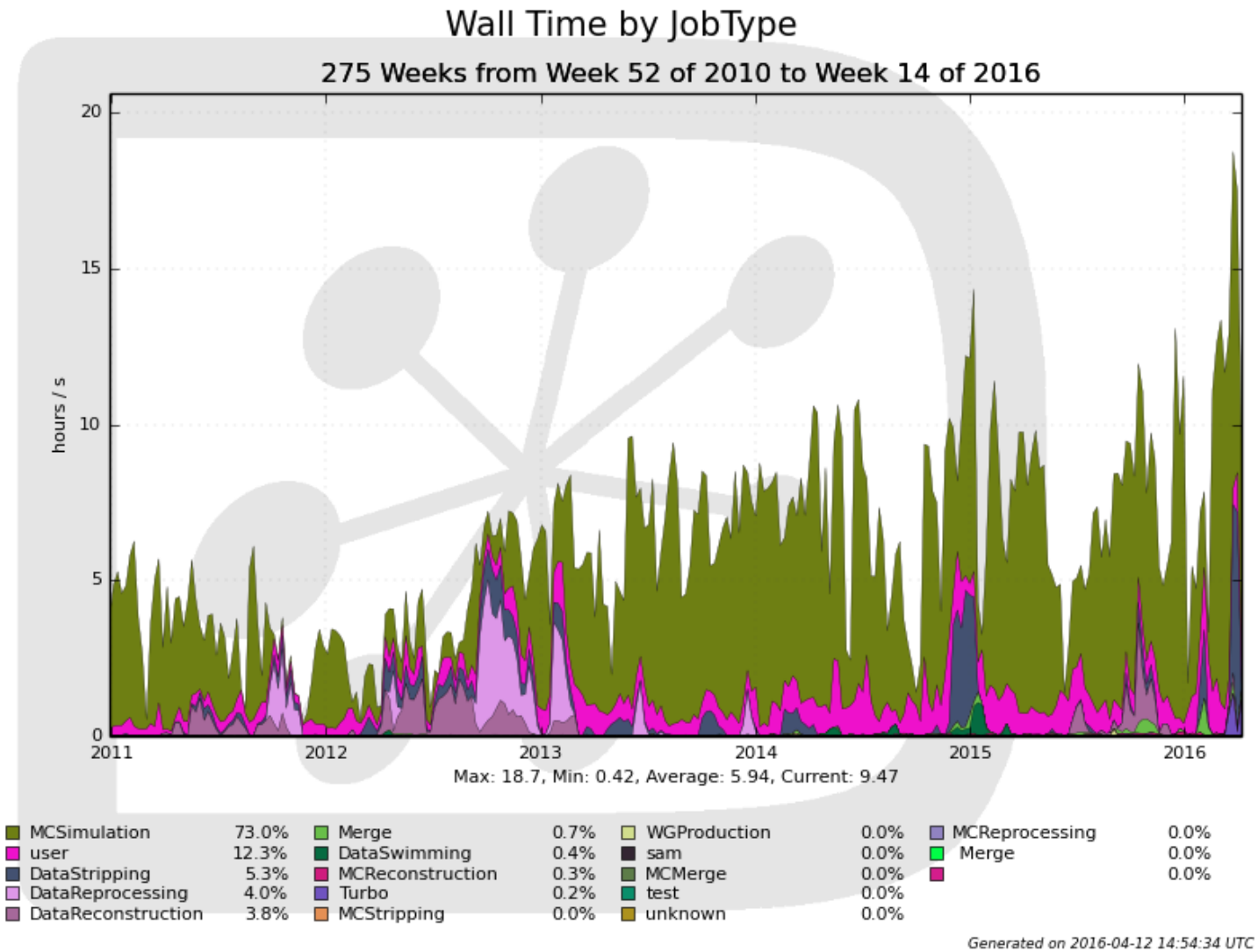
.uk is the largest single country at 16.6% (just)



CPU power for LHCb jobs by UK site



JobTypes since 2011



Lots of Monte Carlo.
This will increase absolutely and as a share.

Increasing user analysis.

Prompt reconstruction during data taking.

Well defined reprocessing during planned campaigns.

Avoiding reprocessing in Run 2

- During LS1, major redesign of LHCb HLT system
  - HLT1 (displaced vertices) run in real time
  - HLT2 (physics selections) deferred by several hours
    - Run continuous calibration in the Online farm to allow use of calibrated PID information in HLT2 selections
- Automated validation of online calibration for use offline
  - Includes validation of alignment
  - Removes need for “first pass” reconstruction
- Green light from validation triggers ‘final’ reconstruction
  - Foresee up to two weeks’ delay to allow correction of any problems flagged by automatic validation
- **So no end of year reprocessing campaigns - just restripping**
  - ~2017 *may* need to park some data if insufficient resources to do reco

Protocols

- LHCb aims to access resources in the formats sites prefer
  - but doesn't want a proliferation of new interfaces
- For job execution:
  - Pilot jobs to CREAM, ARC, or HTCondorCE
    - All of our CERN jobs will soon be via HTCondorCE
  - "Vacuum platform" VMs
- For storage:
  - We currently require SRM - but this will be dropped
  - We require xrootd (or POSIX) file access from jobs
  - We have looked at WebDAV access/federation

T2-D storage

- 12 sites with another one, Liverpool, funded by GridPP5
- 2.5PB available now, and another 600TB agreed
- 2016 LHCb Tier-2 requirement is 2.8PB. In 2017, 3.8PB
- Has achieved stated aim of providing another “Tier-1” amount of storage
- Working well, but some labour intensive periods with specific sites
- Problems tend to be associated with sites that are new to hosting data for WLCG, running DPM/dCache etc
- In the future, Monte Carlo simulation will increasingly become a bottleneck
- We will continue to welcome T2-D storage in line with our stated requirements, but we don’t want to artificially favour Tier-2 storage over CPU for simulation

Tier-2A + Storage Element concept

- Basic idea is to split T2-D into validated LHCb Tier-2A sites, hosting zero or more validated LHCb Storage Elements
- “A” is for analysis jobs, and these sites are reliable enough to run user jobs on, either
  - accessing local data on a Storage Element
  - accessing remote data if they have invested in suitable network
  - running CPU intensive user jobs (eg big fitting scenarios, exploring a parameter space)
- They’re also good Tier-2 sites to use for centrally managed data processing
- And will still be running Monte Carlo of course
- LHCb updating criteria note/Twiki: in particular, sites will need 300TB to go live (after initial testing) to favour sites with storage experience

WLCG developments

- LHCb supports rollout of Machine/Job Features: uniform way of notifying job about HS06, cpu, time limits, memory etc
  - Technical Note published: HSF-TN-2016-02
  - Revised/complete Torque/PBS implementation. HTCondor next
- LHCb supports Information Systems Evolution (simplification)
  - We would be happy to work without BDII
  - We can operate with the proposed additions to GOCDB about queues etc
- We're aiming to run multiprocessor (8-way) jobs and VMs soon
  - Initially with 8 payloads per pilot job or VM
  - Our interruptible MC jobs should help using up capacity during draining when going from single to 8-processor slots

Remote data access

- DIRAC already provides user jobs with a list of replicas to try
  - First one (at the target site) used by default
  - Failover to other copies
- So we effectively have a kind of xrootd federation
  - Just via the DIRAC File Catalogue
- This mechanism could be used to run user jobs at sites with good networking but no storage
- However, we don't have a huge need to do that
  - Priority is increasingly going to be volume of Monte Carlo
- We also already stream data in to jobs on Tier-2 WNs (or VMs) for stripping, reprocessing, reconstruction
  - Data comes from a random T1/T2-D

GPUs

- During Run 2 / LS 2 we won't need GPUs for processing LHCb data or for simulation
  - Probably not for Run 3 either
- However, we do have research groups who are starting to use GPUs for high level fits in user grid jobs
  - e.g. the CHEP paper/poster about this: *"Search for matter-antimatter asymmetries in multi-body decays with GPUs"*
- So it may be worth making GPU resources generally available
  - e.g. via a "gpu" queue?
- Not obvious how GridPP should account for this though

Ideas for Upgrade / Run 3

- Computing upgrade TDR now being prepared
- Early versions of some ideas may start to appear during Run 2 / LS 2
  - Distributed computing will continue to be based on DIRAC
- Ideas relevant to GridPP include
  - Event indexes instead of creating new files via stripping
  - Analysis job trains to group file accesses at each site
  - Putting (extending) parameterised job handling into DIRAC
    - Some of the functionality of Ganga?
  - New Gaudi components to make multiprocessor jobs more efficient than current prototype

Summary

- The UK sites make the largest national contribution to LHCb
- LHCb workload increasingly dominated by Monte Carlo
- Run 2 HLT changes mean an end to reprocessing campaigns
  - Still stripping and reconstruction while running
- T2-D becoming Tier-2A plus SE
- LHCb encourages Machine/Job Features rollout, Information System simplification, access to GPUs from user jobs
- Remote data access already possible in user job and used in stripping etc at Tier-2s.
- Towards end of GridPP5, early versions of some Computing Upgrade changes may start to appear
- **LHCb aims to use resources in the formats the sites prefer, but doesn't want a proliferation of new interfaces**