LHCb in GridPP5

Andrew McNab
University of Manchester

Overview

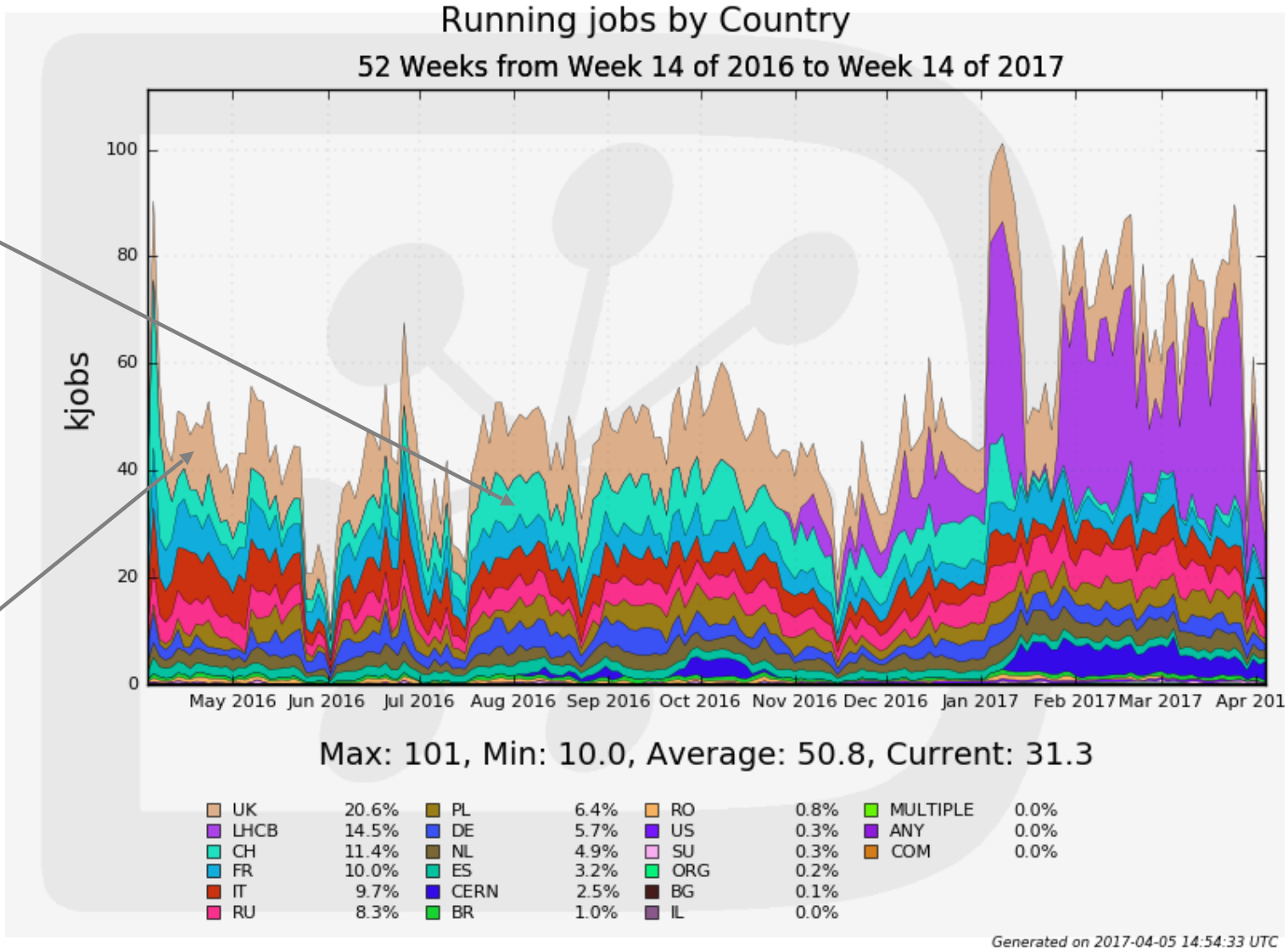
- Current state
- LHCb upgrade for Run3
- Key points
- Shift to more and more MC
- Other topics
  - Aims and protocols
  - Remote data access
  - Containers
  - CPU extensions and GPUs
  - WLCG developments

CPU power for LHCb jobs by country

.ch was mostly CERN Tier-0

.lhcb is the HLT farm, which is enabled for offline work whenever possible

.uk is the largest single country at 20.6%



Upgrade for Run3

- Major hardware upgrade of LHCb is during LS2 for Run3 not during LS3
  - End of GridPP5 in spring 2020 is ~year before Run3 is supposed to start
- Expect to be recording 5 to 10 times as much data
  - We know the offline system can record at this level, based on experience with 2016 ion runs
  - Already do reco in HLT with final calibration, so don't do reprocessing
  - But we need more storage and above all CPU to do the corresponding Monte Carlo, and user analysis / stripping / indexing
- Computing upgrade TDR now being prepared
- Distributed computing will continue to be based on DIRAC
  - Changes to DIRAC driven by scalability and requirements from applications
  - Major rewrite of applications underway to support multiprocessor operation

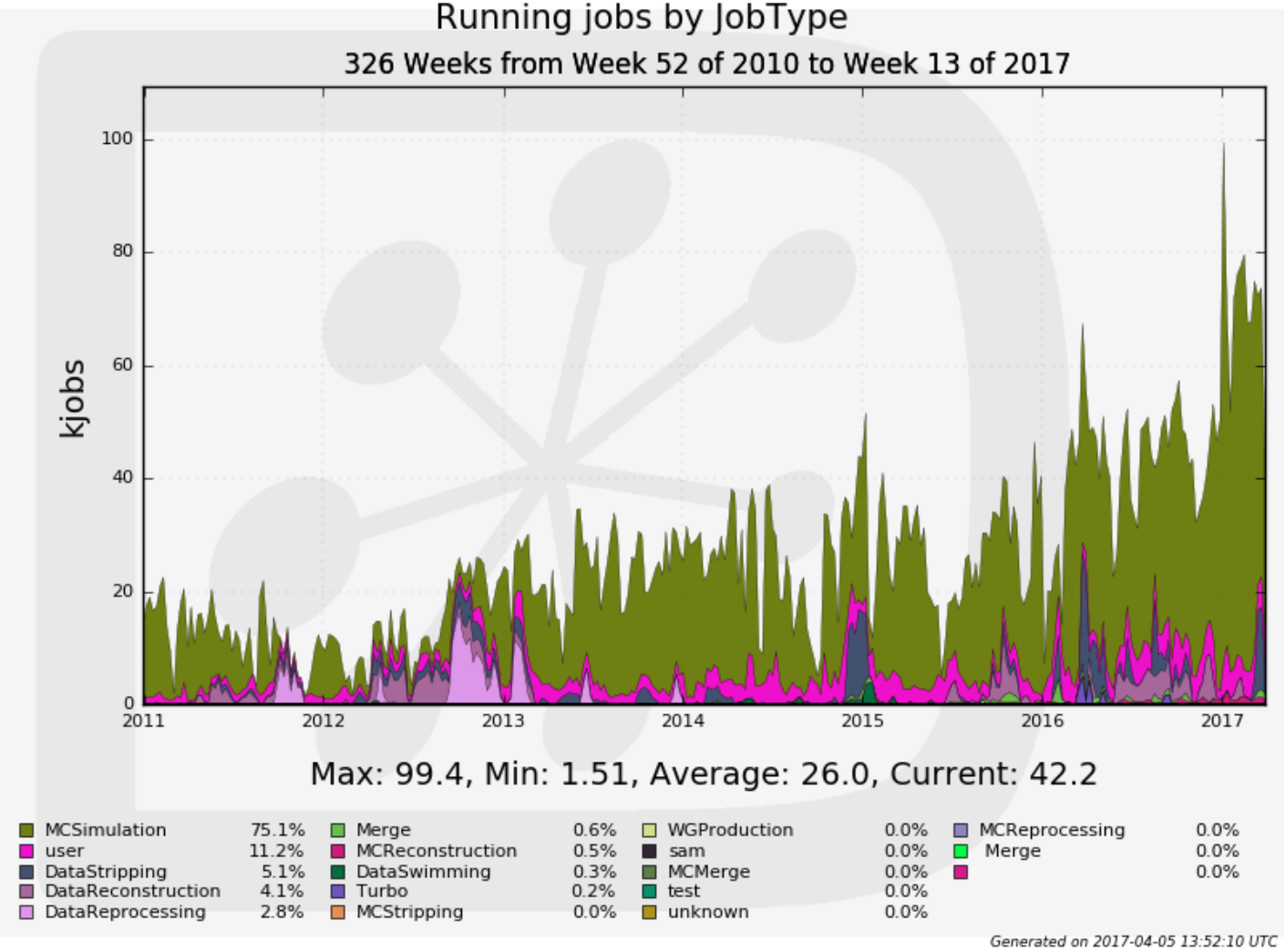
Key points

- LHCb mainly uses WLCG sites for running Monte Carlo
  - This is true for all sites, even CERN
  - This emphasis is going to continue during Run2 and into LS2 and Run3
- Tier-2 storage is only used by user jobs (~10% of all jobs)
  - Emphasis for user jobs is at Tier-1s
- Intermediate files are only stored at Tier-1s
  - eg some stripping jobs are run at Tier-2s, but they stream the data in and out from Tier-1s

LHCb in Tier-2 Evolution Document

"LHCb currently requires Tier-2 sites with storage to have a minimum of 300TB of disk, providing additional long-lived replicas of MC and data DST files for analysis jobs. The per-site minimum will rise in line with LHCb's overall request for Tier-2 storage. Nevertheless, ***all sites that support LHCb are used primarily for MC generation****, which runs without any use of local storage. LHCb strongly encourages concentrating effort spent supporting storage and user analysis jobs at* ***a few Tier-2 sites and the Tier-1****."*

JobTypes since 2011



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

Increasing user analysis.

Prompt reconstruction during data taking.

2011-13 had well defined reprocessing in planned campaigns, but these stopped for Run2

Other topics …

Aims and 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

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 pull data in to jobs on Tier-2 WNs (or VMs) for stripping, reprocessing, reconstruction
  - Data comes from a random T1

Containers as logical machines

- We are running at two sites which use LHCb containers
  - Andrew's system at RAL, and Skygrid at Yandex
  - Both use (different) containers derived from LHCb DIRAC VMs
- We've developed a generic LHCb container definition based on this experience
  - Uses Docker
  - Uses cernvm root image (ie via cvmfs)
  - LHCb cvmfs and /init script to run inside the container also provided via volumes
  - This is a format which will be supported by Vac and (inside a generic Docker VM) by Vcycle
    - See tomorrow's Vac/Vcycle talk

CPU extensions and GPUs

- Want to build applications with SSE4.2 asap
  - Needs correct platform matching by DIRAC
  - Will run old builds on the ~5% of hardware that doesn't support it
  - Expect to use this strategy for further desirable extensions in the future

- During Run 2 / LS 2 we won't need GPUs for processing LHCb data or for simulation
  - No plans for this in Run 3 either
- However, we do have research groups who using GPUs for high level fits in user grid jobs
- 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

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
  - Torque/PBS, HTCondor and Grid Engine implementations
- 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

Summary

- The UK sites make the largest national contribution to LHCb
- LHCb workload increasingly dominated by Monte Carlo
- Expect to continue to with model of user jobs at Tier-1 and a few large Tier-2 sites with LHCb storage
- LHCb encourages Machine/Job Features rollout, Information System simplification, ability to 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**