LHC Run 2 Computing

Alastair Dewhurst, Andrew Lahiff, Raja Nandakumar

Introduction

- Trying to provide one integrated talk rather than 3 consecutive talks on ATLAS, CMS, LHCb.
- Computing Models
- Start of Run 2
- Distributed Data Management / Placement
- Federated XrootD
- Jobs / CPU usage
- New Things

Computing Models

LHCb Computing model

- Data storage
  - Tape at CERN + 7 Tier-1 sites
    - ~30% of Tier-1 storage at RAL
  - Disk at CERN + 7 Tier-1 sites + 11 Tier-2D sites
    - Each T2D with > 300TB storage and local LHCb contact
    - DPM and dCache Storage
- CPU
  - Jobs run at all sites which allow LHCB
  - Jobs without need to access data run everywhere
    - Sometimes throttled at T1 when reprocessing is being done.
  - Jobs needing to access data send to where data is available
    - Remote access when local copy is not available

LHCb Computing model

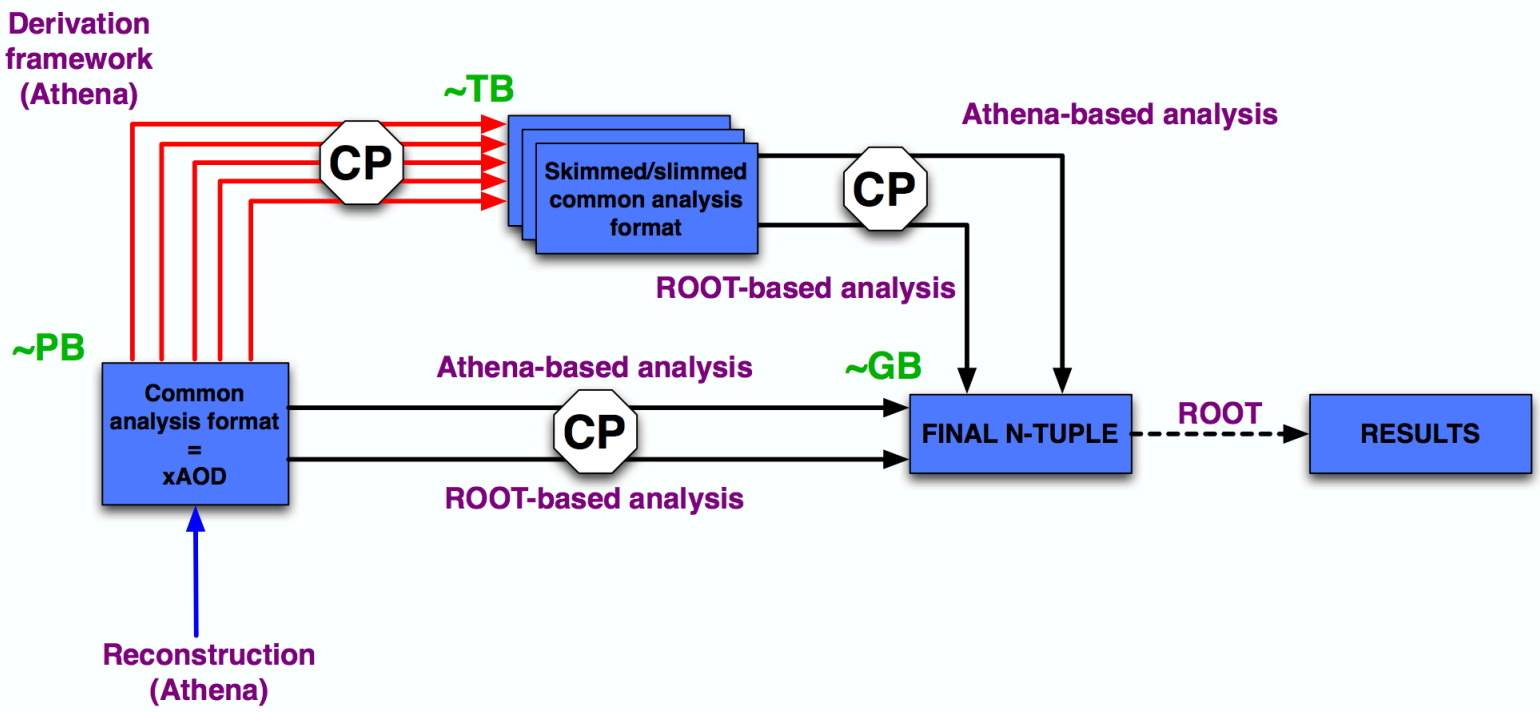
- LHCbDIRAC to handle jobs on the grid
  - DIRAC + LHCb specific customisations
  - LHCb initiated and till now, the main developer of DIRAC
- All production and large-scale data operations through DIRAC
  - Good interfaces to most grid services
- Ganga for user interface to grid
  - Instrumented with knowledge of LHCb applications
  - Submits jobs to DIRAC which then runs them and returns the output to the user through Ganga

ATLAS Computing model

- Slow evolution away from strict model from Run 1
  - LFC long gone.
  - Transfers allowed directly between T2D.
  - Cloud boundaries slowly being removed.
- New analysis model for users (next slide)
- More aggressive data cleanup and archival to tape
- Trying to make use of as many opportunistic resources as possible.

ATLAS analysis model

- Run 1 model was not efficient for users
- AOD and D3PD replaced by a root readable xAOD format.
- Analysis will be performed on small derived formats that will be produced centrally by the production system.



CMS Computing model

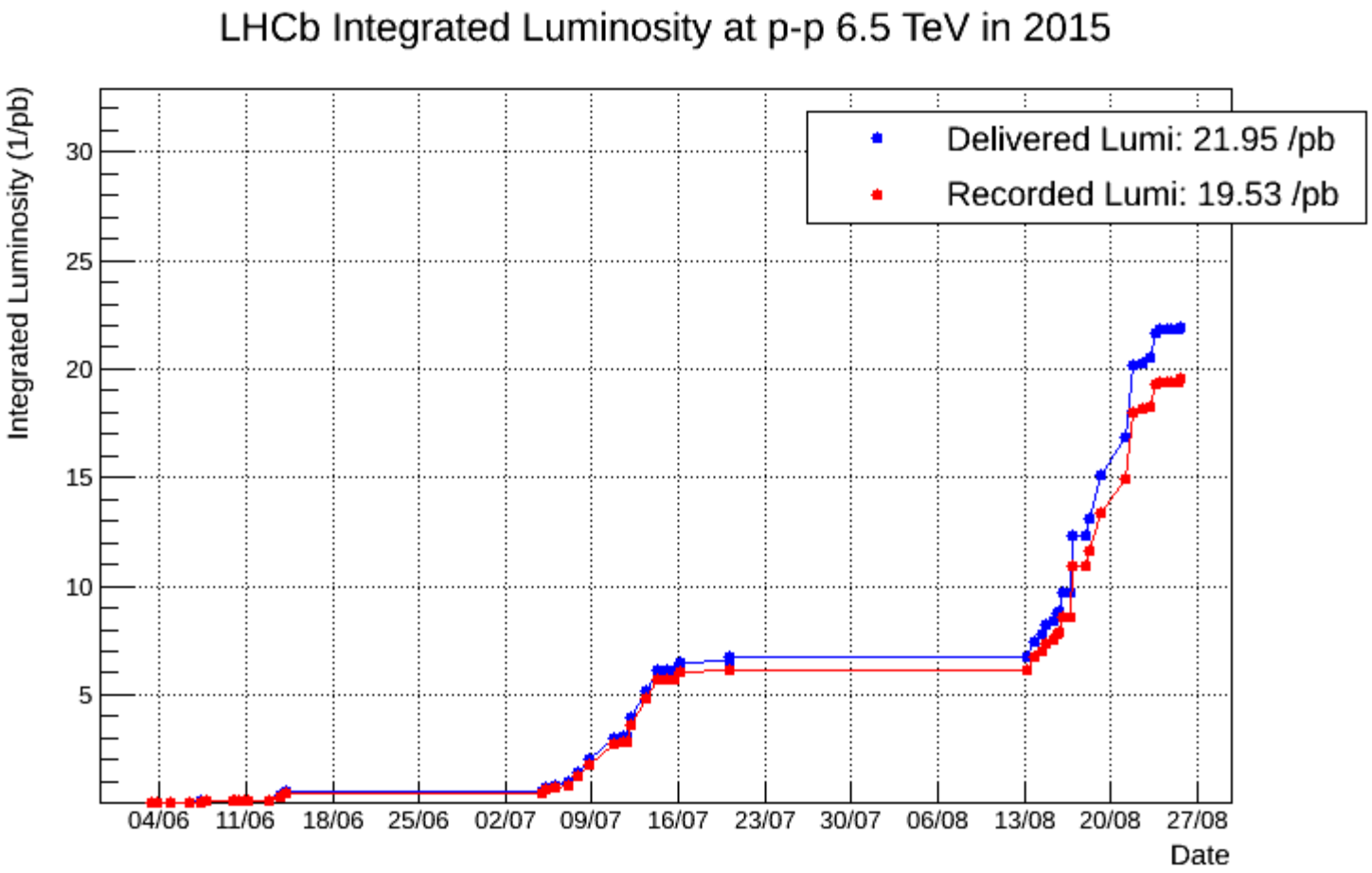
- Data storage
  - tape at CERN + 7 Tier-1 sites
    - 8% of Tier-1 storage at RAL
  - disk at CERN + 7 Tier-1 sites + Tier-2 sites
- Moved away from the regional model used during Run 1
  - Run 1: MC production run in regional groups of Tier-2s and with output data stored at associated Tier-1
  - Run 2: MC production run across all Tier-1 and Tier-2 sites with output data stored at any Tier-1
- Introduction of MiniAOD data tier in 2014
  - serve the needs of mainstream physics analyses
  - small event size (30-50 kb/event)

CMS Workload management

- Global pool
  - single HTCondor pool for all CMS analysis & production
  - separate Tier-0 pool, but jobs can flock to the global pool
- WMAgents (production) & CRAB3 servers (analysis) submit jobs to the global pool
- glideinWMS submits pilots to sites
  - run HTCondor startds which join the global pool

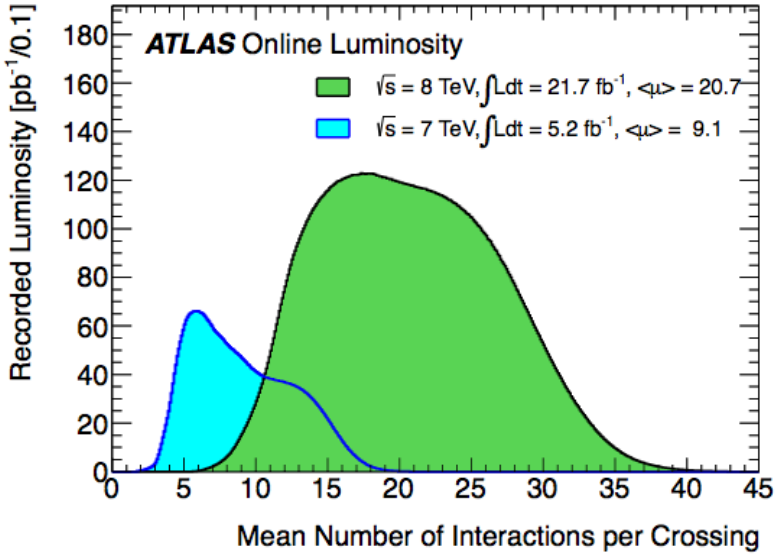
Start of Run 2

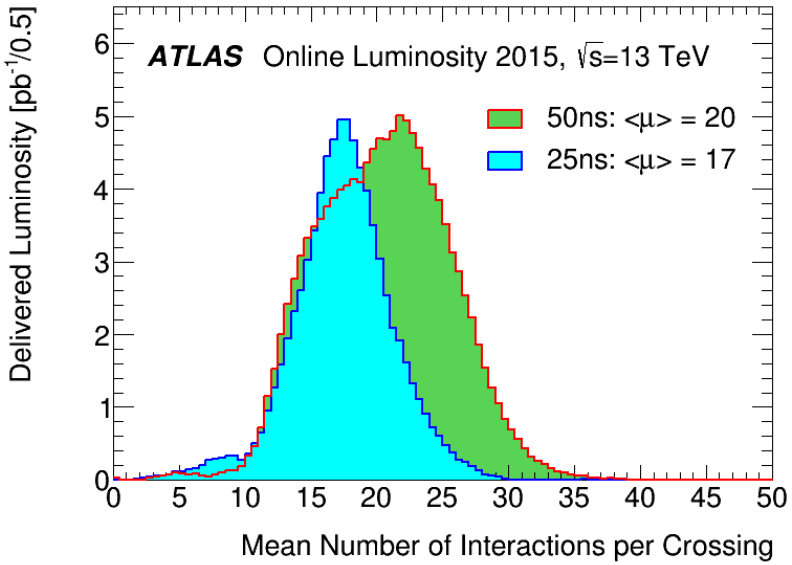
Data taking


LHCb Integrated Luminosity at p-p 6.5 TeV in 2015
Integrated Luminosity (1/pb)
Delivered Lumi: 21.95 /pb
Recorded Lumi: 19.53 /pb
30
25
20
15
10
5
04/06
11/06
18/06
25/06
02/07
09/07
16/07
23/07
30/07
06/08
13/08
20/08
27/08
Date

Data taking

- More interactions per bunch crossing mean larger memory + CPU footprint for reconstruction.
  - Number of interactions will increase significantly
  - LHCb de-focus beam for fewer interactions





Distributed Data Management / Placement

ATLAS DDM

- Rucio is the Distributed Data Management system.
- Started replacing DQ2 in production from December 2014.
- Was probably introduced before it was ready. Missing features have now been added:
  - DaTRi (Data transfer interface for users) has now been replaced with R2D2
  - Recovery service for lost files now working
- Automated file consistency service being worked on.

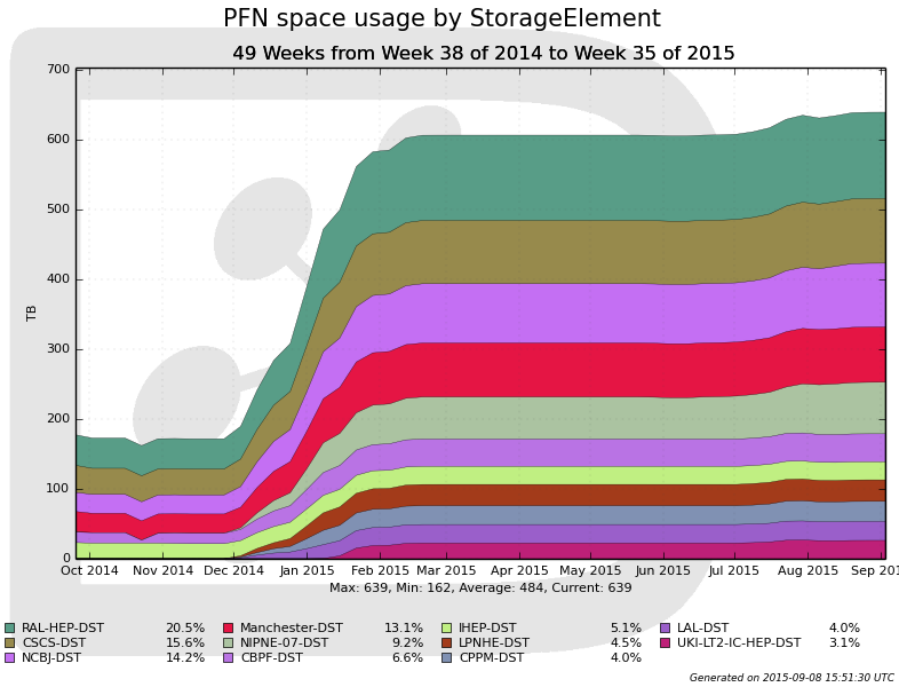
ATLAS Data Placement

- Russian T1 is starting to accept tape backed data, Taiwan T1 tape endpoint is being decommissioned.
- New data placement model evolving for run 2:
  - T2s are selected from the following criteria: more than 400TB capacity in DATADISK and more than 90% availability for analysis in the last 3 months.
  - (x)AODs 1 copy across Tier 1, 1 copy across T2
- More replicas can be produced based on popularity.
  - Files cleared up if not used for a certain period of time.
- The sum of each group derivations are allowed ~5% of AOD total.

LHCb Data Placement

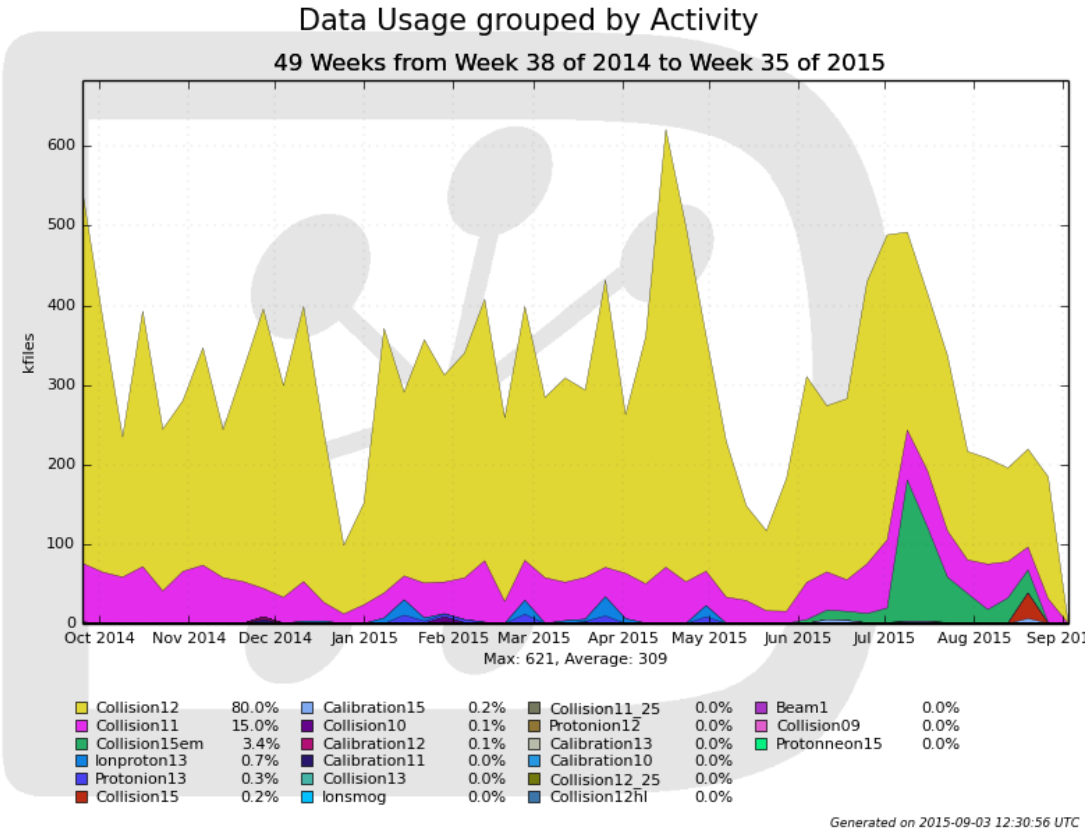
- Latest (nth) processing :
  - Data : 4
  - Simulation : 3
- n-1th processing :
  - Data : 2
  - Simulation : 2

Earlier processing : Not on disk

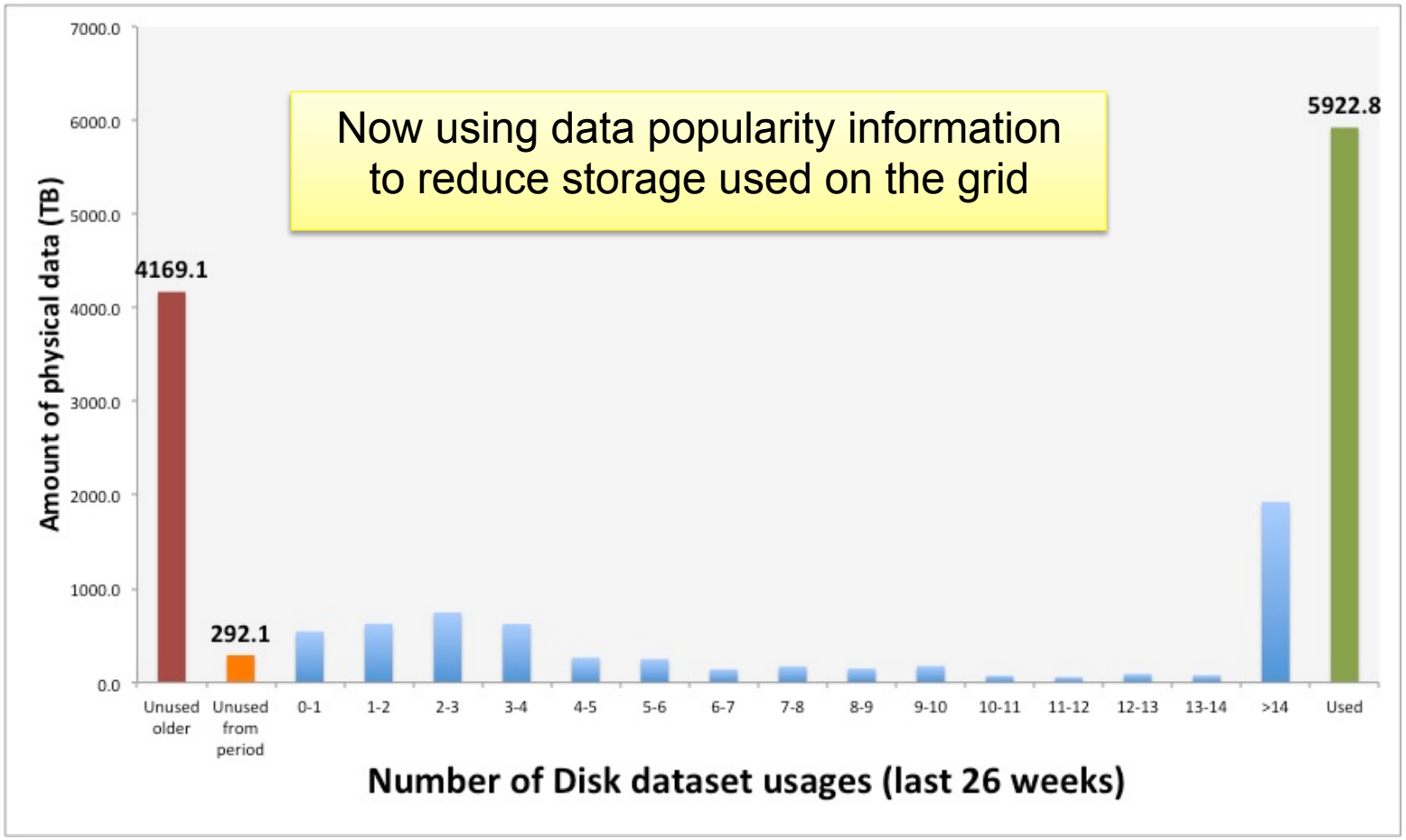


LHCb Data usage

- LHCb data usage by type.
  - Green peak = 2015 data taking



LHCb Data popularity

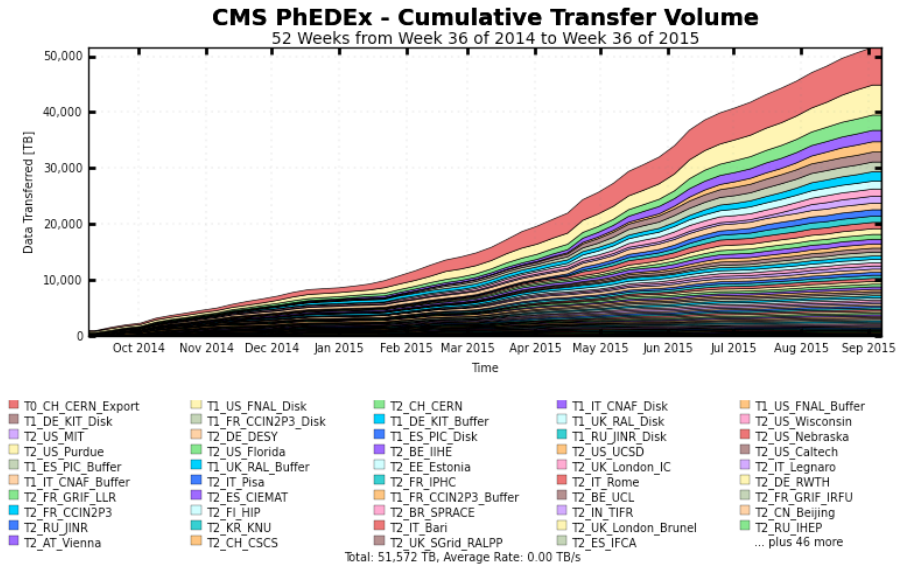


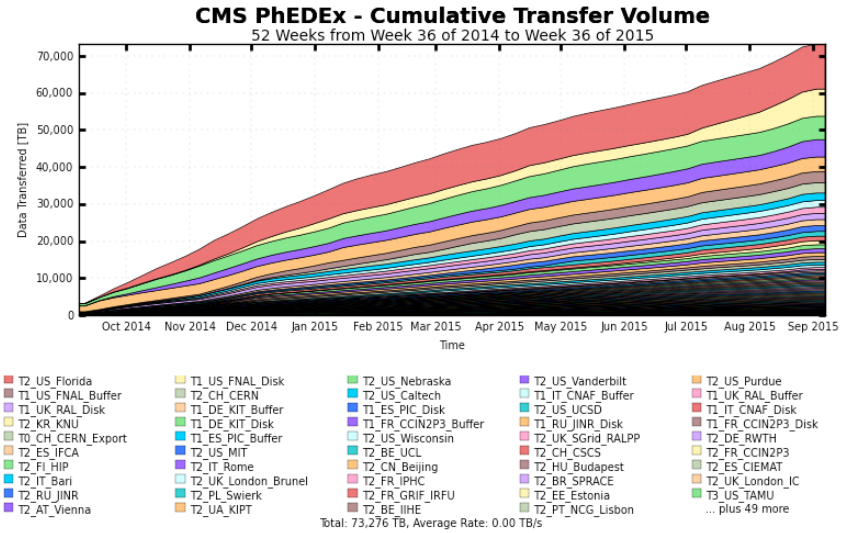
CMS Dynamic data management

- All Tier-1/Tier-2 storage is treated as one big distributed cache – dynamic data pool
  - has been in operation since January 2015
- Data popularity
  - usage of datasets is logged
- Data replication
  - based on dataset popularity additional copies are created
- Cache release
  - prevents sites from filling beyond 90%
  - least valuable dataset replicas are deleted (keeping at least one copy in the system)

CMS Production vs debug

- CMS data transfers over the past year
  - ~50 PB production
  - ~70 PB debug





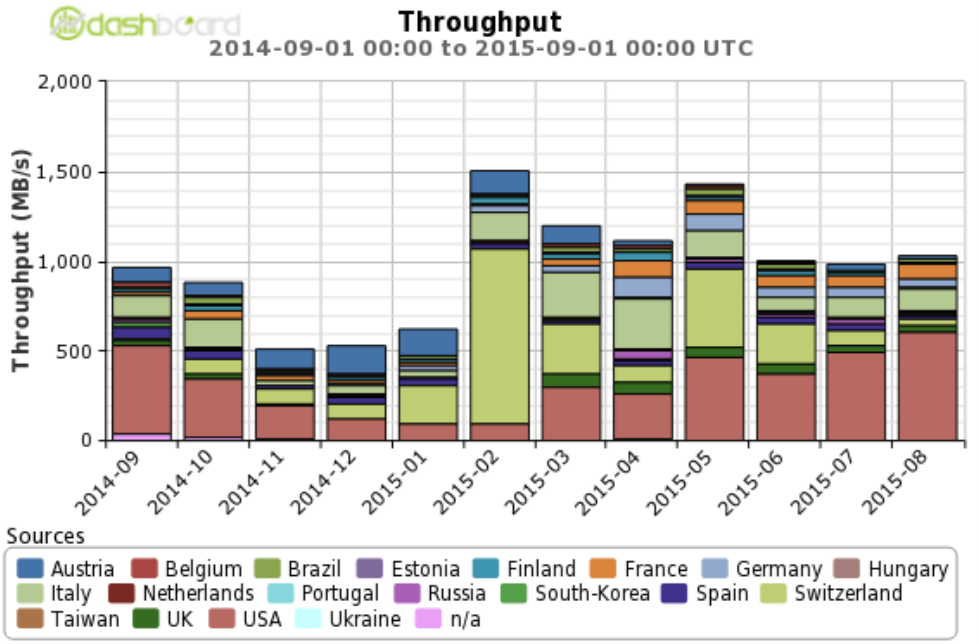
Federated XrootD

Federated XrootD

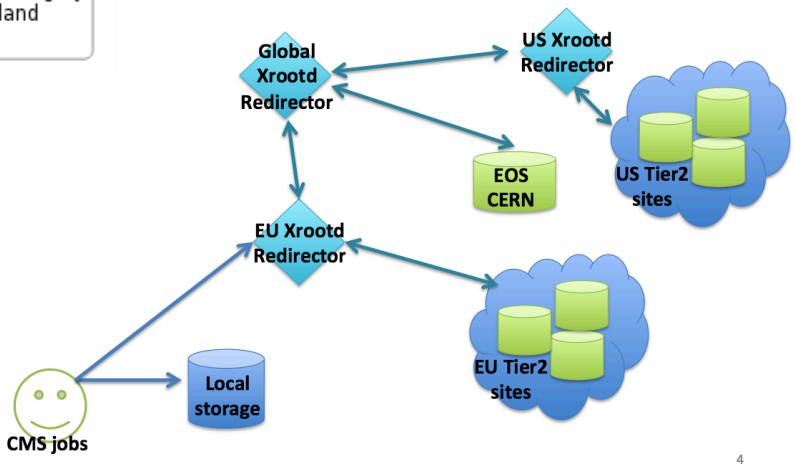
- ATLAS, CMS and LHCb all use Federated XrootD
  - FAX for ATLAS
  - AAA for CMS
  - 'Phil' for LHCb (They haven't given it a name yet)
    - LHCb have no monitoring yet but have seen failover working in job logs

	ATLAS	CMS	LHCb
Fallback	Yes	Yes	Yes
Overflow Jobs	< 10%	Yes	No
Opportunistic Resources	Yes	Yes	In Future

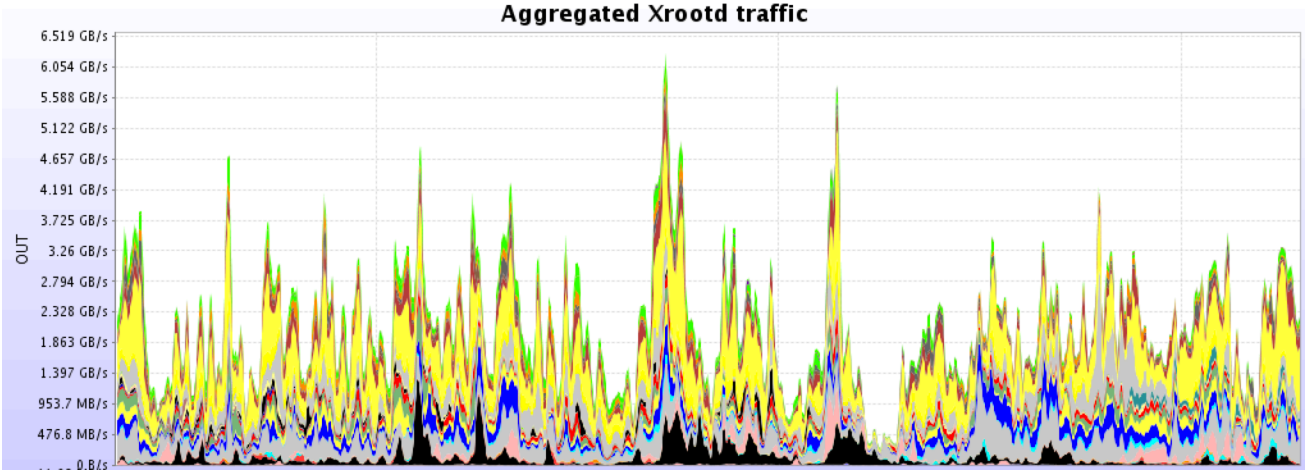
AAA


dashboard
Throughput
2014-09-01 00:00 to 2015-09-01 00:00 UTC
Throughput (MB/s)
2,000
1,500
1,000
500
0
2014-09
2014-10
2014-11
2014-12
2015-01
2015-02
2015-03
2015-04
2015-05
2015-06
2015-07
2015-08
Sources
Austria Belgium Brazil Estonia Finland France Germany Hungary
Italy Netherlands Portugal Russia South-Korea Spain Switzerland
Taiwan UK USA Ukraine n/a

ar


Global Xrootd Redirector
US Xrootd Redirector
EOS CERN
US Tier2 sites
EU Xrootd Redirector
EU Tier2 sites
Local storage
CMS jobs

FAX usage



Show 25 entries Search:

Site	Jobs	WithFAX [files]	WithoutFAX [files]	WithFAX [GB]	WithoutFAX [GB]
unknown: AGLT2_SL6	1	1	0	0.19	0.00
unknown: BU_ATLAS_Tier2_SL6	1	1	0	0.19	0.00
unknown: CA-SCINET-T2_MCORE	4	8	12	0.05	0.08
unknown: DESY-ZN_MCORE	1	1	0	0.17	0.00
unknown: FMPhI-UNIBA_MCORE	2	2	0	0.44	0.00
unknown: IFAE_MCORE	2	2	0	0.42	0.00
unknown: IFIC	39	39	0	17.02	0.00
unknown: IHEP_PROD	13	13	0	0.18	0.00
unknown: INFN-FRASCATI	1	1	0		
unknown: INFN-MILANO-ATLASC_MCORE	3	30	0		
unknown: INFN-ROMA3_MCORE	14	14	0		
unknown: LPC	16	16	0		
unknown: LRZ-LMU	1	1	0		
unknown: LRZ-LMU_MCORE	1	19	0		
unknown: MWT2_SL6	29	29	0		
unknown: OU_OCHEP_SWT2	1	1	0	0.19	0.00
unknown: praguelcg2	3	3	0	0.59	0.00
unknown: ROMANIA07	13	13	0	2.54	0.00
unknown: SLACXRD	1	1	0	0.19	0.00
unknown: Taiwan-LCG2	29	45	20	8.12	0.01
unknown: Taiwan-LCG2_VL	1	1	0	0.11	0.00
unknown: TECHNION-HEP_MCORE	2	2	0	0.09	0.00
unknown: UKI-LT2-Brunel_MCORE	7	14	0	0.28	0.00
unknown: UKI-SCOTGRID-GLASGOW_SL6	1	1	0	0.28	0.00
unknown: UKI-SOUTHGRID-BHAM-HEP_SL6	82	82	0	1.56	0.00

Showing 1 to 25 of 25 entries

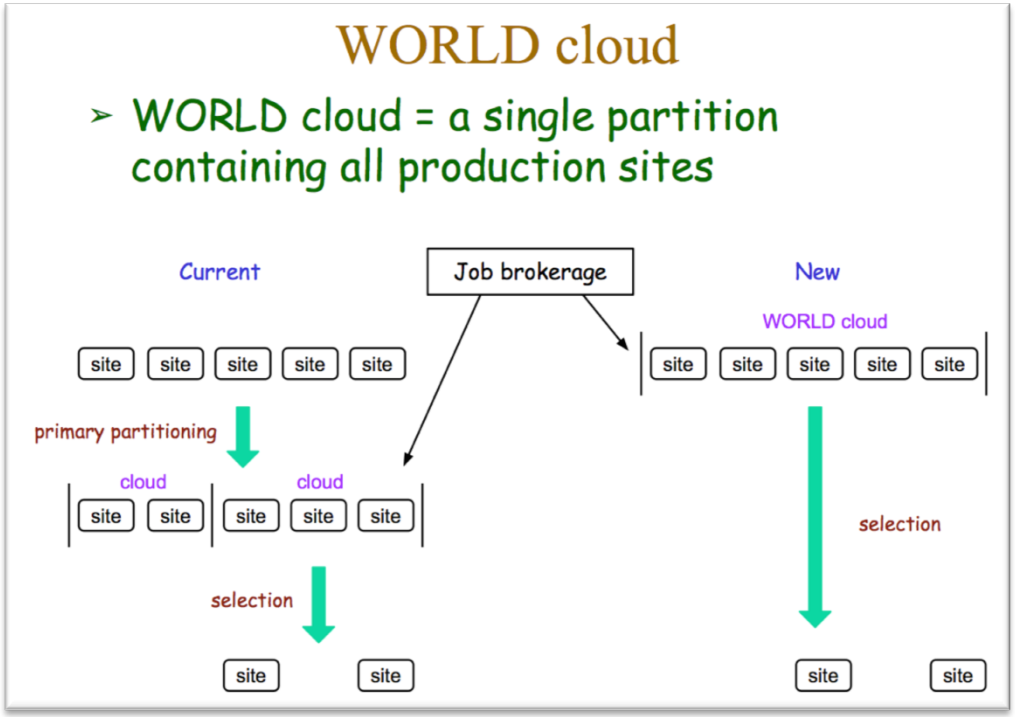
Jobs using FAX as failover in last 6 hours

10th September 2015

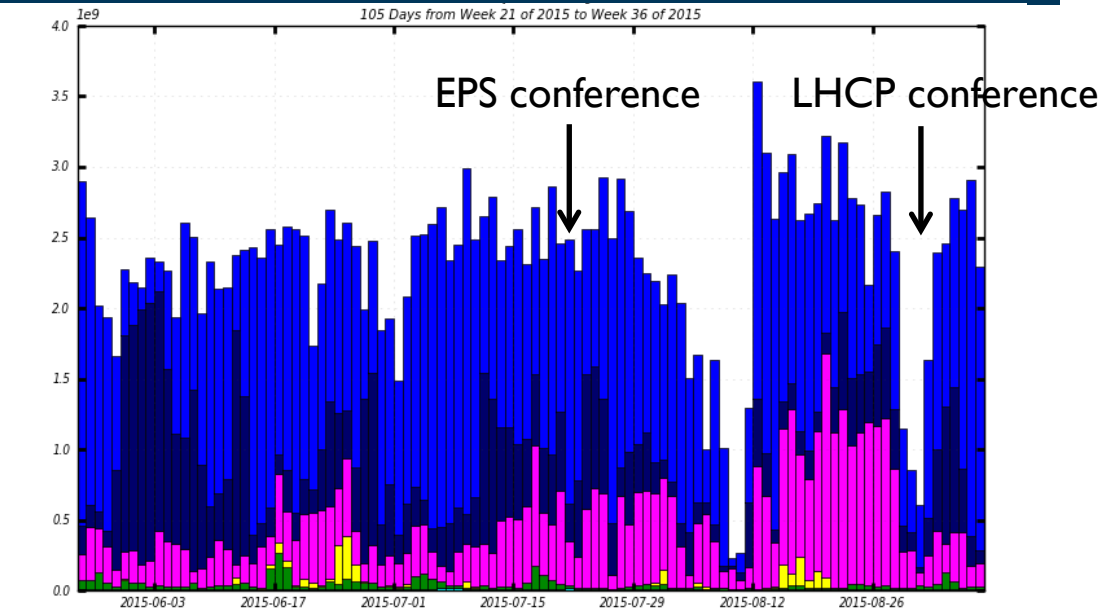
Jobs / CPU usage

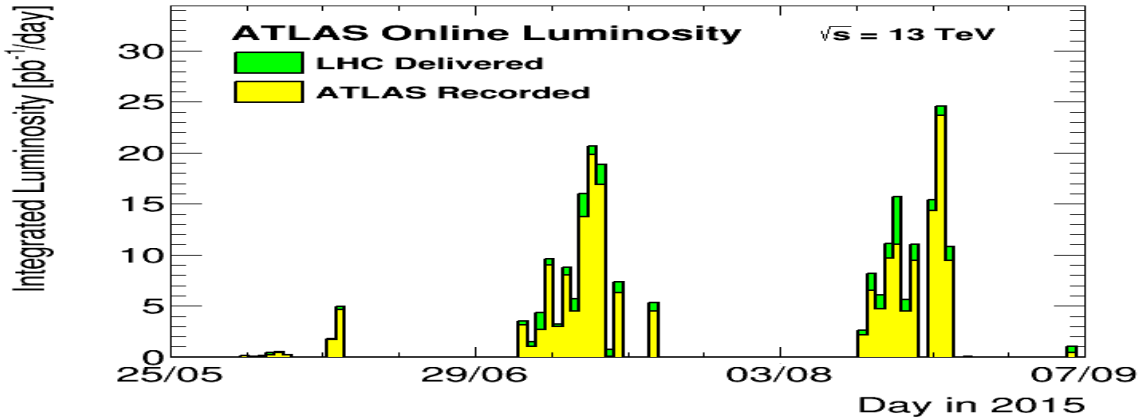
PanDA

- PanDA is the production and Distributed Analysis system
- Migrated in July 2014, stable running.



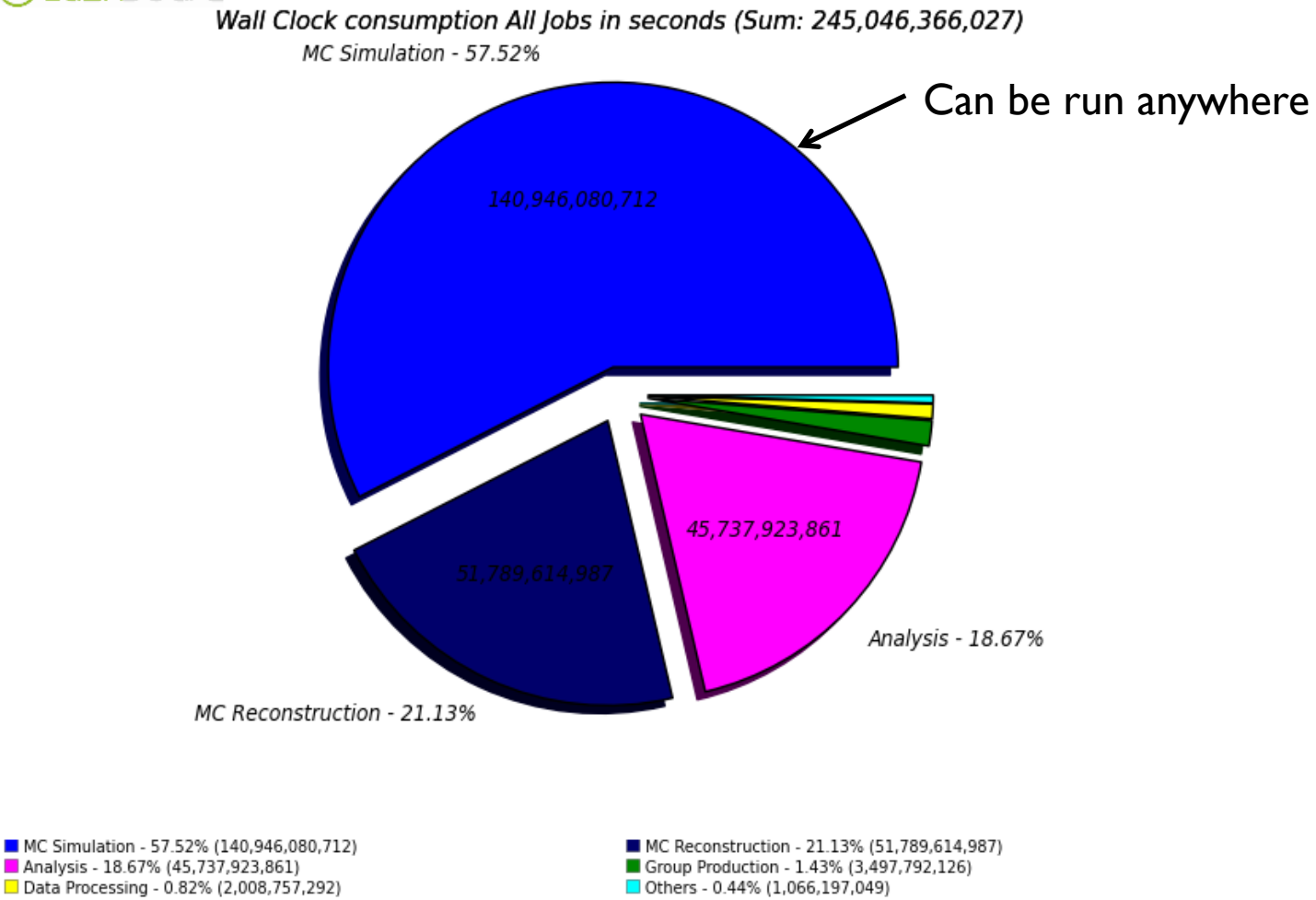
ATLAS Jobs in 2015





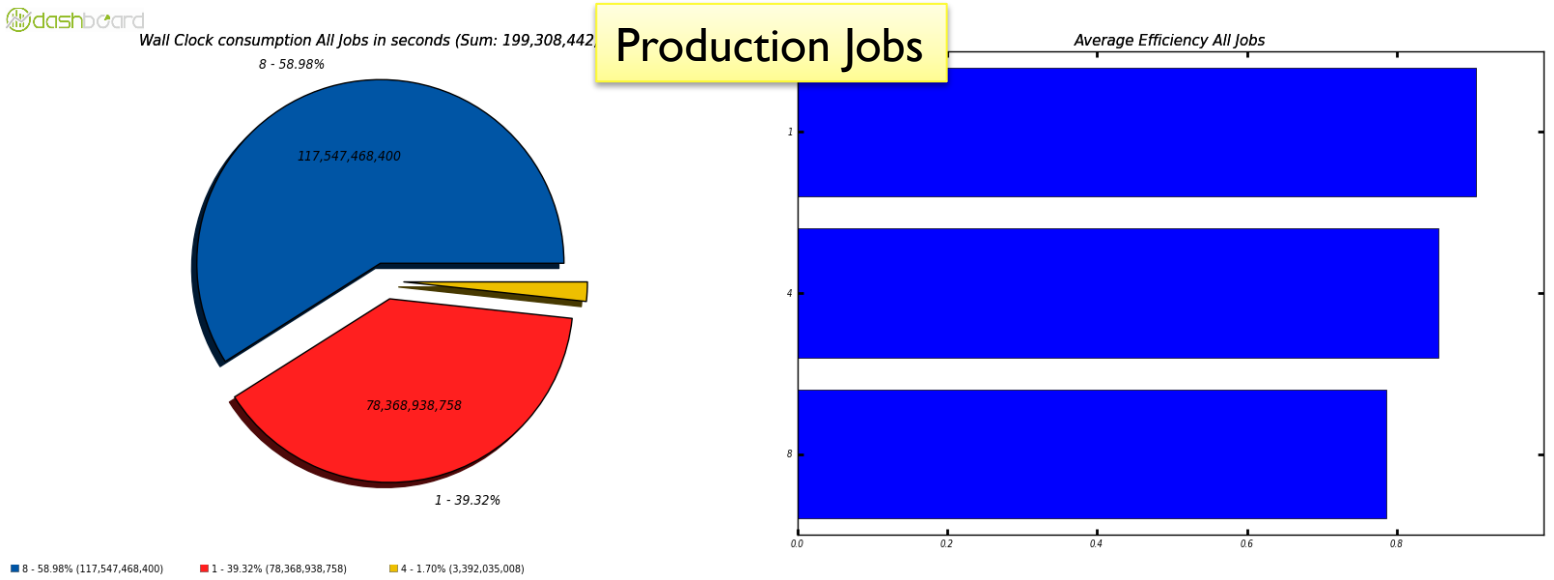
10th September 2015

ATLAS Job types

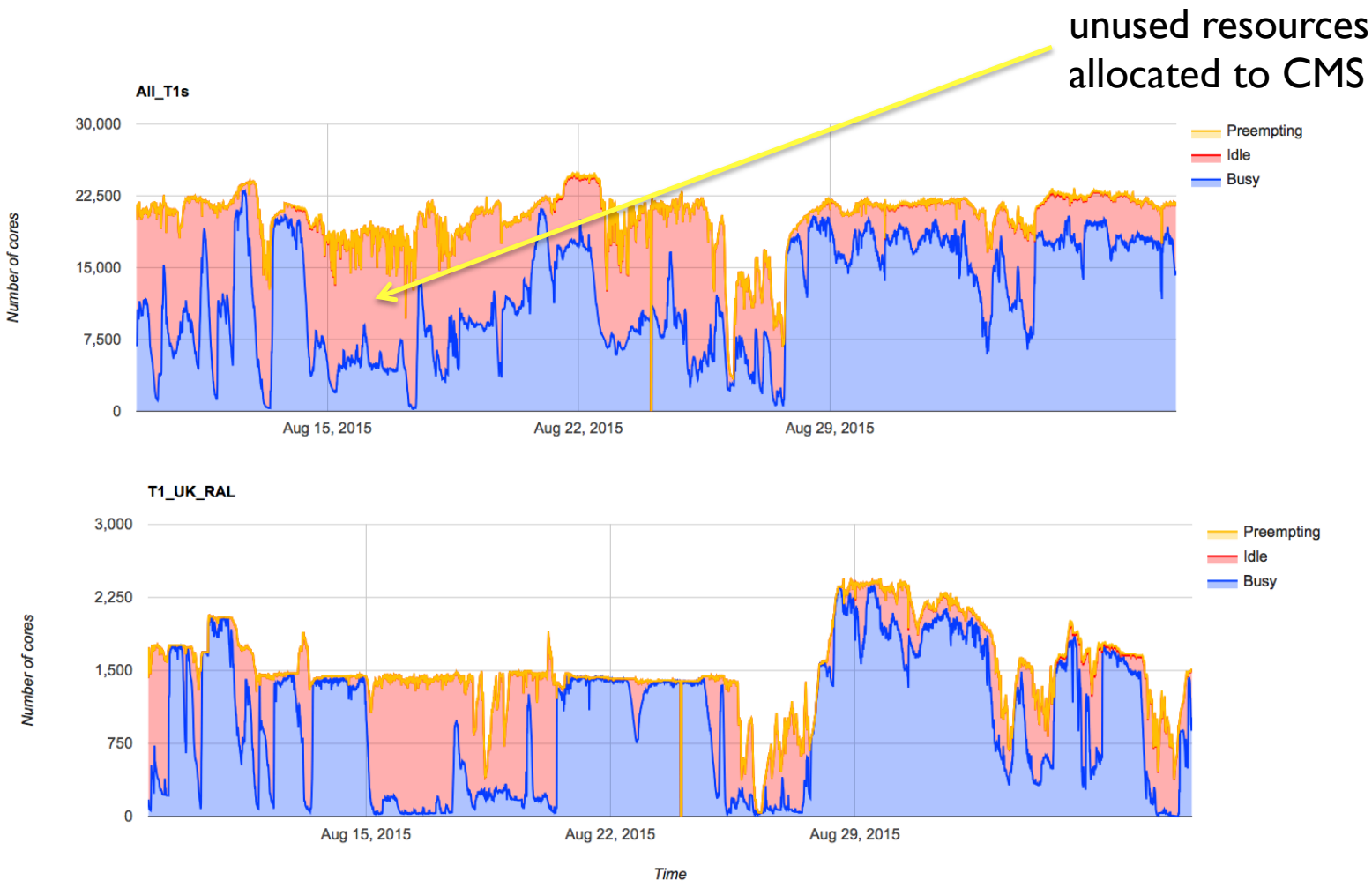


ATLAS MultiCore jobs

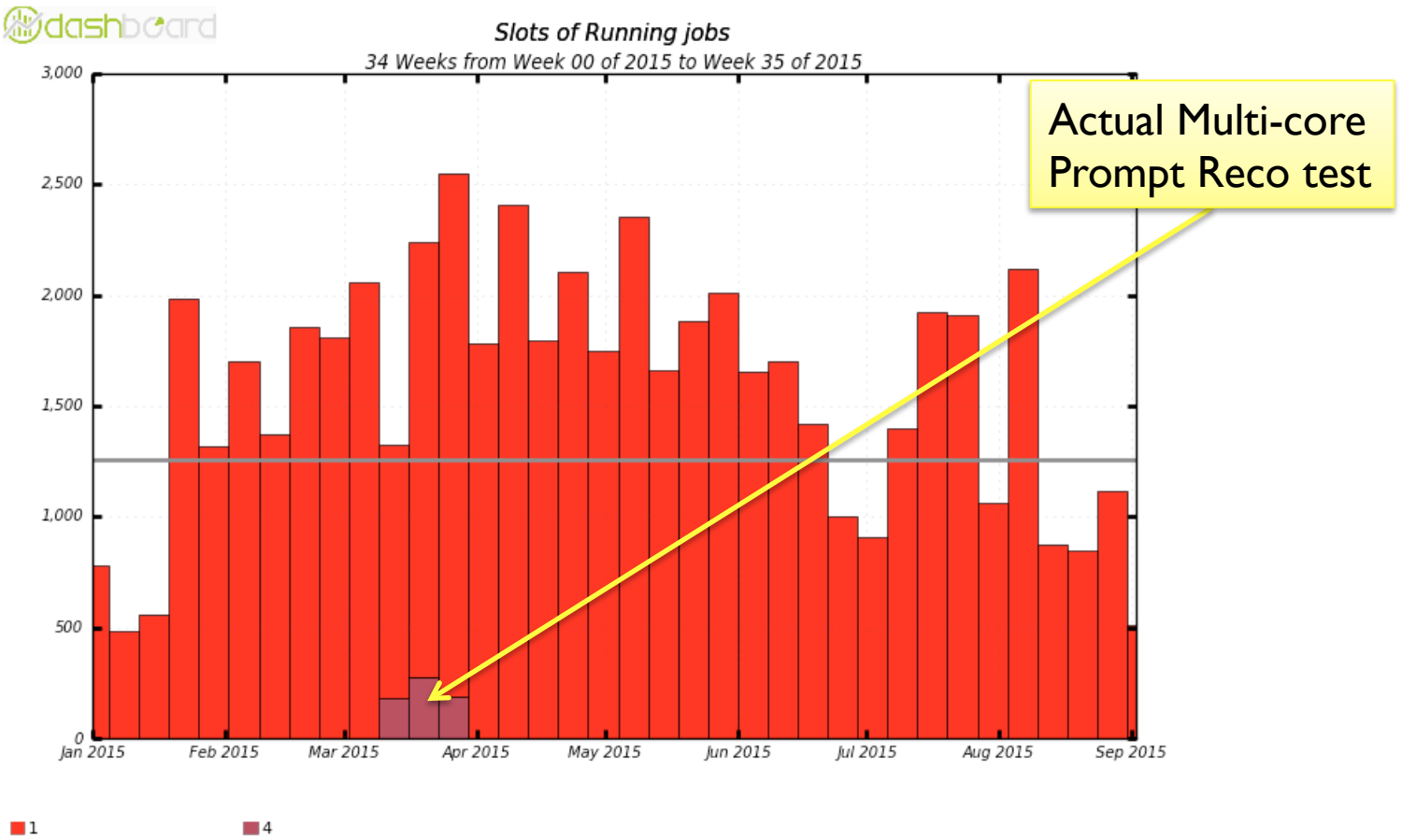
- AthenaMP is the name of the ATLAS multi-process software.
  - Less memory hungry per core
- Most production work now MultiCore
- Some analysis use cases but want to avoid having another queue at all sites.



CMS multi-core ‘jobs’



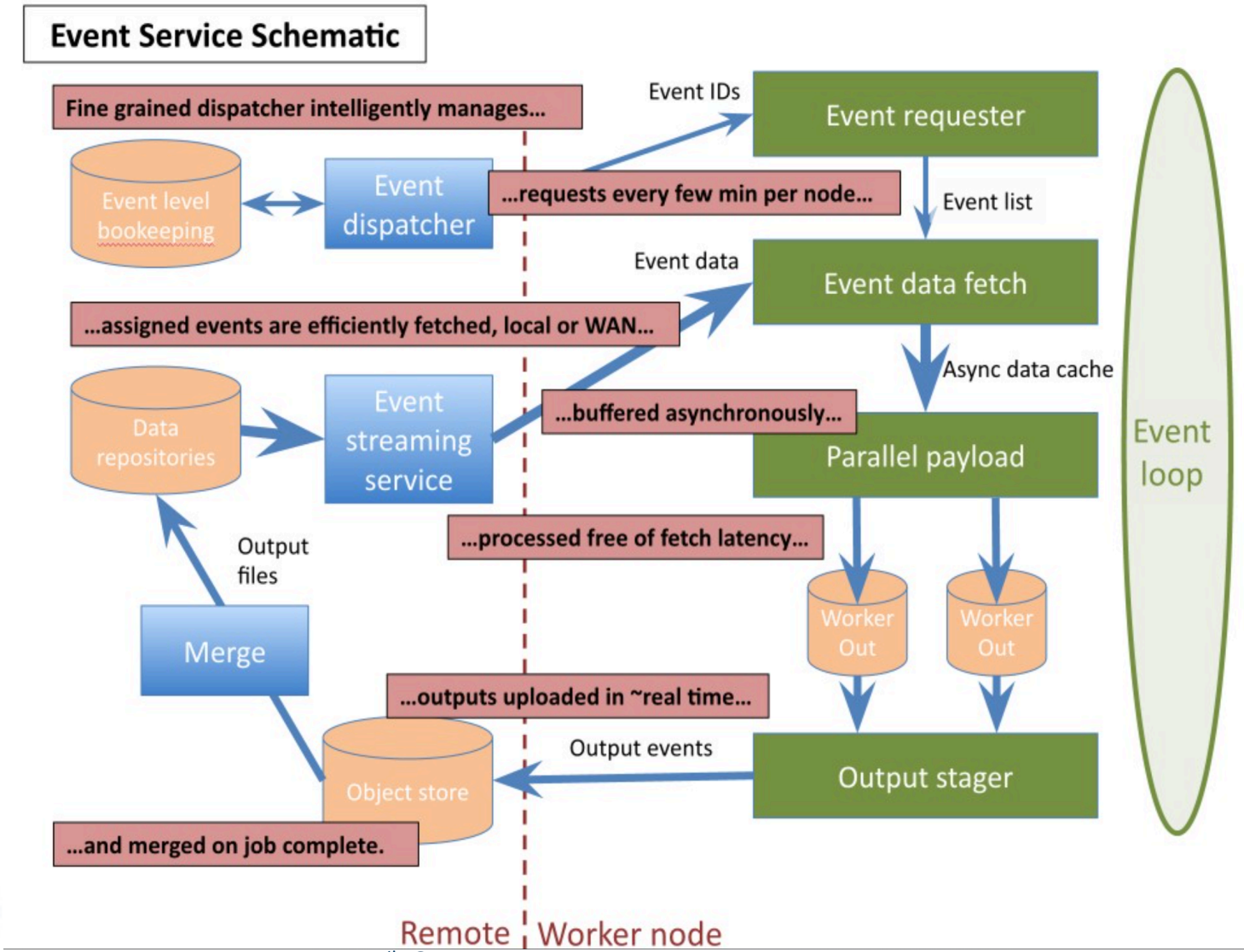
CMS multi-core jobs at RAL



10th September 2015

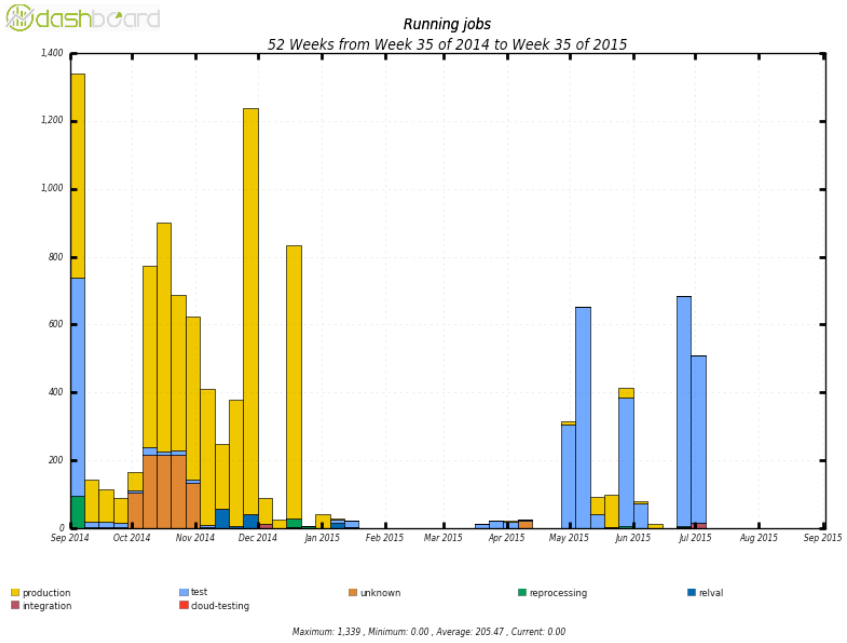
New things

ATLAS Event Service

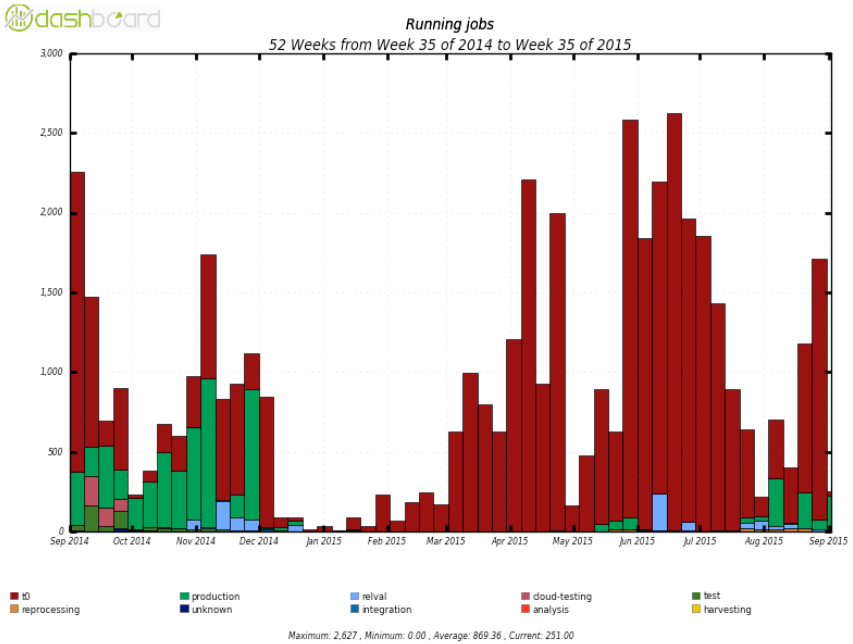


Cloud usage in CMS

- In production:
  - CERN AI (CMS Tier-0)
  - HLT cloud

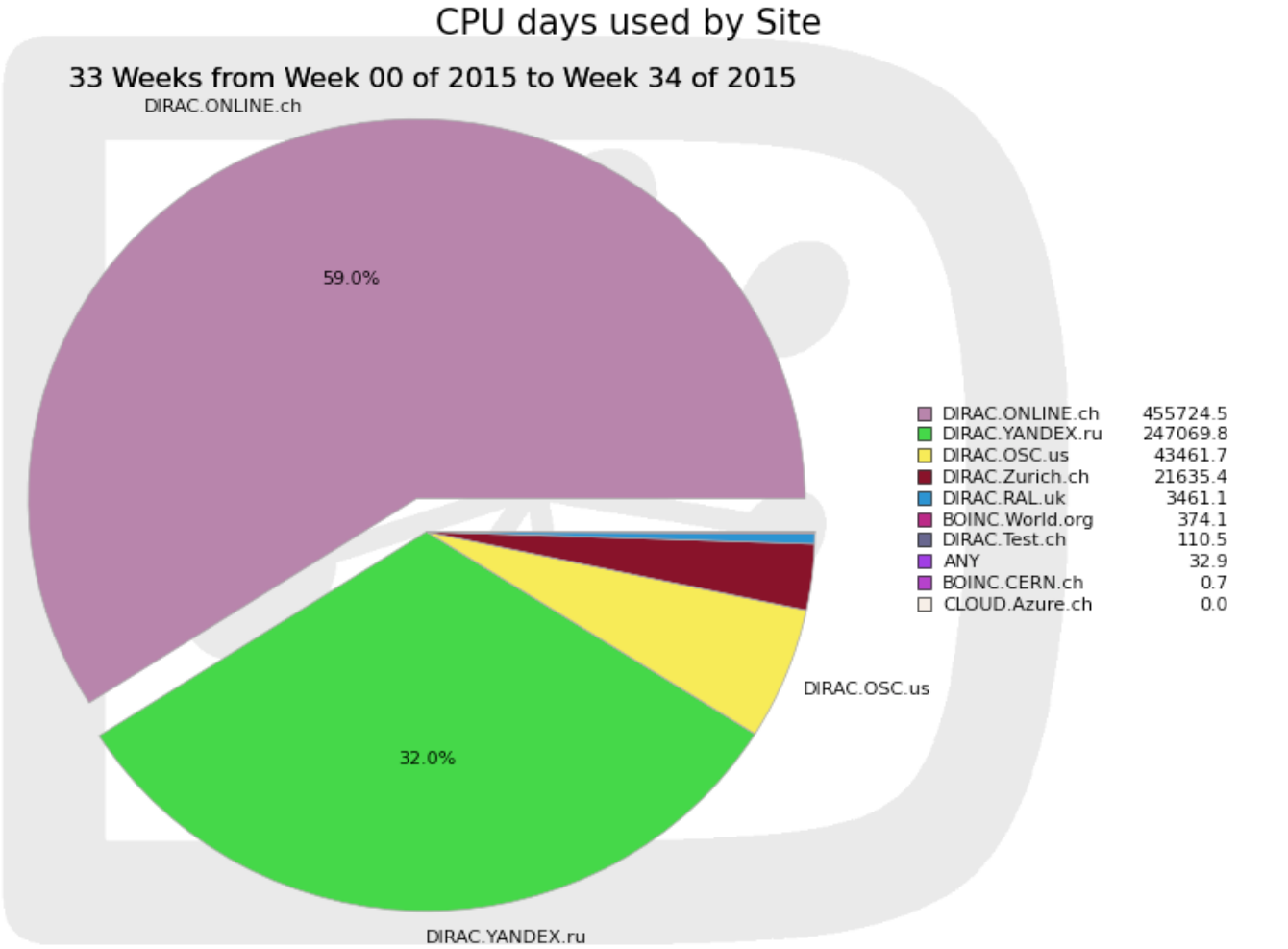


HLT cloud



CERN OpenStack

LHCb Non-pledged



Summary

- Computing models have been updated for run 2.
- Computing didn't stop during the LHC downtime
  - Even more to do now run 2 has started.
  - Very fast turn around between data taking and producing results
- Many improvements still possible.