Meeting with LHCC

M. Girone, D. Lange, D. Bonacorsi

September 22nd, 2015

Through LS1 into Run-2

Computing requirements in Run-2 considerably larger than Run-1

- event rate to storage 1 kHz (~x2.5), higher PU (~x2.5)
- without any improvement after Run-1, we would have needed ~x6 increase in CPU for reco

Goal is to fully realise the physics potential of the experiment by deploying sufficient resources, but also by capitalising on efficiency gains obtained while running the currently deployed systems.

Resources at the Run-2 start-up:

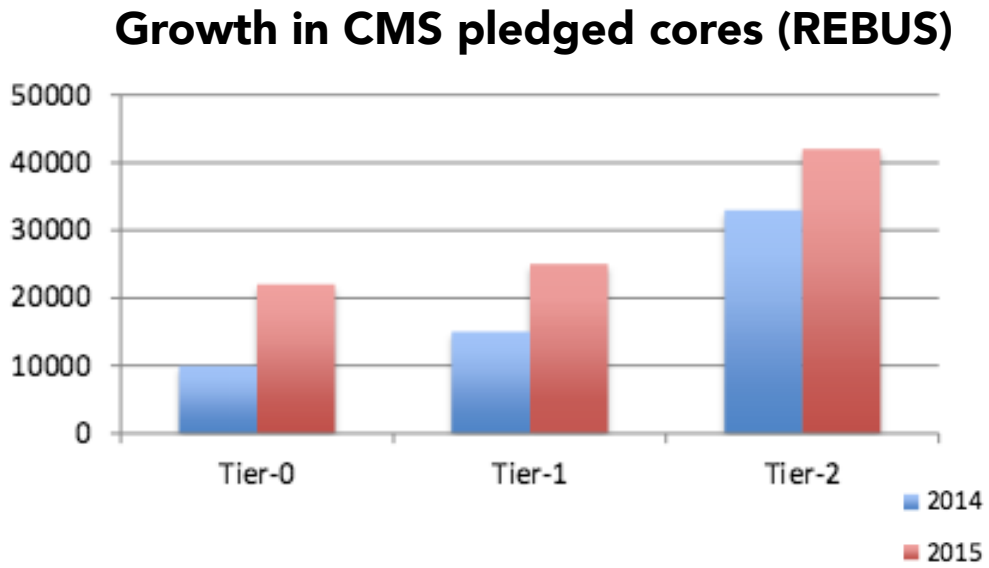
- processing capacity +>50%
- disk capacity +17%
  - (~doubled T0, slower ramp for T1/2 - bigger increases expected in 2016/17)
- tape capacity +35%
  - (Run-1 and before are included)

LS1 was used to prepare for Run-2 (e.g. **threaded framework**, **reconstruction code improvements**, …) and to modernise our computing by adding (in several ways) increased **flexibility in the model**, thus containing the resource requests.

Resources in 2015

CERN

- CPU 22k cores as Tier-0
  - 15k - sometime - as HLT
- disk 15 PB, tape 31 PB
- network 10-100 Gbps to T1s



Tier-1

- 7 facilities primarily at national labs or large computing centres
- CPU ~25k cores, disk 27 PB, tape 74 PB
- network 1-100 Gbps to T2s

Tier-2

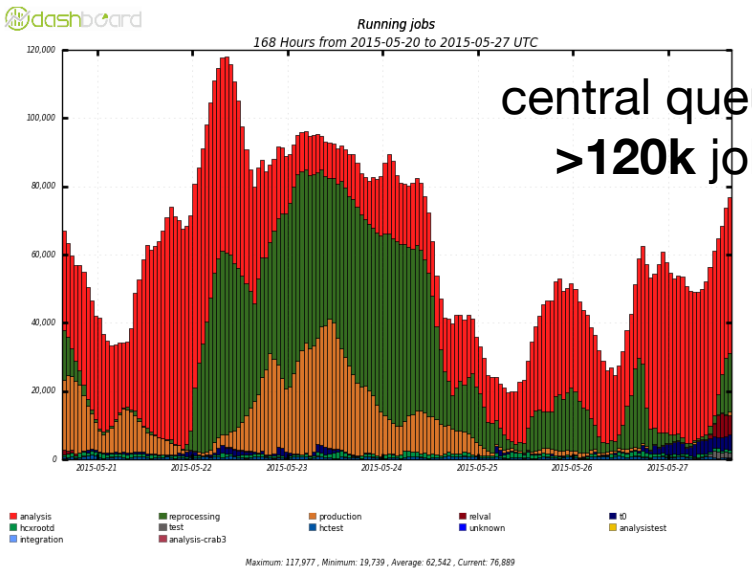
- ~50 facilities primarily at university centres
- CPU 80k cores, disk 31 PB

T1/T2 resources utilisation

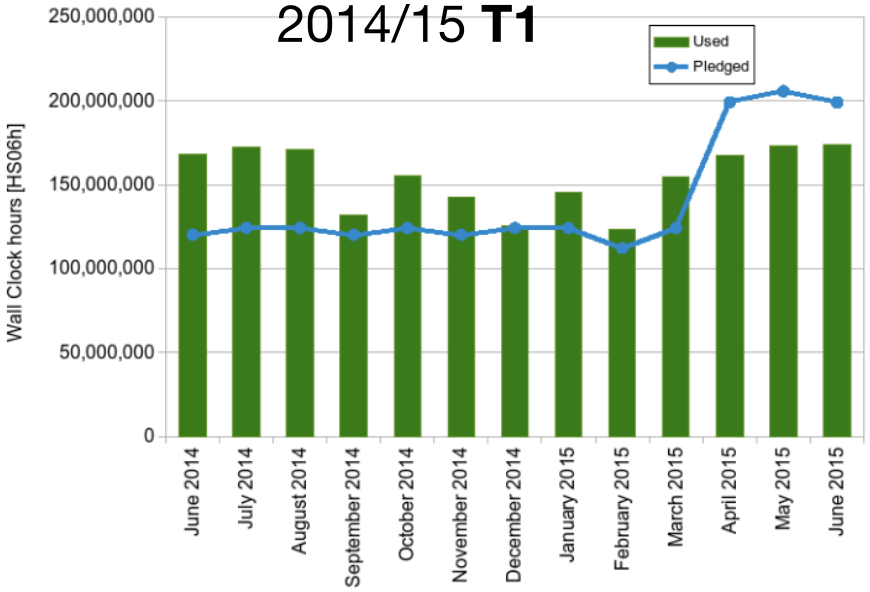
Average use of Tiers over the year:

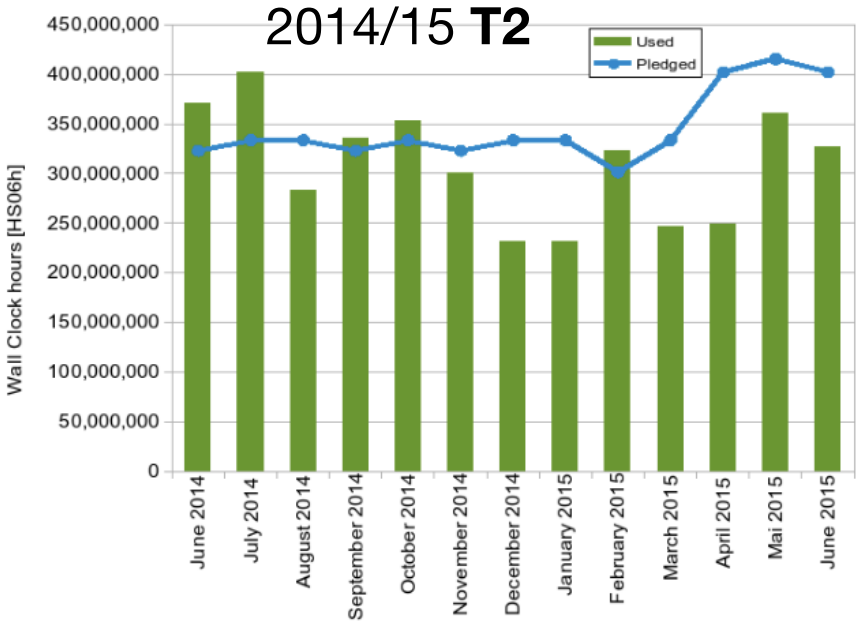
- T1s: **108%** of the pledge
- T2s: **88%** of the pledge
  - Wide variation in average use across countries

Even with changes in machine performance and improving code the CMS resource utilisation remains **high**



central queue peaked at **>120k** jobs running





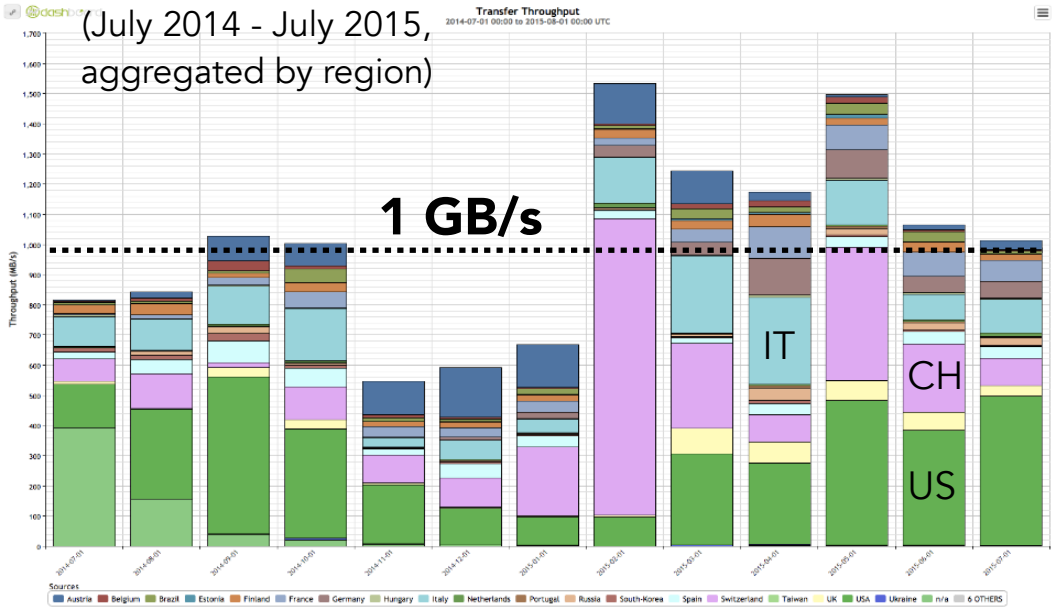
Data transfers

PhEDEx resumed robust operations also in Run-2:

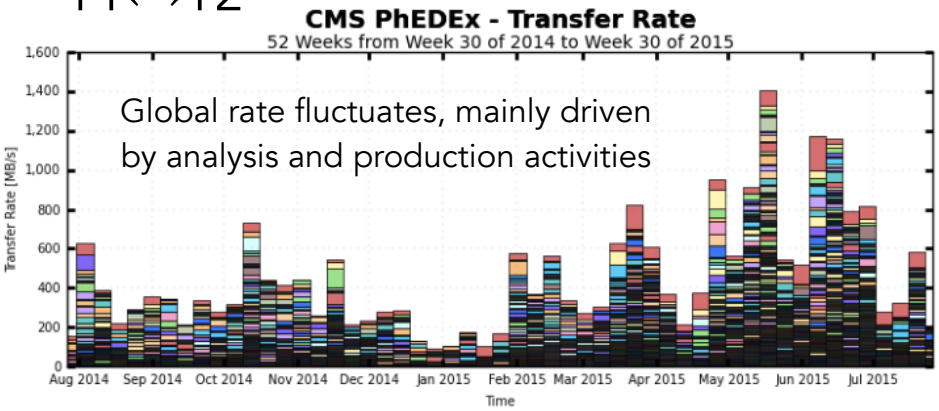
- Run-1: moved 150 PB
- Run-2 so far: stably ~**2 PB/week** among ~60 sites, efficiency at >95%, >3.3K commissioned links in the PhEDEx topology

Higher activity visible as Run-2 started

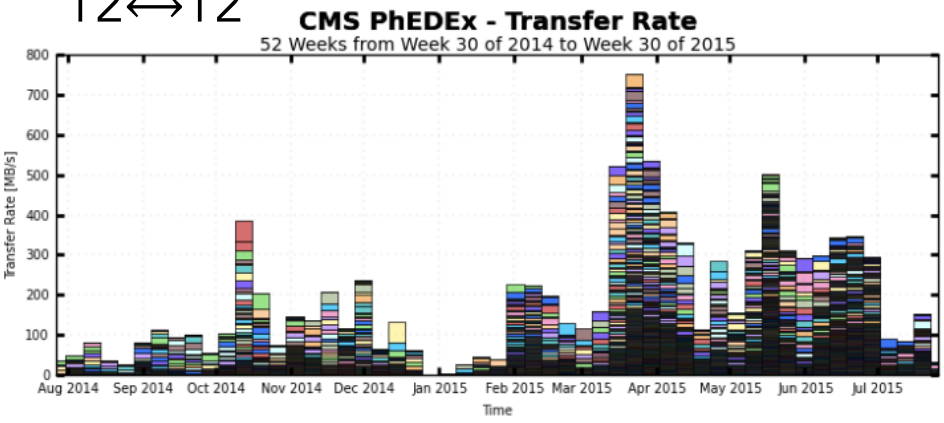
xrootd monthly traffic



T1↔T2



T2↔T2



Model evolutions towards flexibility

Anydata, Anywhere, Anytime (**AAA**) Data Federation at full speed

- CMS applications can read data efficiently over WAN
- Relaxation of constraints on datasets location and workflow execution

Disk-tape separation at Tier-1 sites

- More control over what datasets are available on T1 disk
- Through AAA, T1 data can be used in workflows anywhere

A more **Dynamic Data Management**

- Automatic transfers of new datasets, deletions of less useful replicas, replication of most popular datasets
- Optimised use of disk space at all Tier levels

Global Pool for resource provisioning via glideInWMS

- Allows central control of job priorities, simplified infrastructure
- Demonstrated scaling to operate all T1/T2/opportunistic resources in a single pool

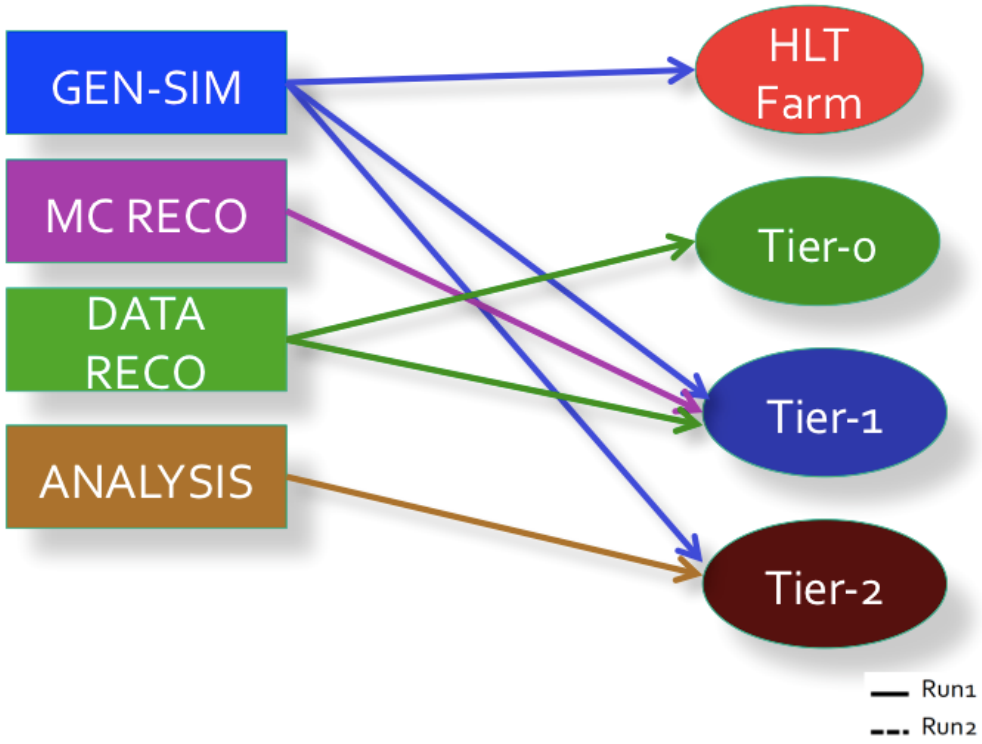
Ability to provision **cloud infrastructures** via glideinWMS

- Allows use of HLT and potentially opportunistic and commercial clouds
- Ability to burst into extra resources if necessary

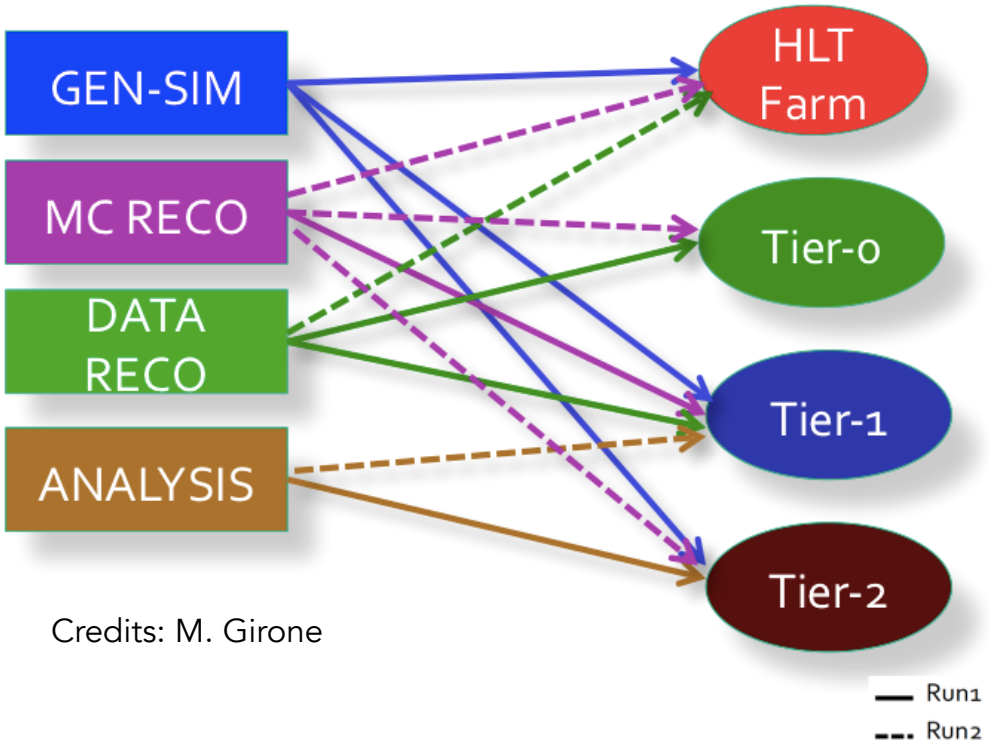
MiniAOD: new analysis format for Run2

- Compact format: ~30-50 Kb/evt (10% of the analysis datasets used in Run1)
- Can serve ~80% of all CMS analyses

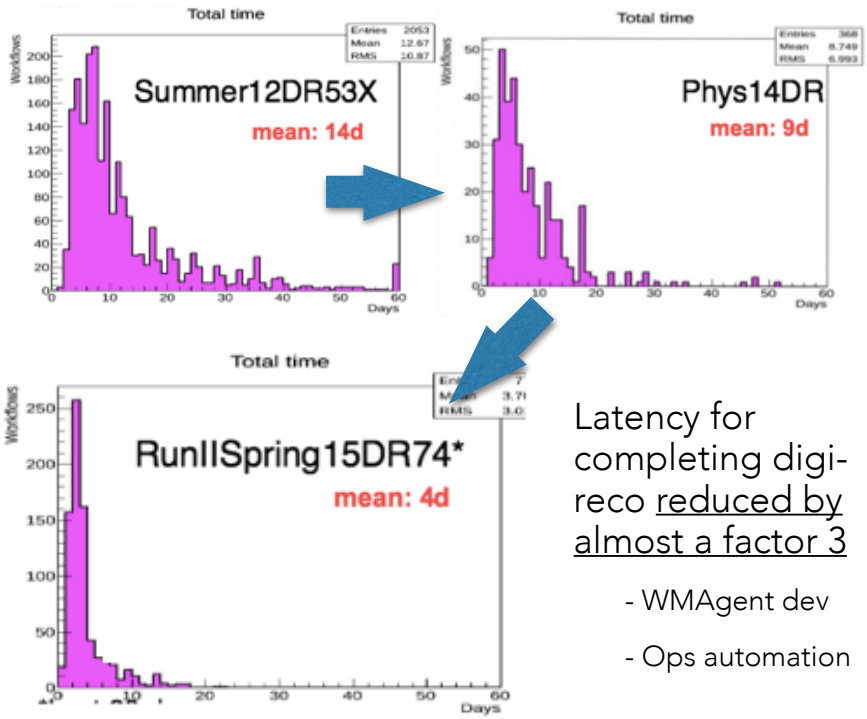
Impact of facilities and workflows



Impact of facilities and workflows



Credits: M. Girone



Latency for completing digi-reco reduced by almost a factor 3

- WMAgent dev
- Ops automation

Larger flexibility for facilities operations and workflows execution

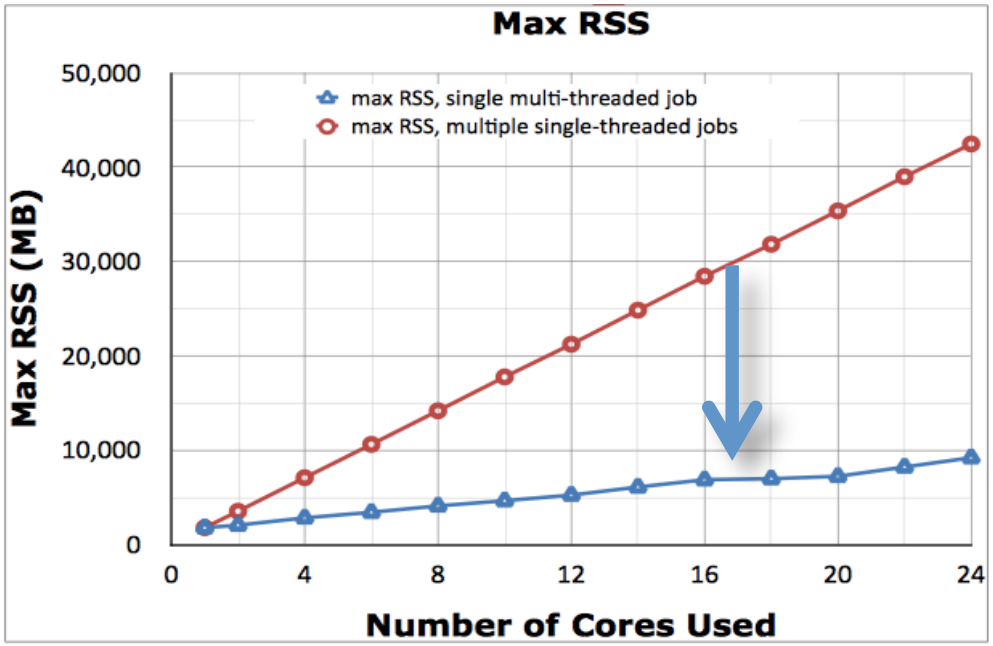
- transparent access to CMS data across the Grid thanks to the **Data Federation**
- **improved networks** have been key to this
- one **central queue** for all resources and workflows facilitates prioritisation (e.g. analysis vs prod)
- integration of the **HLT farm** (size ~Tier-0) outside of LHC running
- T2 sites commissioned to perform reco, previously reserved to T1s
- Analysis jobs can run at more than T2 sites

Breaking boundaries among Tiers, less restrictions than ever before

CMS T0 multithreaded in 2015

Threaded framework and reconstruction algorithms developed during LS1 now deployed in production at Tier-0

- large Run-2 trigger rates processed with low latency
- dramatic memory saving from threading



Memory savings allow us to produce a full suite of outputs directly from the PromptReco application

- Reconstructed events in full format and two analysis formats (AOD and MiniAOD)
- Monitoring histograms for data certification
- Detector and physics skims

On-going development priorities for 2015

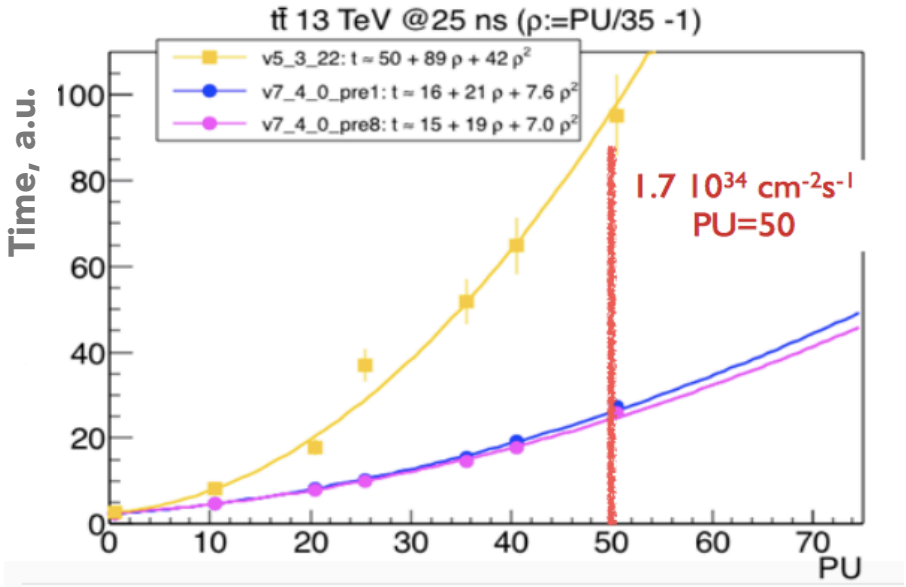
- Finalize 25 ns reconstruction configuration
- Evolve MiniAOD format to follow analysis needs. As MiniAOD can be re-derived from the AOD format, CMS can quickly reproduce them upon needs

Gains in CPU usage for SIM/RECO

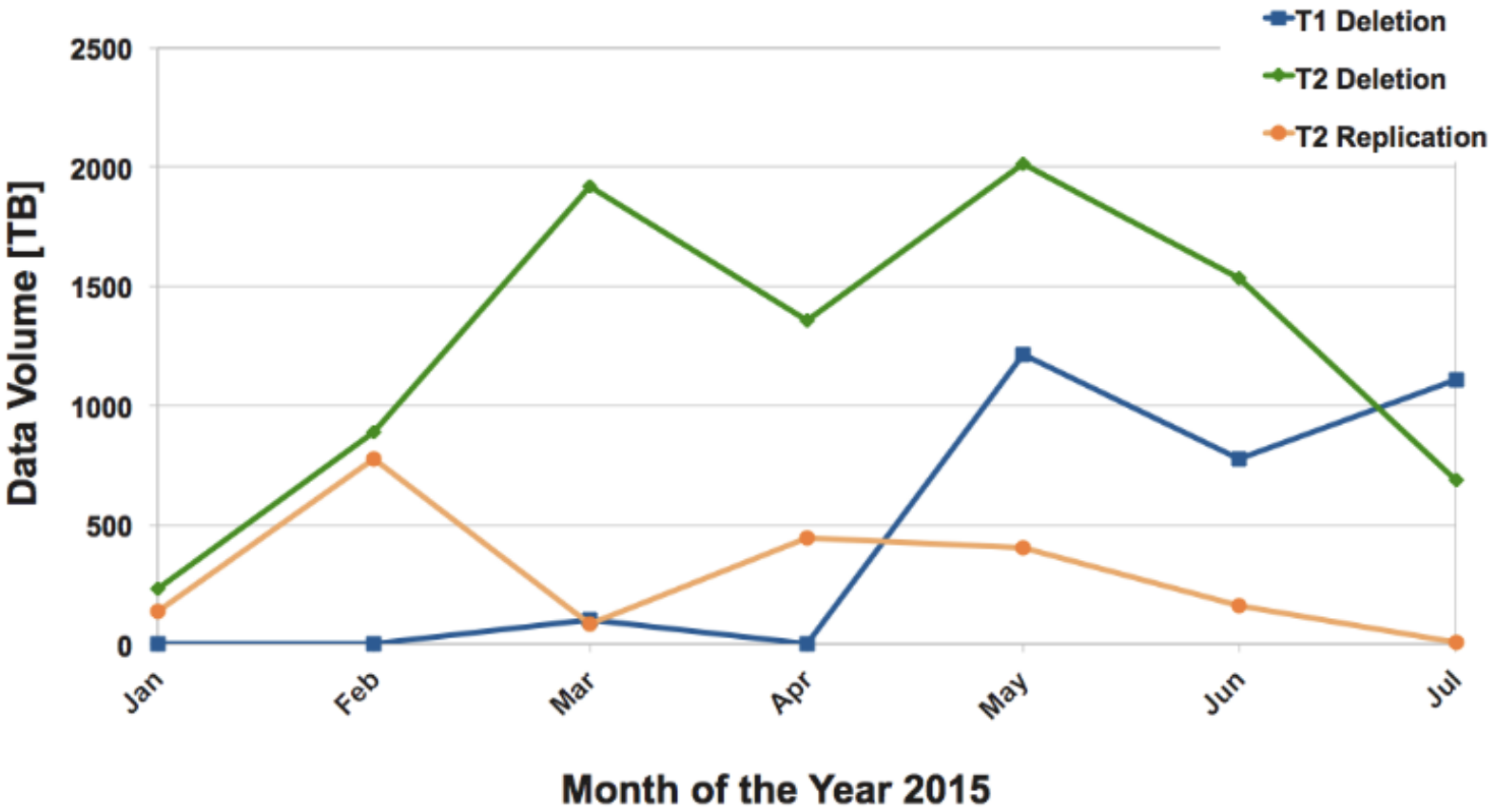
Large technical performance gains achieved during LS1

- Simulation: Factor of 2 gain in CPU utilization, primarily from Russian Roulette sampling algorithm to reduce time spent tracking low-energy particles in Geant4
  - Visible improvements already in the number events/month produced for CSA14(CMSSW6_2) and 2015 production RunIIWinter15GS (CMSSW7_1)
- Reconstruction: Large gains, particularly in tracking area and algorithms appropriate for 25 ns conditions

Crucial achievements to face Run-2 challenges within resource constraints



Dynamic Data Management



Goal: optimise resource utilisation (disks at Tier-1/2)

In production since Jan 2015. Outcome of first 6 months of operations:

- dynamic deletions: **3.2 PB** (T1 disk-only) and **8.6 PB** (T2) deleted
- dynamic replications: **2 PB** of most popular datasets (T2)

Resource requests 2016/17

Resource Utilisation and Resource Requests docs submitted to the C-RSG at the end of August

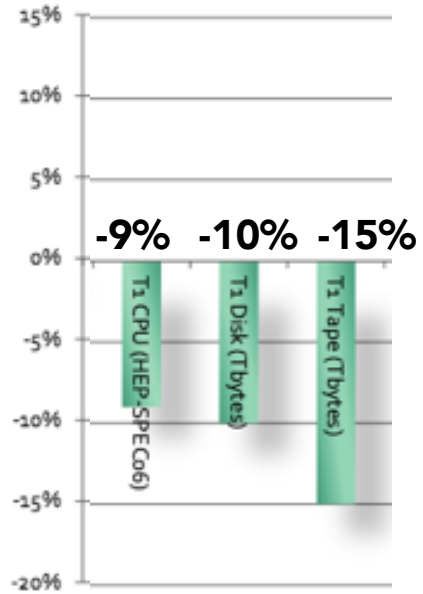
- currently being scrutinised (-> October RRB)

For 2016 we are systematically low on T1 pledges (and it is getting worse over years)

- part of the T1 deficit mitigated by some T2 over-pledge (e.g. CPU) in some regions
  - this emphasise again the value of adding flexibility to our workflow execution

Pledged vs Required

2016



in 2015 was:	**-8%**	**-7%**	**-10%**
in 2014 was:	**-1%**	**-8%**	**-12%**

Requests for 2017 also include parking and Upgrade simulations

- balanced with the evident improvements on the software performance side
- summing all up, main request is disk at T1 sites

MiniAOD actual impact on the CMS evolved model will be learnt as Run-2 progresses

- on this, a change w.r.t Fall 15 requests is possible, on the time scale of the April RRB

CMS resource requests (2015/17)

	Pledge 2014	Increase from 2013	Pledge 2015	Increase from 2014	2016 (C-RSG Apr 15)	Increase from 2015	2017	Increase from 2016
Tier-0 CPU (kHS06)	121	0%	256	111%	292	14%		
Tier-0 Disk (TB)	7000	0%	3200	Reallocated to CAF	3200	0%		
Tier-0 Tape (TB)	26000	0%	31000	31%	38000	23%		
T1 CPU (kHS06)	175	0%	300	71%	400	33%		
T1 Disk (TB)	26000	0%	26000	4%	35000 (33000)	30%		
T1 Tape (TB)	55000	11%	74000	34%	100000	35%		
T2 CPU (kHS06)	390	14%	500	25%	700	40%		
T2 Disk (TB)	27000	4%	29000	16%	40000 (38000)	37%		

2017: a modest increase - inst. lumi not expected to change significantly in 2017 wrt 2016

2017: no need for increase - trigger rate is the same, small change in the size

2017: increase - driven by the new data

2017: increase, coming on the tails of a previous +33% - need for reprocessing capacity for data and simulation

2017: increase, coming on the tails of a +30% - reaches a better balance of CPU and disk

2017: increase, custodial storage of new data

2017: a modest increase (less than before)

2017: same as CPU

CMS analysis model concentrates primarily on current plus previous year's data. 2016 is the largest relative increase

Subject to changes for the April RRB:

- if the <PU> increases
- on the experience we will have collected with the new MiniAOD format