LHCb Status

A.Tsaregorodtsev, CPPM

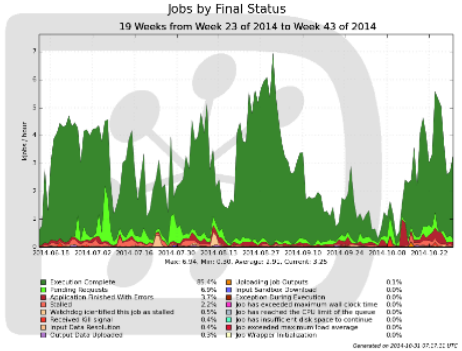
LCG-France, IRFU, 1 Nov 2014

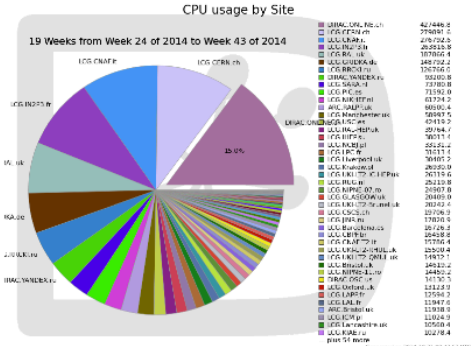
Outline

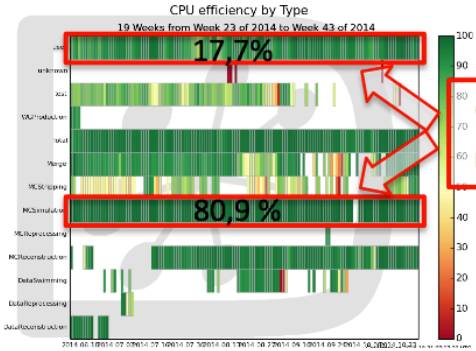
- Current operations
- Updated Computing Model for Run2
- T2D status
- Clouds, Vac, Vcycle
- Parallel calculations
- Federated storage
- Catalogs
- Data popularity and storage optimization

Grid usage in general

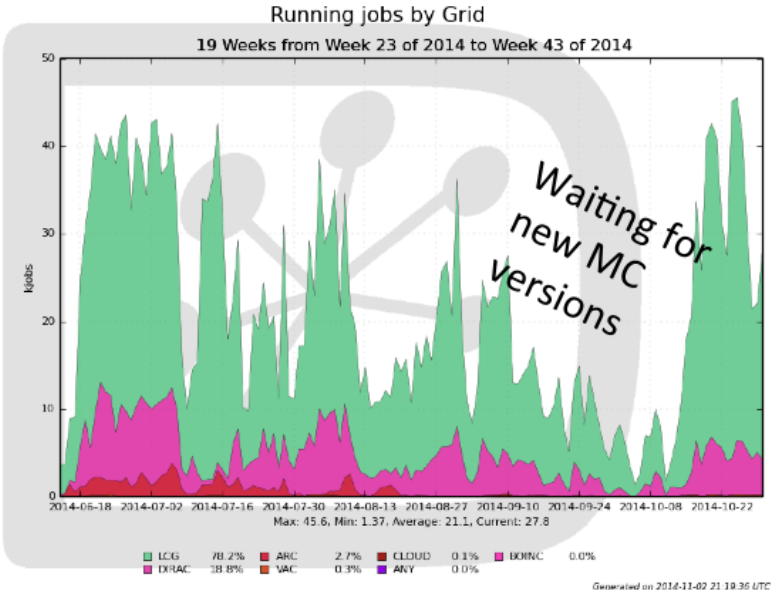
- Main metrics for this period stay unchanged
  - Dominated by Monte Carlo
  - > 91 % of all payloads successfully executed
  - All Tier1 sites in the top of used sites (+ HLT & Yandex)
  - Main activities with very high CPU efficiency ~ 100 %



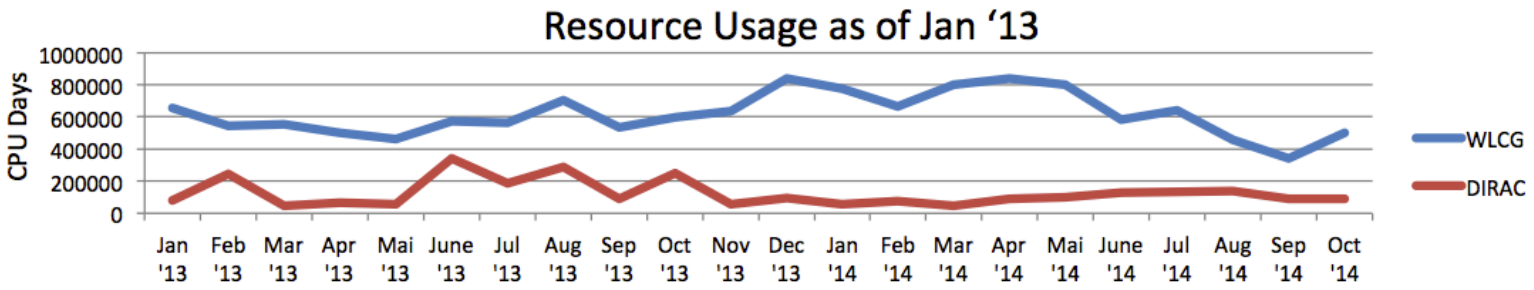




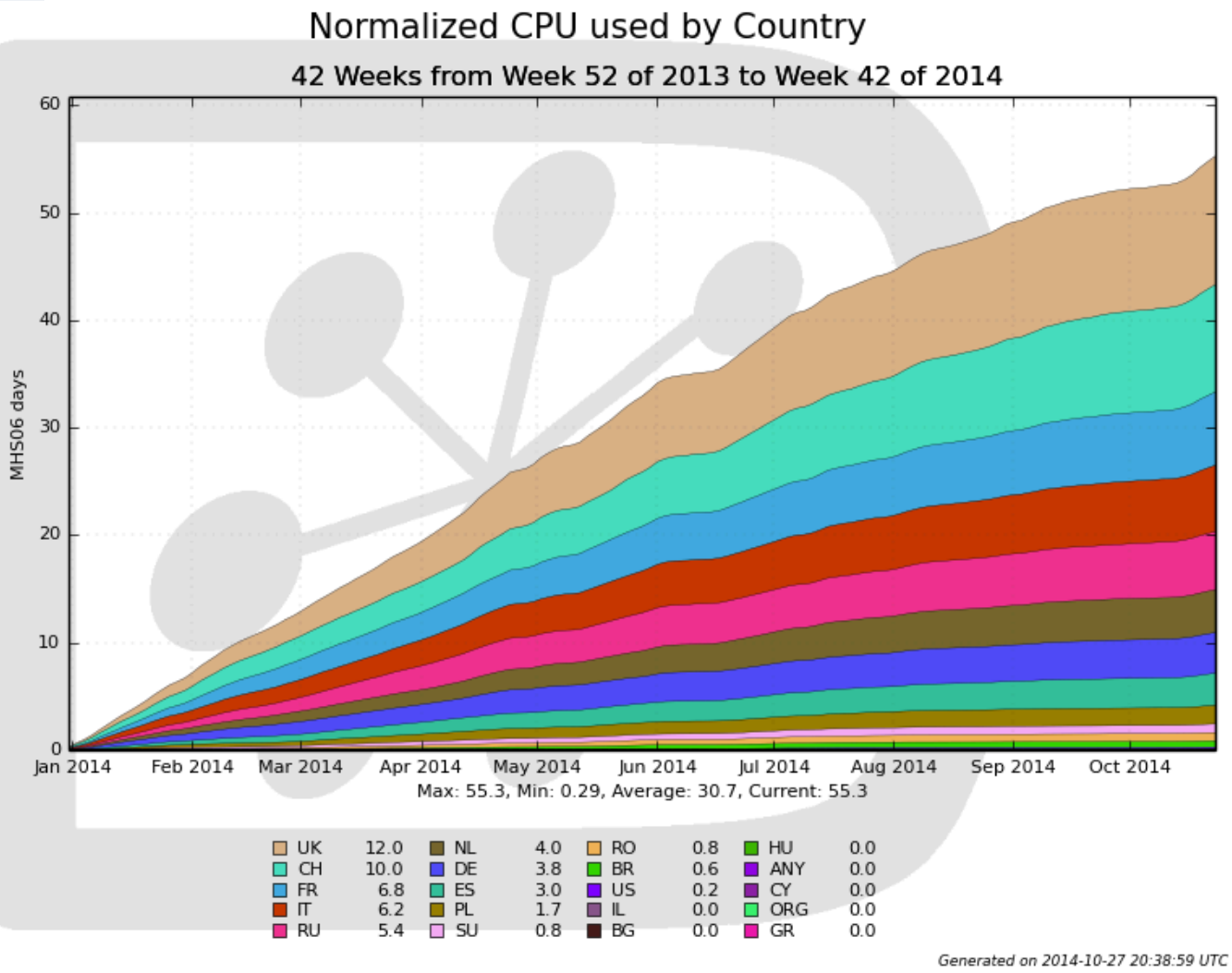
Resource Usage



- Low usage of resources in this period
  - Physicists waiting for new MC models to submit productions



Normalized CPU by Country



France: 12%
18.6 KHS'06
2340 CPU-years

Lyon T1 center

- **Pledge/consumed Oct 2014**
  - **CPU: 21700 HEPSPEC'06 years, consumed 12500**
    - Pledged ~18% of all the T1's
    - Consumed ~13% of all the T1's
  - **Disk: 1.35 PB (allocated 1.4), consumed 1.15**
  - **Tape: 1.8 PB, consumed 1.35**

- **CC/IN2P3 preliminary pledges and percentage of the LHCb request**
  - **CPU: 23 000 HEPSPEC'06 (~19%)**
  - **Disk: 1.88 PB (~15%) (complemented by T2D storage)**
  - **Tape: 4.36 PB (~18%)**
- **The pledges 2015 are very satisfactory**

Current operations

- Currently running Stripping 21 campaign to produce “legacy Run I dataset”
  - As the number of stripping code lines doubled, fighting with memory requirements (~ 3GB/job) – under control
- Using FTS3 for prestageing data in the BUFFER Storage Elements
  - Great simplification of operations and workflows
- Setting up Federated storage setup
  - HTTP/Webdav access

Assumptions for 2015

- RAW files will be produced in 2015 with a trigger rate of ~ twice 2012 (10 kHz RAW + 2.5 kHz turbo)
  - Assume 25 TB of RAW / day
- Prompt reconstruction will be delayed for using final calibrations
  - Assume 2 weeks delay
- The size ratio FULL.DST/RAW will remain the same as 2012
  - 45 TB of FULL.DST / day
- Stripping will hopefully be running almost in parallel with reco
  - Avoid having too large buffer used
- As soon as the RAW is in CERN-RAW, replicate to:
  - XXX-RAW (XXX being a Tier1, selected according to the RAW shares)
  - And in addition, to:
    - YYY-BUFFER: if it is decided to reconstruct the run at YYY
- Main difference with now:
  - Processing input SE selected at replication time (by run)

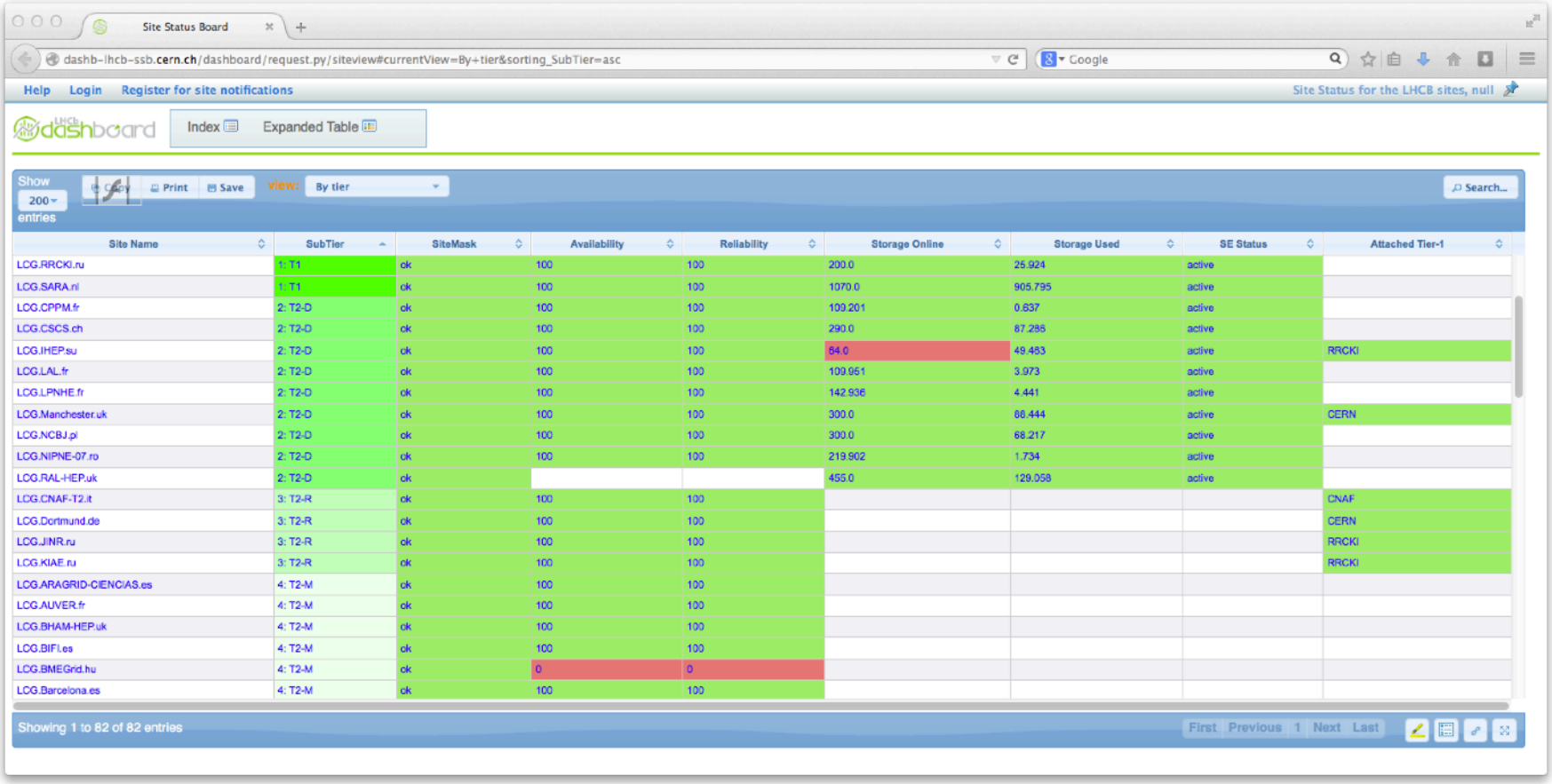
Assumptions for 2015 cont'd

- Reconstruction:
  - Run from Tier1-BUFFER
  - Upload FULL.DST to YYY-BUFFER
  - Replicate to the same YYY-RDST (or another)
- Stripping
  - Run from YYY-BUFFER replica of FULL.DST
  - Remove FULL.DST from BUFFER
    - Issue a removeReplica request in the job
  - Merging
    - As now, upload preferentially to selected YYY-DST but can be to any Tier1-DST (no need for FAILOVER)

Job matching

- Jobs can run anywhere, preferentially where data is, but not necessarily if download policy:
  - Introduce new concept in JDL
    - Primary sites (SE is accessible by protocol for remote file opening)
    - Secondary sites (SE is allowed for downloading to the WN)
  - Matching:
    - If no job available in matching by a pilot using the site as as Primary Site, try and match as Secondary site
    - If no specific explicit site matching available, match from ANY
  - User jobs can only use primary sites
- This is a challenge for efficient matching

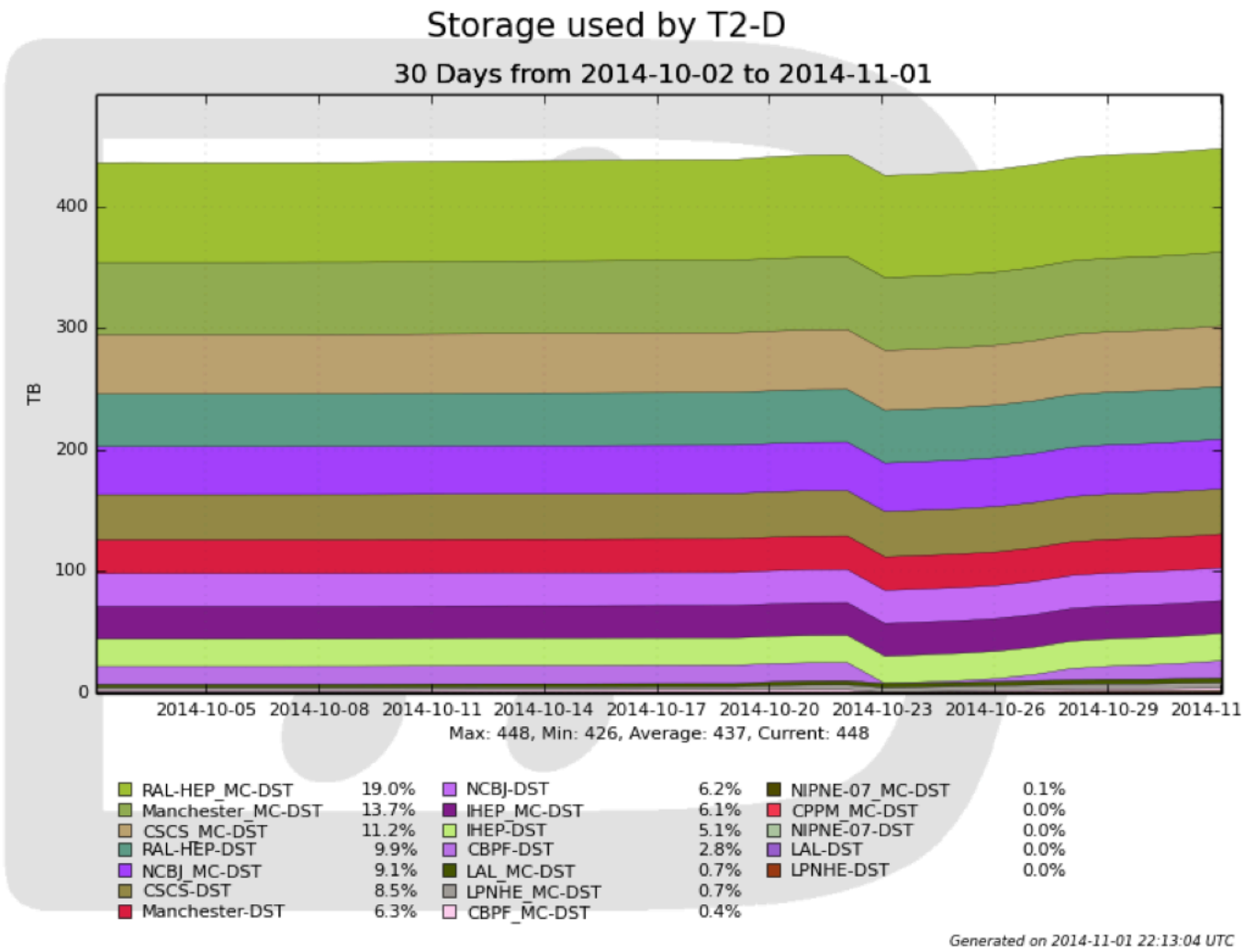
Side Status Board



T2D status

- 9 official T2D sites
  - dCache: CSCS.ch, IHEP.su, RAL-HEP.uk
  - DPM: CPPM.fr, LAL.fr, LPNHE.fr, Manchester.uk, NCBJ.pl, NIPNE-07.ro
- LAL.fr and LPNHE.fr share some infrastructure and offering 300TB together
- CPPM.fr offers ~100TB now
- T2D sites already provide 1991 TB disk storage for analysis jobs which meets the 2015 requirements
- DPM xrootd problem fixed and deployed at all DPM sites
- Mainly used for user analysis of DST data

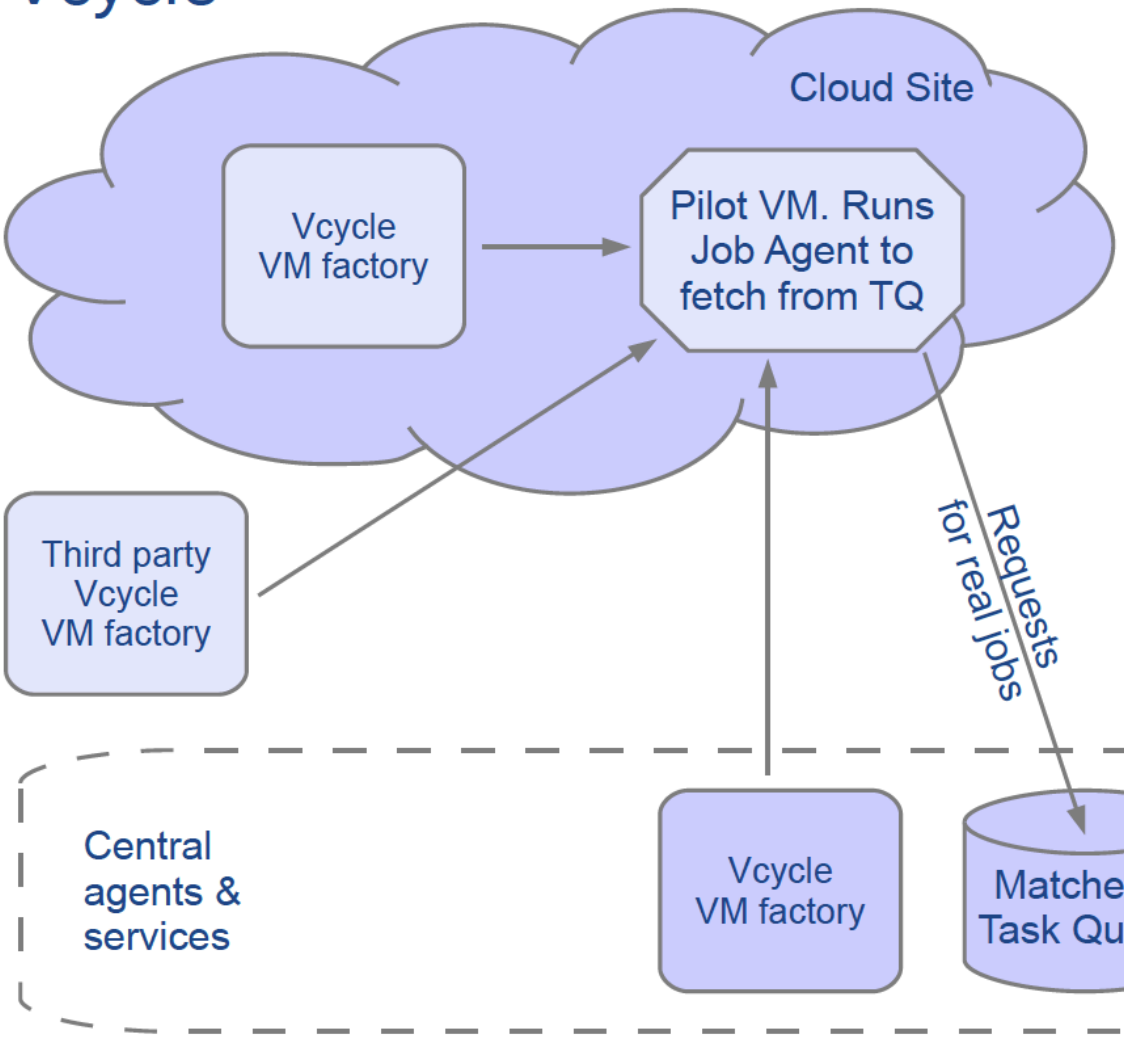
Storage used at T2-Ds



Cloud and virtualized resources

- **LHCb users Vac and Vcycle**
- Both VM Lifecycle Managers
- Vac is a standalone daemon you run on each worker node machine to create its VMs
- Vcycle manages VMs on IaaS Clouds like OpenStack
  - Can be run at the site, by the experiment, or by regional groups like GridPP
- Both developed at Manchester as part of GridPP Clouds/VMs effort
  - With help from Lancaster, Oxford, IC, CERN, LHCb and ATLAS
- Both make very similar assumptions about how the VMs behave
  - The same LHCb VMs working in production on Vac and Vcycle

Vcycle

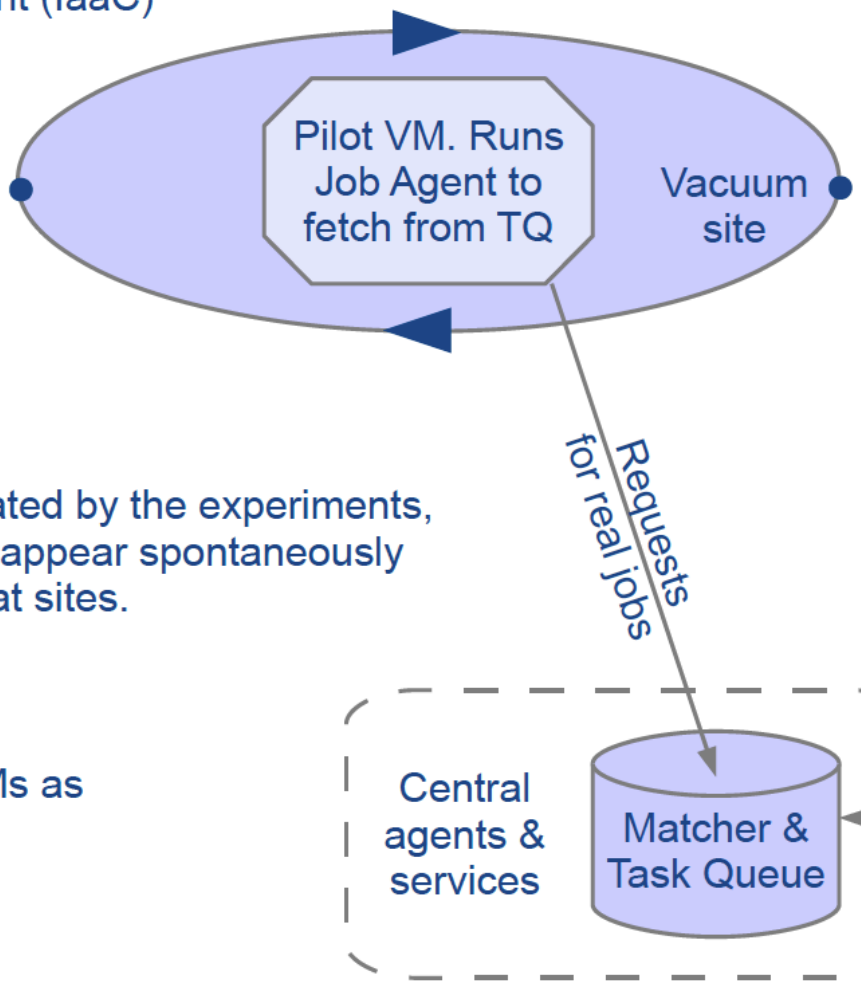


Vcycle is one implementation of an external VM factory that manages VMs

Can be run centrally by experiment or by site itself or by a third party

Vac's Vacuum Model

Infrastructure-as-a-Client (IaaC)



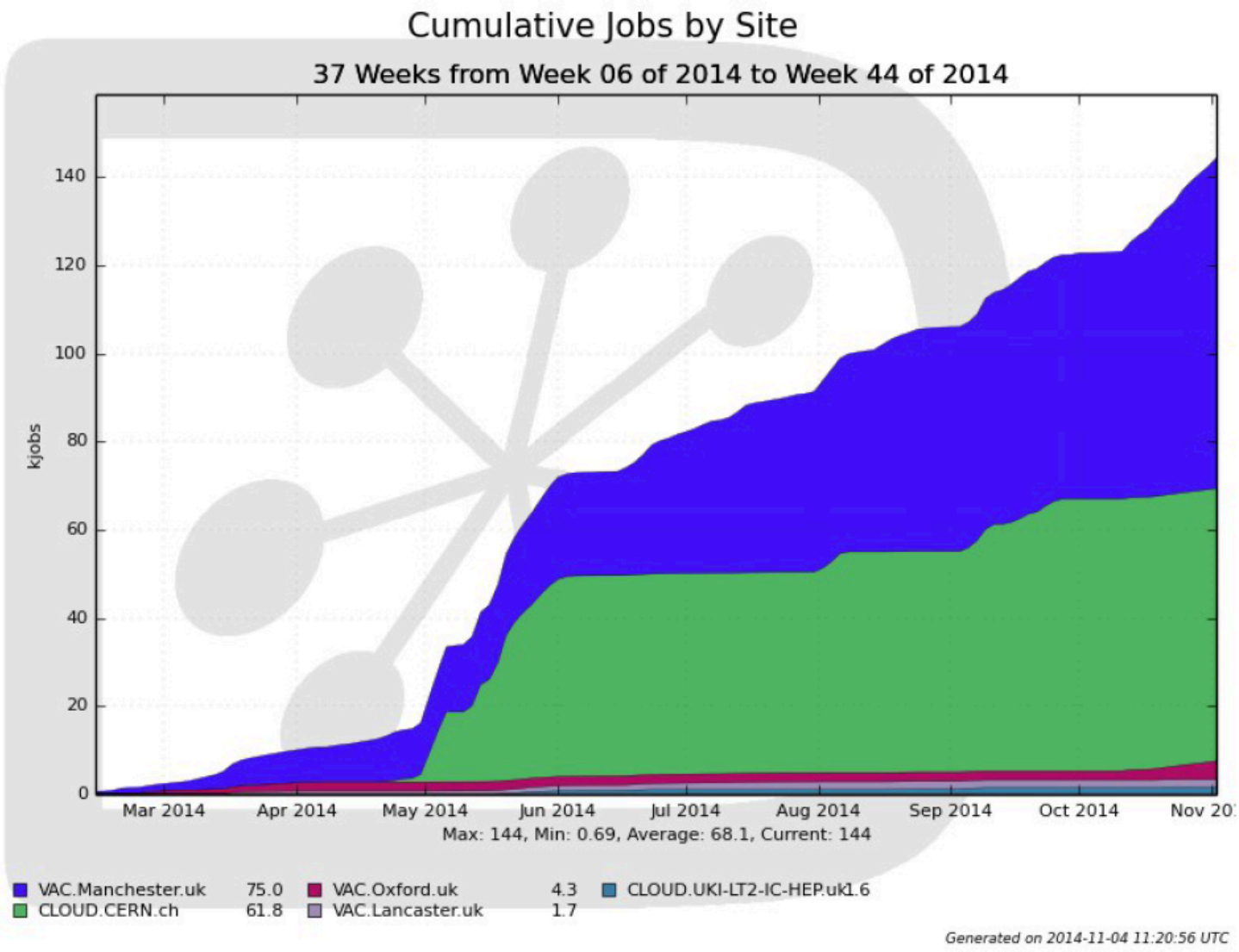
Since we have the pilot framework, we can do something much simpler.

Strip the system right down and have each physical host at the site create the VMs itself.

Instead of being created by the experiments, the virtual machines appear spontaneously "out of the vacuum" at sites.

Ideally use same VMs as with IaaS clouds

LHCb jobs from Vac and Vcycle VMs



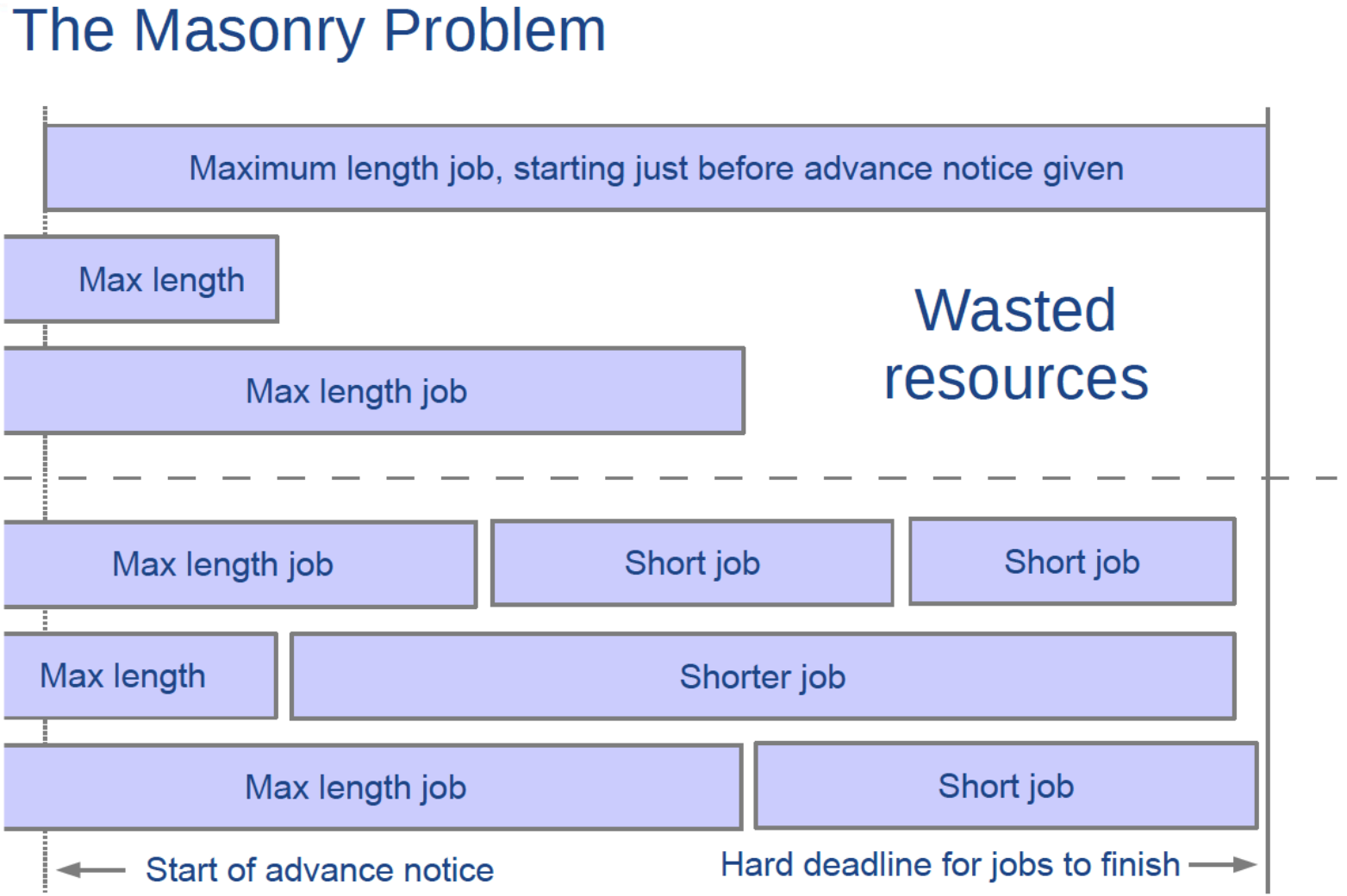
Vac and Vcycle

- Both Vac and Vcycle assume the VMs have a defined lifecycle
  - The VM runs and its state is monitored
  - VM executes shutdown -h when finished or if no more work available
- Can use machine/job features
  - Information on VM resources, e.g. number of cores
  - Information on remaining CPU, etc
- Machine/Job features are intended to be used in the DIRAC pilot running in the LHCb VMs
  - Knowledge of available resources
  - Graceful termination before the slot is exhausted

Vac and Vcycle for multi-core VMs

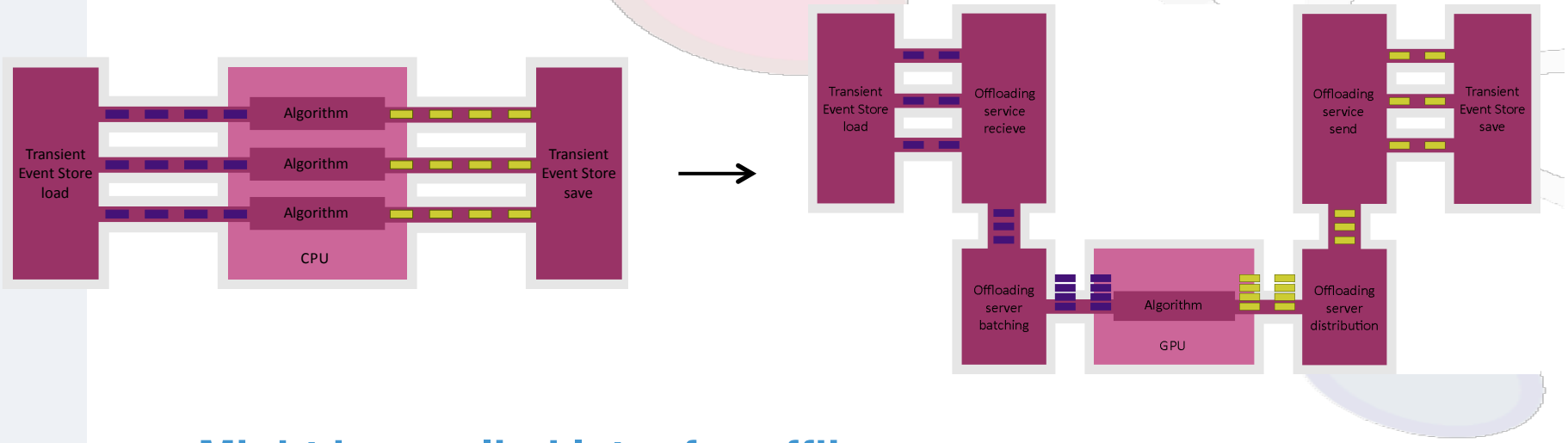
- Both Vac and Vcycle can create multiple processor VMs
- LHCb VM architecture is a simple wrapper around dirac pilot script, so running one multiprocessor payload in a VM should be straightforward
  - Using DIRAC Pilot 2.0 architecture
    - On the fly configurable pilots in terms of parameters and functionality
    - Executing a customizable, e.g. VO dependent, list of commands
- Managing multi-core slots is done by one or more JobAgents running inside the pilot
  - JobAgent is running a small local “batch system” for managing cores in its VM
  - Problems to solve
    - How to reconcile multi-core job requirements and priorities
    - Masonry problem

Multi-core node scheduling “Masonry” problem



Exploring parallel calculations

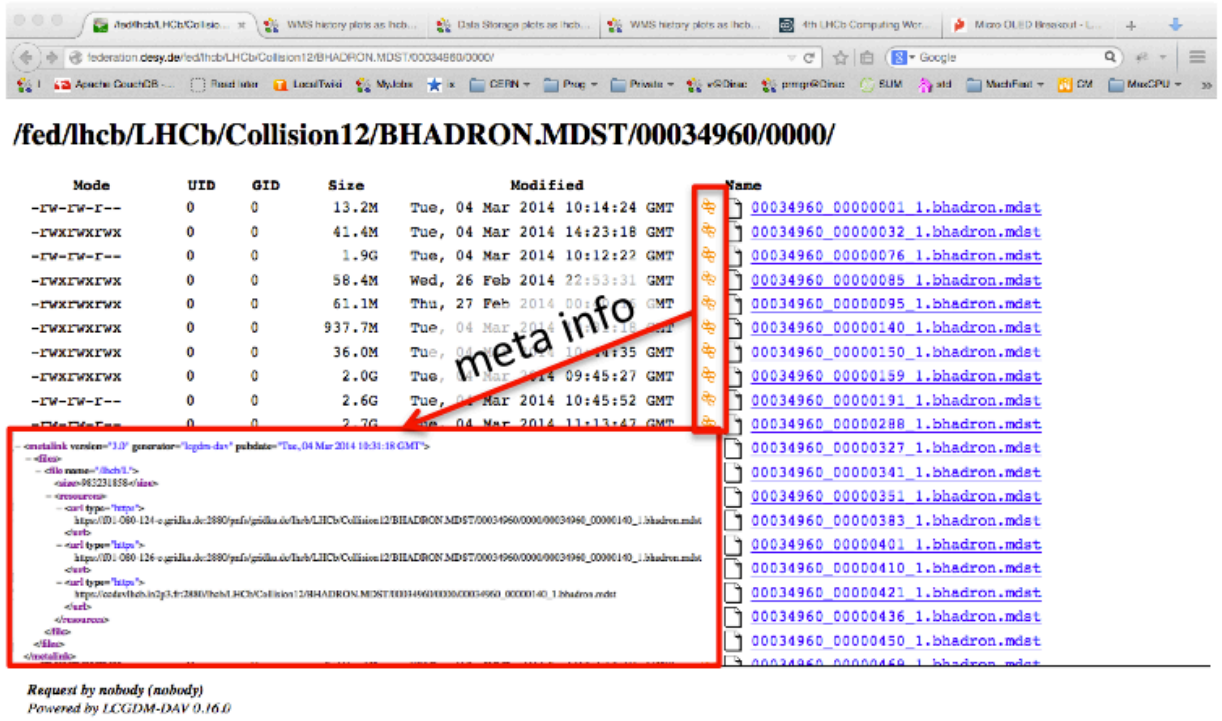
- LHCb Gaudi based applications capable of using multiple cores
  - How to get such jobs scheduled on the grid
  - Mixture of jobs with different number of cores requirements
- The use of GPUs is explored in the context of the HLT – software trigger
  - Gains and overheads



- Might be applied later for offline

HTTP protocol access & federation

- Currently setting up a prototype for http/webdav to all LHCb storages
  - HTTP/WEBDAV can become a main access protocol in the future
  - HTTP federation on top of these accesses
  - All but 5 sites available

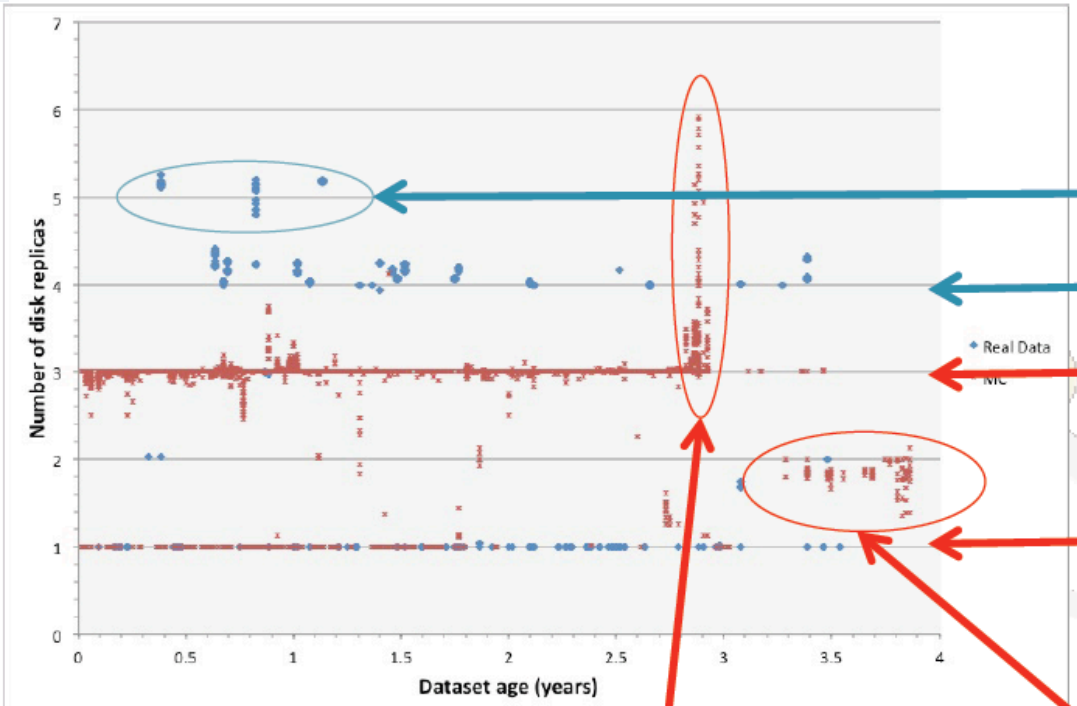


LFC to DFC migration

- Even if Federated storage is a success, the File Catalog is still needed
  - Access control rules
  - Metadata standard and user defined
  - Smart data placement policies
  - Efficient storage usage reports

- LHCb has chosen to replace the LFC by the DIRAC File Catalog
  - The last among the LHC experiments
  - The migration will be done in January 2015, before the RUN 2
    - 4 hours pause in operations to copy over the contents of LFC to DFC
    - Possibility to use LFC and DFC in parallel for a transition period
  - The tests, functionality and performance, are ongoing
    - Results are optimistic
  - DAV access would be needed to couple nicely with Federated storage

Number of replicas vs Computing Model



Mistakes!

Real Data

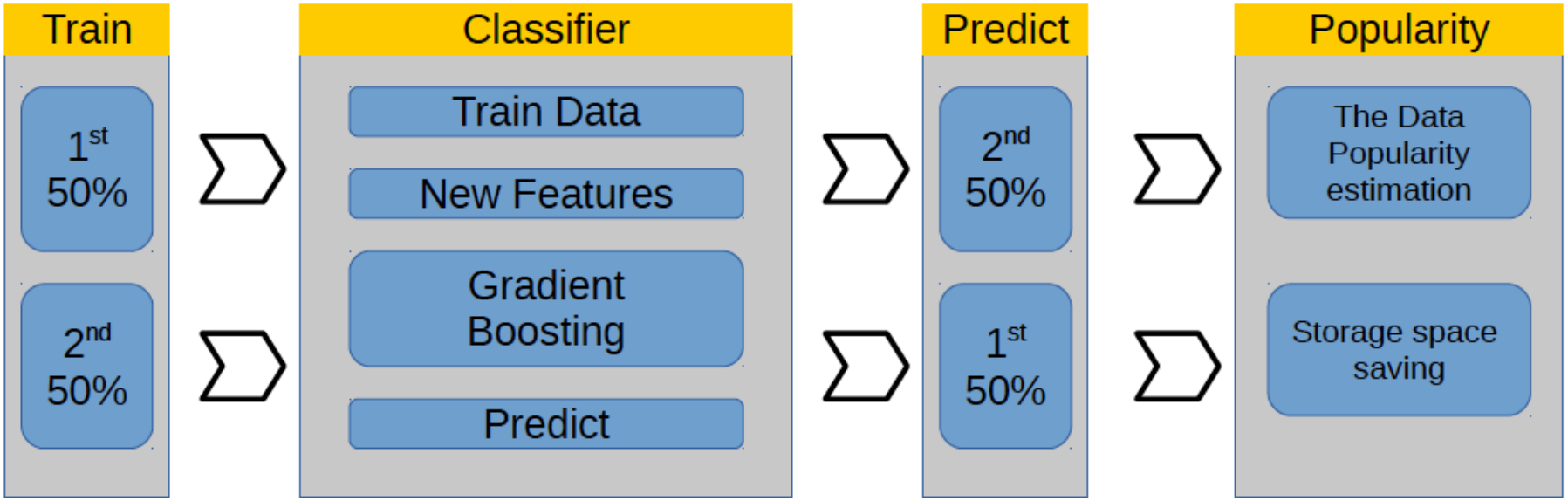
MC

Tests / WG prods

Manually reduced replicas

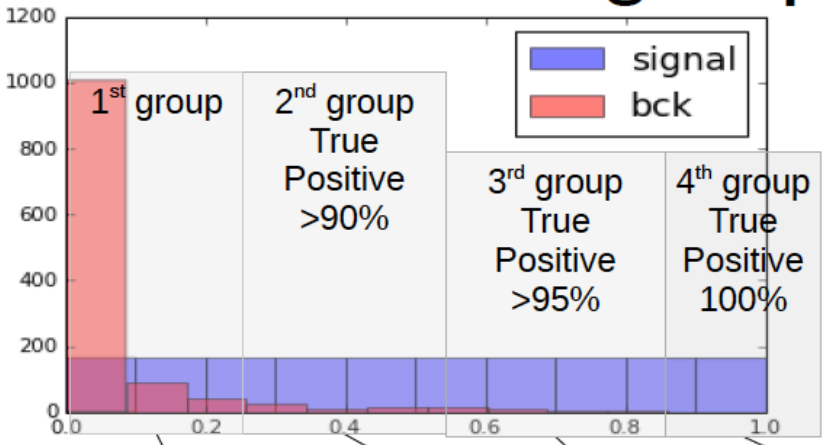
Wrong replication policy + no babysitting

Data Popularity for storage usage optimization



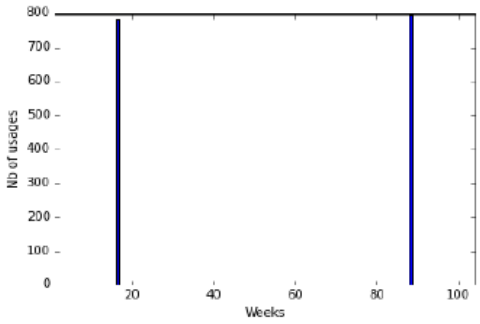
Data Popularity for storage usage optimization

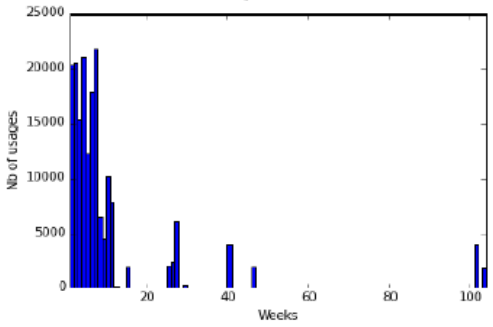
Storage space saving

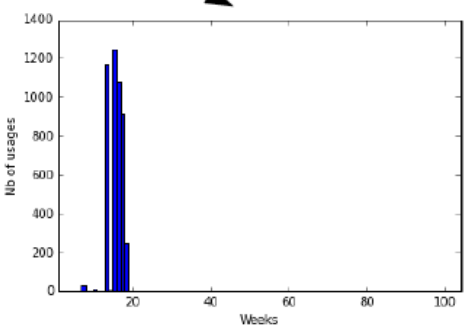


1st group: select the datasets with low usage rate (after selection).
2nd group: leave some replicas on disk.
3rd group: leave small number of replicas.
4th group: leave 1 replica or less on disk.

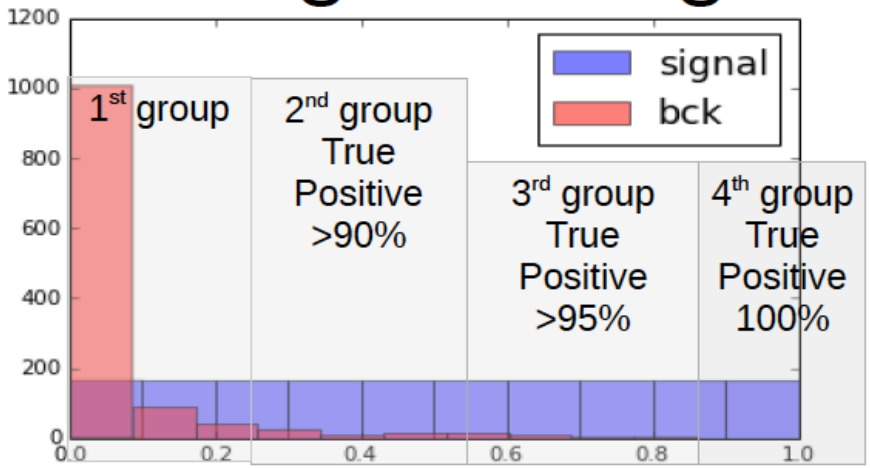
Selection:
nb_peaks < 4;
inter_max > 26.







Storage saving strategies examples



	3-4 groups	2Nd group	1St group	Total
Dataset that was not used from 1st to 78th weeks. Remove from disks. ~1077 TB	Transfer to the tapes ~1008 TB	Leave 1 replica ~670 TB	Leave 1 replica	~2782 TB of 8894 TB
	Leave 1 replica ~1277 TB		~27 TB	~2381 TB of 8894 TB

Conclusions

- Very good efficiency of operations in 2014
- 2015 pledges of French T1 resources are very satisfactory
- T2D sites are getting into production
  - **Will be fully used with RUN2 start**
- Using cloud virtualized resources with Vac and Vcycle in production on several sites
  - **More to come**
- Using multi-core nodes is possible
  - **More intelligent scheduling to increase efficiency is needed**
- Using Federated storage is demonstrated integrating 13 out of 19 LHCb sites
- Replacing LFC by DFC is being prepared
- Analysis of data popularity can help to reduce the usage of disk storage