Status of CMS

*Mostly based on information from CMS Offline & Computing Week November 3-7

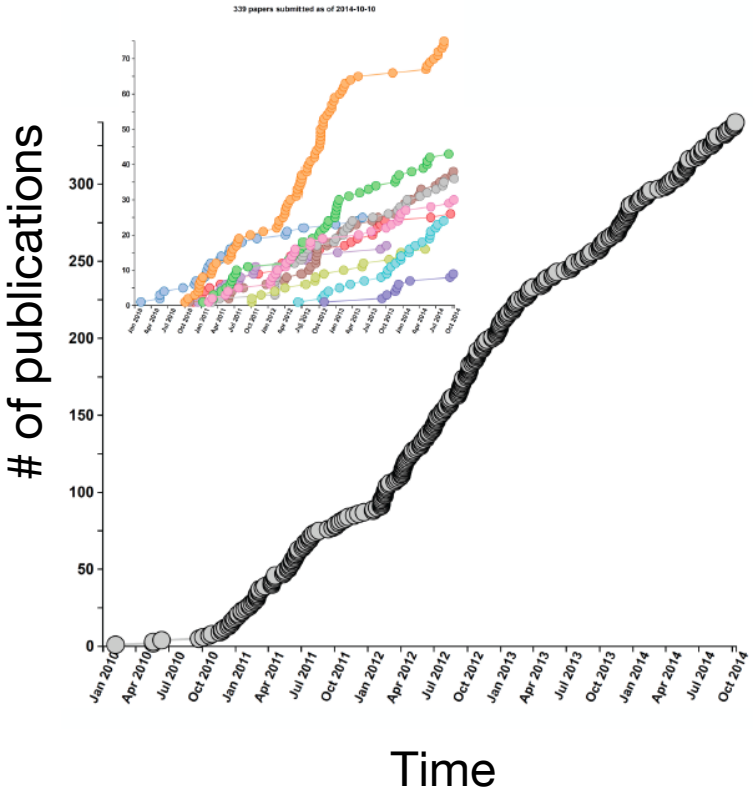
Matthew Nguyen

Recontres LCG-France

December 1st, 2014

Winding down Run 1 physics

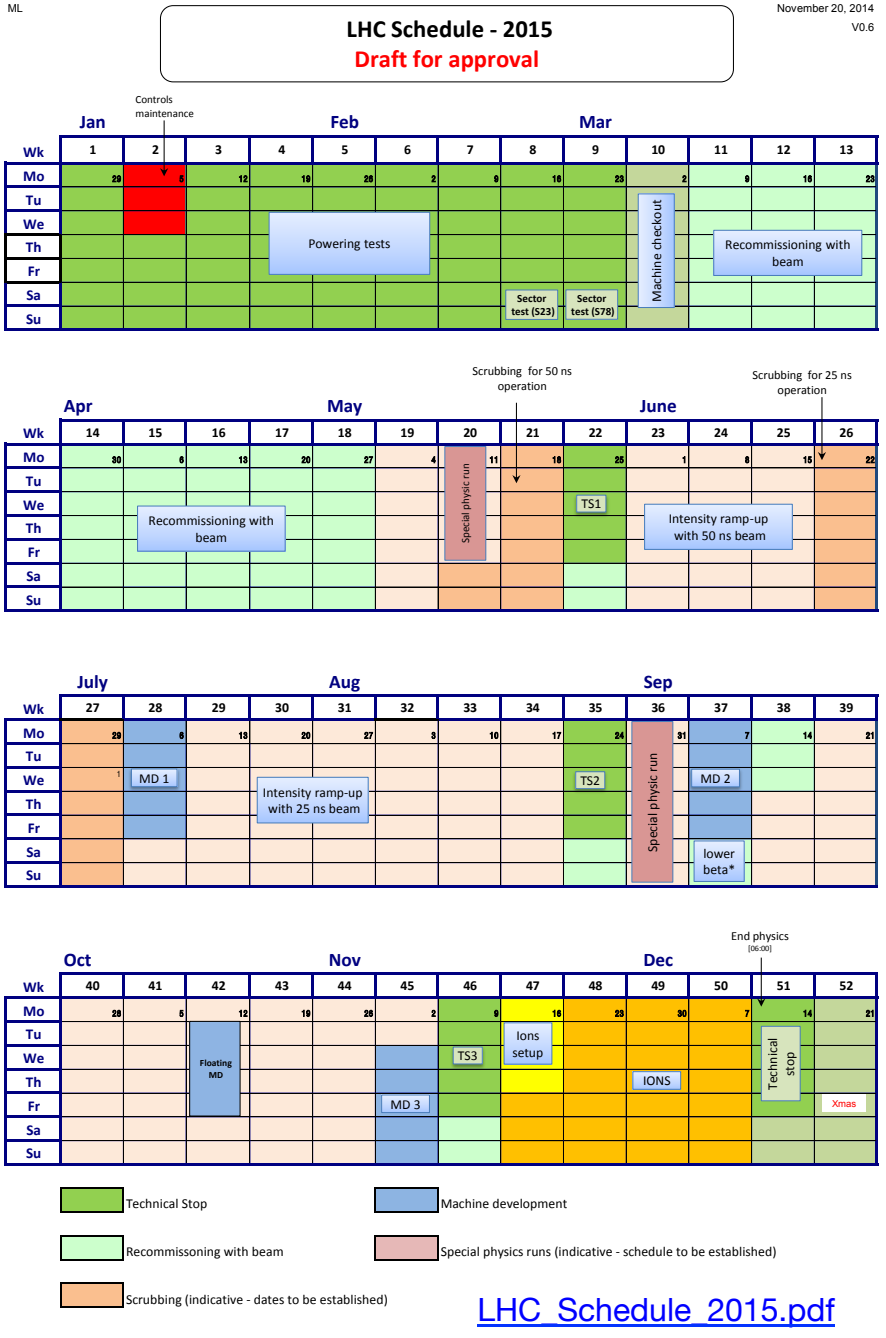
- Mostly ‘discovery physics’ groups (Higgs, SUSY, exotica) winding down publication Run 1 publications this year
- Others (standard model, b-physics, heavy ions) continue to publish Run 1 physics into 2015
- No remaining large MC requests or reprocessing
- No major software developments
- Continued need or analysis resources into 2015



LHC schedule (as of today)

- June 1: Start of collisions for physics
- 3 weeks of 50ns bunch spacing (as in Run 1)
- Few week technical stop for “beam scrubbing”
- 6 weeks @ 25 ns, large β*
- 3 week technical stop / special runs
- 7 weeks @ 25ns, nominal β*
- 1 month heavy-ion run (PbPb)

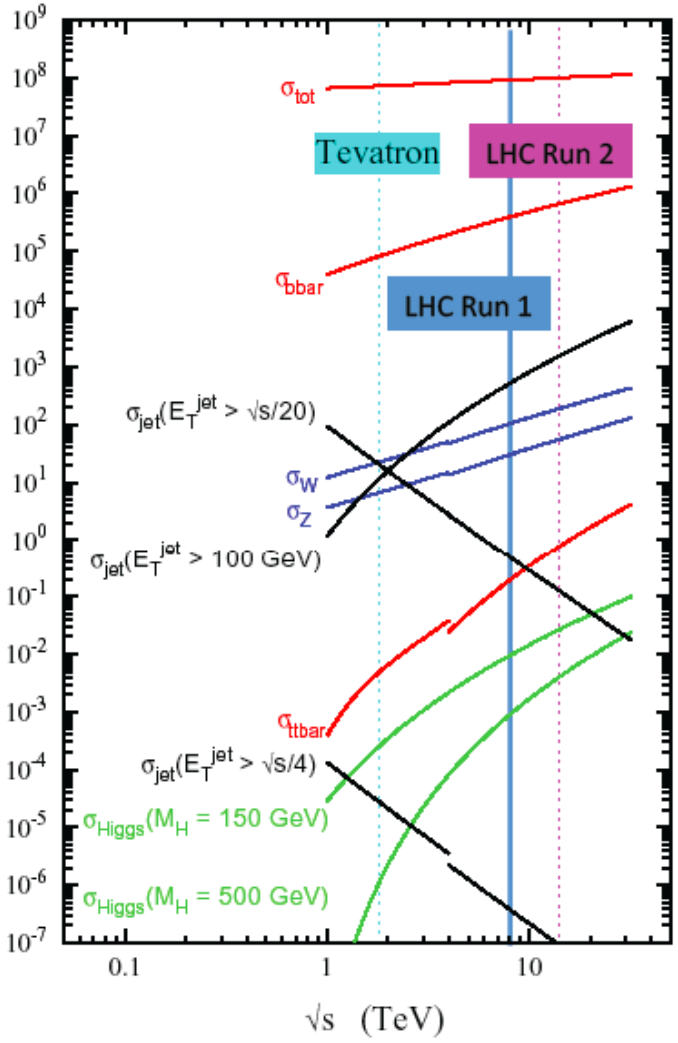
Period	N_{bunch} [10^{11}]	ε* [μm]	k	β* [cm]	L [$cm^{-2}s^{-1}$]	<μ>	Days(*)	∫L [fb^{-1}]
50 ns	1.2	2.2	≈1370	80	5.3×10^{33}	30	21	≈1
25 ns / 1	1.2	2.5	≈2500	80	8.1×10^{33}	26	44	≈4
25 ns / 2	1.2	2.5	≈2500	40	14.7×10^{33}	45	46	≈13



LHC_Schedule_2015.pdf

Run 2 physics goals

- Energy = 13 TeV (still TBC)
- Modest upgrades for CMS from LS1: muon coverage completed, new beampipe, etc.
- Discovery potential depends on channel → specifically 13/8 TeV cross section
- E.g., high mass dijet resonances will be interesting from Day 1
- Precision Higgs meas. by the end of Run 2
- In the early days will have to verify SM predictions at new energy



Challenges for Run 2

- As for all expt's, data volumes increasing faster than resources (Moore's law, flat budgets, etc.)
- Conditions are becoming more demanding, e.g, larger pile-up, with sizeable component coming 'out-of-time' at 25 ns
- LS1 was dedicated to evolution of CMS computing model and software to meet these needs

CSA14

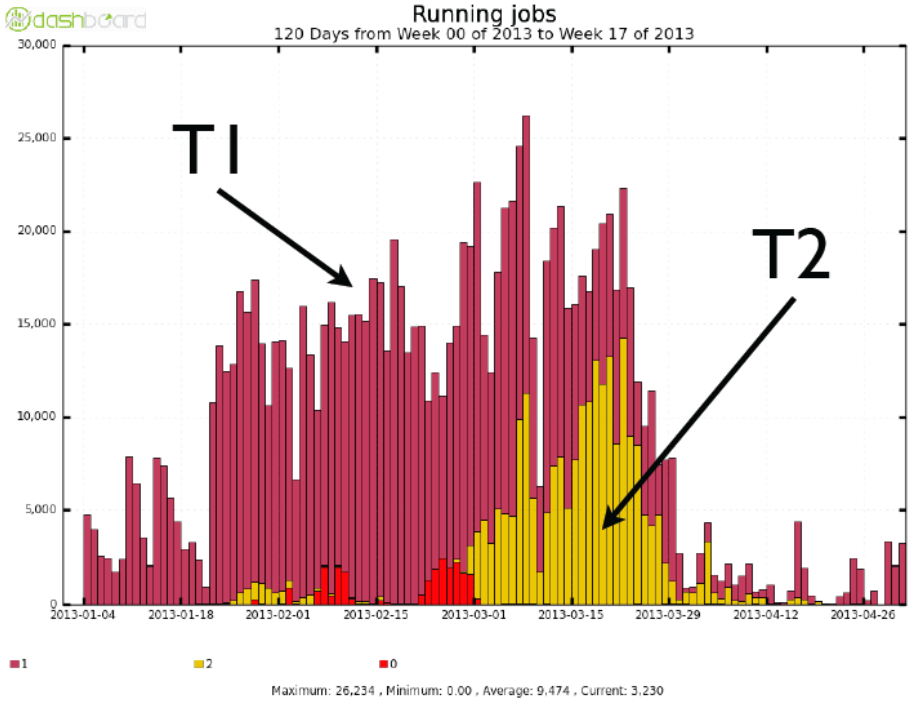
- Computing, Software and Analysis challenge: large-scale tests of complete data processing, software and analysis chain
- Injection of o large samples of simulated events, analyzed using new computing tools and data access techniques
- Samples include different bunch spacing and PU conditions
- Spanned July – September, with > 150 users

2014 Computing Milestones

- ✓ **Data Management milestone: 30 April 2014** **Done!**
  - More transparent data access
    - Disk and tape separated to manage Tier-1 disk resources and control tape access
    - Data federation and data access (AAA)
    - Developing Dynamic Data Placement for handling centrally managed disk space
- ✓ **Analysis Milestone: 30 June 2014** **Done!**
  - Demonstrate the full scale of the new **CRAB3** distributed analysis tool
    - Reduce job failures in handling of data, improved job tracking and automatic resubmission
- **Organized Production Milestone: Ongoing…**
  - Exercise the full system for organized production
    - Cloud-based Tier-0 using the Agile Infrastructure (IT-CC and Wigner)
    - Run with multi-core at Tier-0 and Tier-1 for data reconstruction

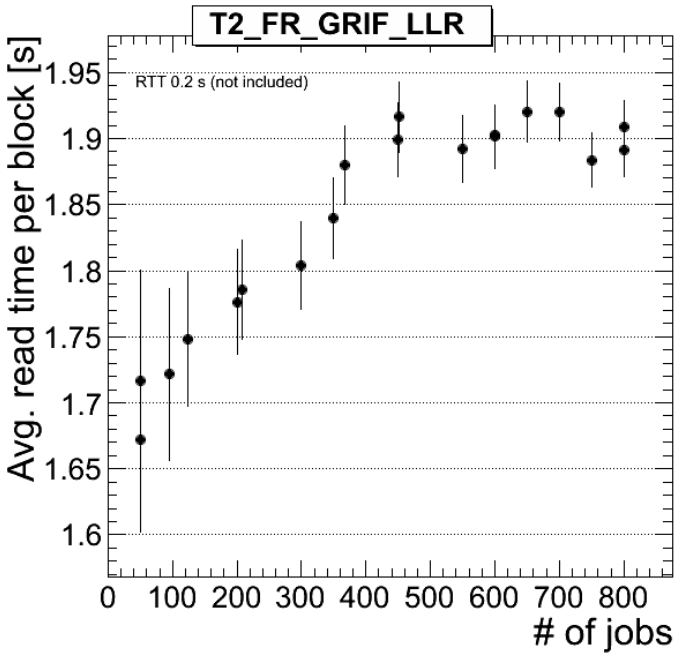
Data Federation

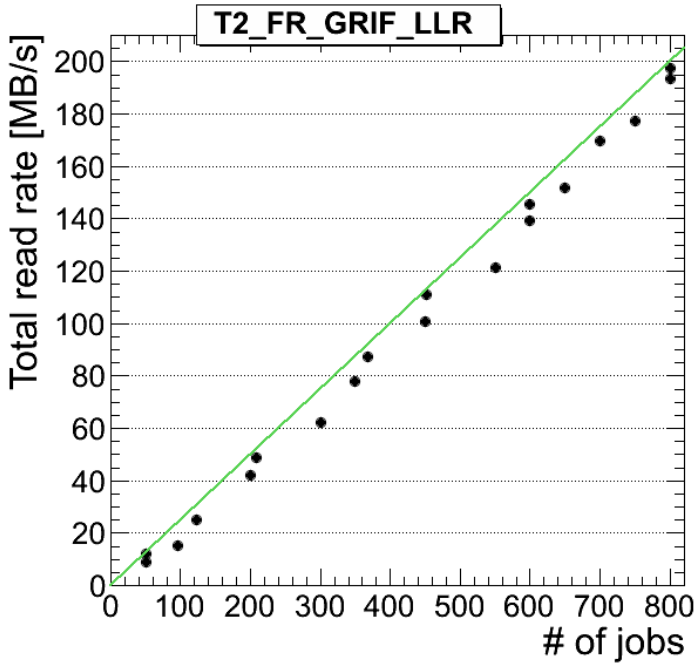
- Relax paradigm of data locality, taking advantage of better bandwidth reliability than anticipated: Any data, Anytime, Anywhere (AAA)
- Implementation:
  - xrootd data federation
  - 'Fall-back' mechanism
  - Disk/tape separation at T1s
- Advantages
  - Slight loss of CPU efficiency, but overall more efficient use of resources
  - Sites w/ storage failures can continue to operate
  - Can imagine disk-less T2s



An early example: "Legacy" reprocessing of 2012 data samples.

Federation scalability

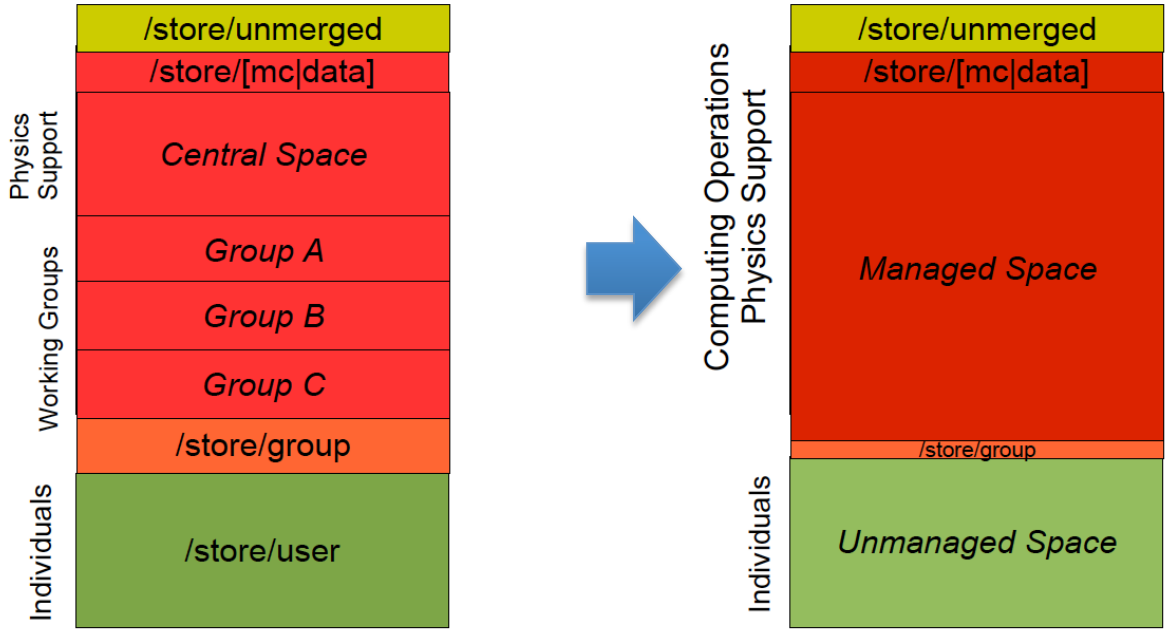




GRIF_LLR scaling very well at least up to 800 jobs

Strategy to optimize performance depends storage system, i.e., may be different at our T2s (DPM) vs T1 (dcache)

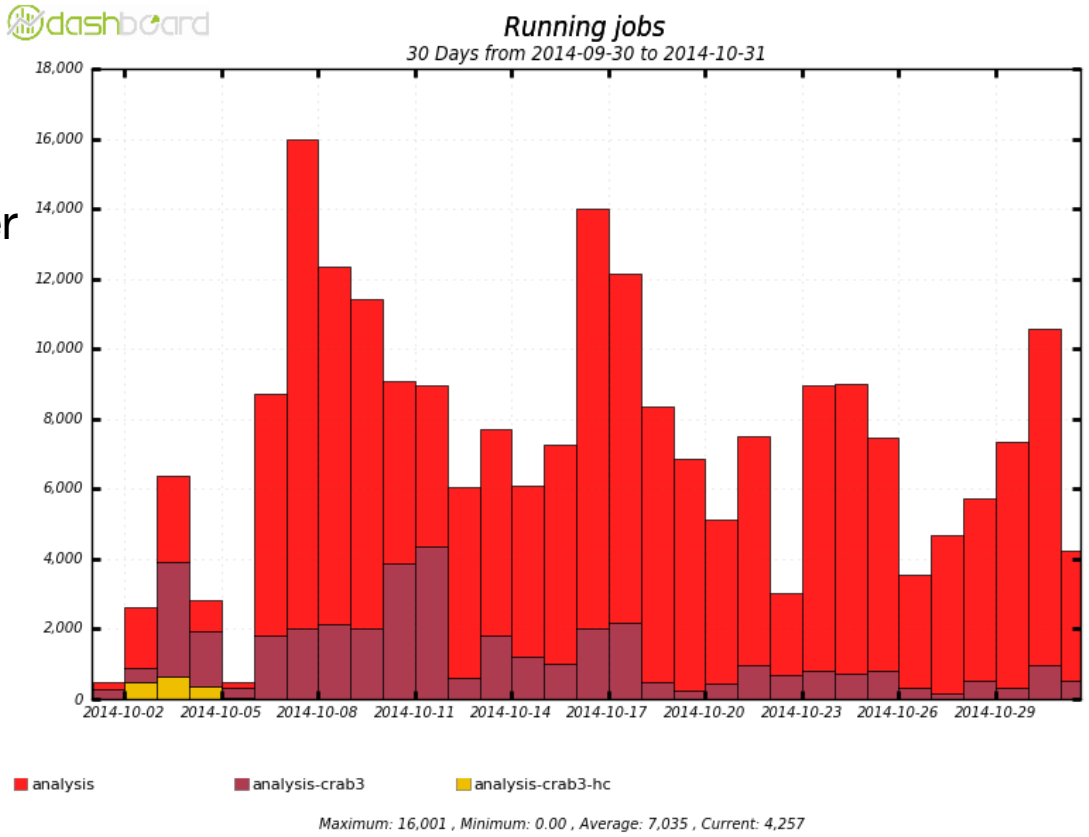
Dynamic Data Management



- CMS manages 100 Pb of disk at 50 computing centers
- ~ 100k datasets (mostly MC), which were distributed essentially by hand
- Idea: Instead use data popularity to determine dataset replication and deletion
  - Release least popular cached copies once 90% of space is used, until 80% usage
  - Create new copies when data becomes popular
- CMSSW reports on file-level access as of version 6

CRAB3

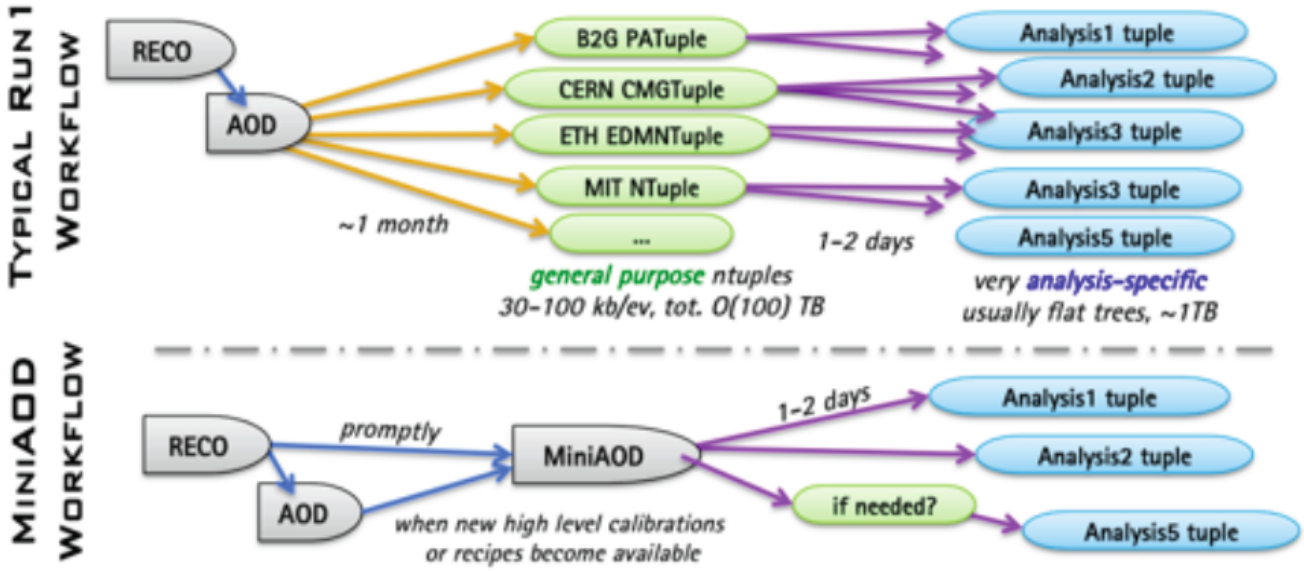
CMS Remote Analysis Builder



- New grid submission tool enables option to ignore data locality, i.e., use AAA
- Tested by artificially forcing jobs to run w/ remote access
- During CSA14: 20k cores in production, 200k jobs/day, average of 300 users/week
- Improves handling of read failures and monitoring

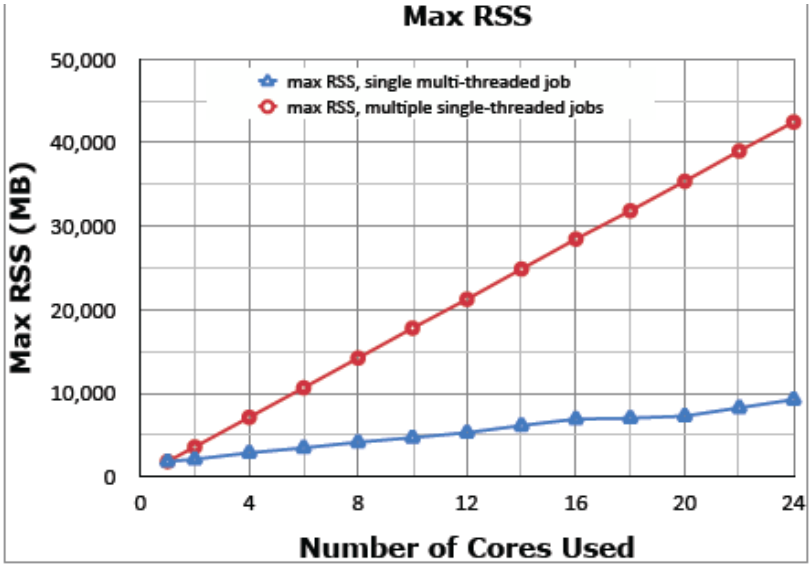
A new data-tier: mini-AOD

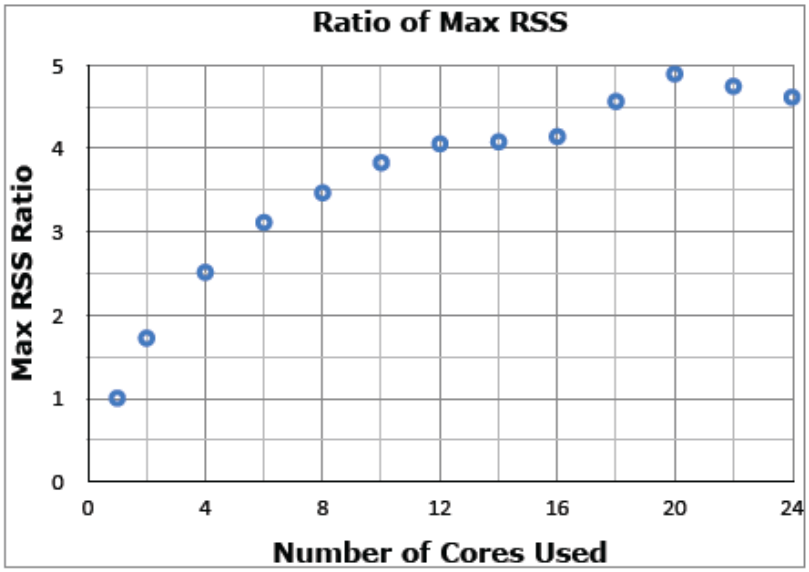
- New thin data tier →10x compression, 30-50kB/event
- Process 2B events w/ 100 slots in 24h
- 800M events produced in CSA14
- Update centrally ~ monthly



Multicore

- Full thread-safety achieved in CMS software version 7
- Different levels of concurrency: Module and sub-module level
- TBB: threaded building blocks
- Multi-core CMSSW available in nightly builds since July
- Reconstruction using threads tested at Tier0
- Work ongoing on simulation
- Performance bottleneck
  - Modules that must be run sequentially
  - Lumi block and run boundaries
- Substantial gains in memory consumption and network load

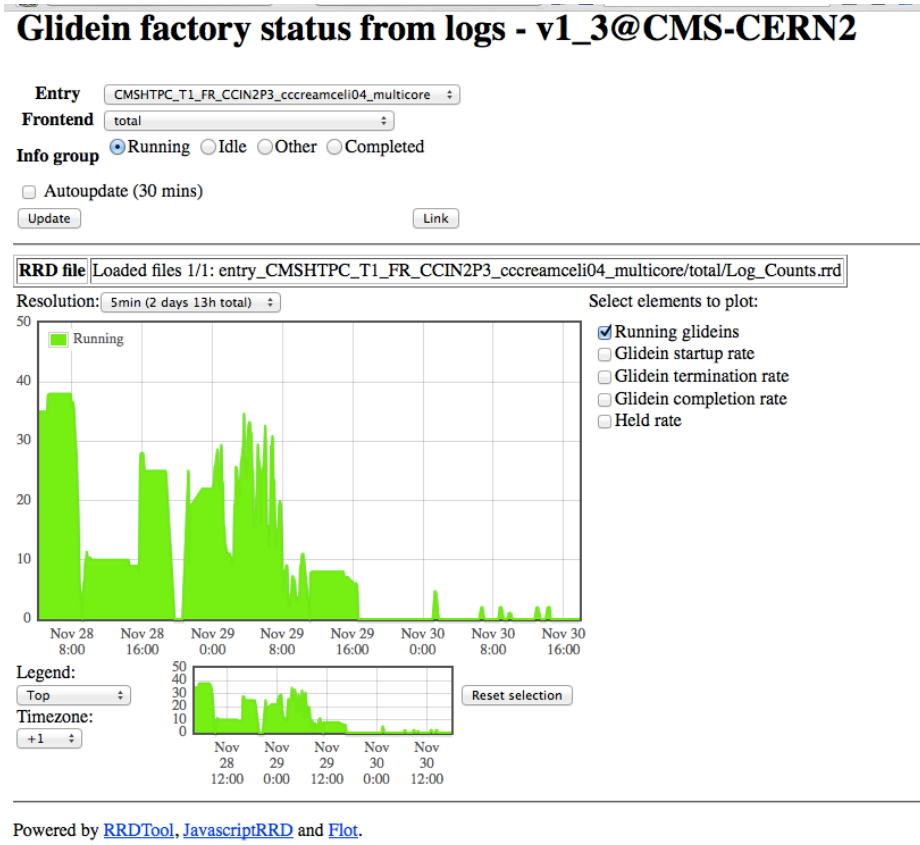




ACAT2014 talk, Sexton-Kennedy

Multicore deployment

- Scaling to large-scale multi-core production not trivial
- Production tool (WMAgent) still being tested
- Many challenges: Pilot optimization, job monitoring, etc.

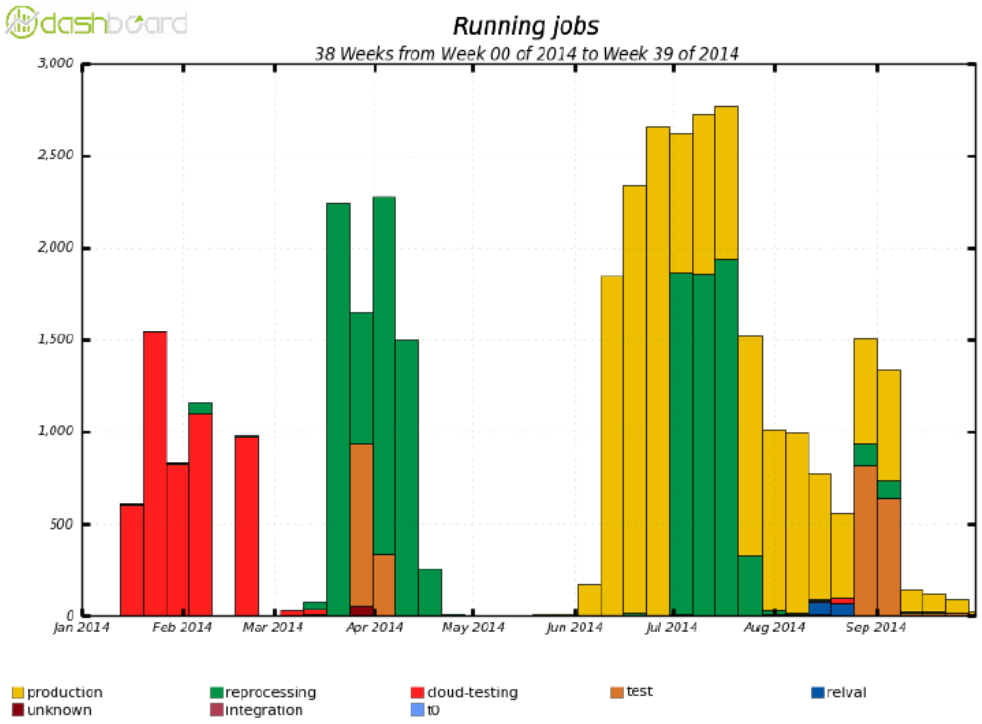


Muti-core pilots at the CCIN2P3 T1

Follow the projection here:
CMSMulticoreSchedulingProject

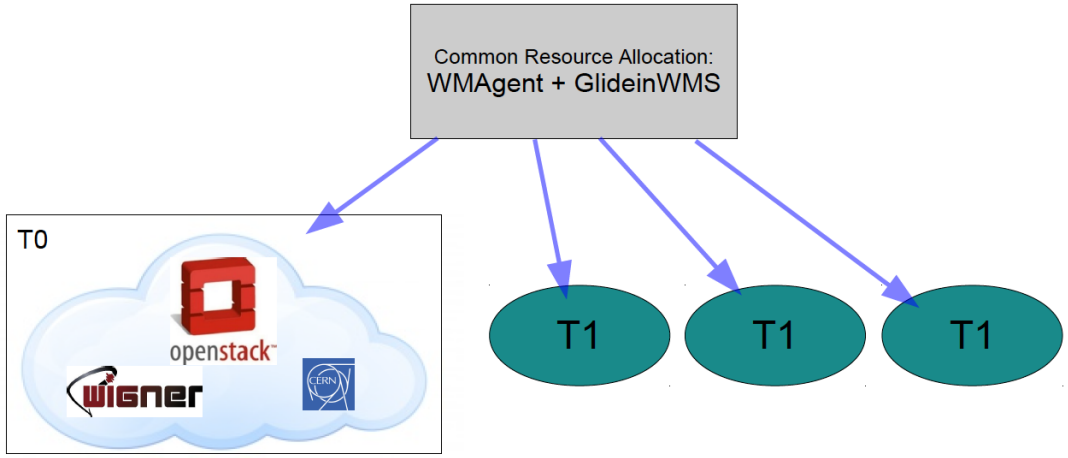
I believe testing is also underway at GRIF

HLT as a cloud resource



- With 10k cores, the HLT farm represents a massive resource
- Implemented cloud middleware (OpenStack) to use HLT during downtime
- Heavy ion reprocessing campaign in April 2014 successful test case
- Steady use for production afterwards
- Use of HLT during inter-fill periods under investigation

Prompt Reco on the cloud



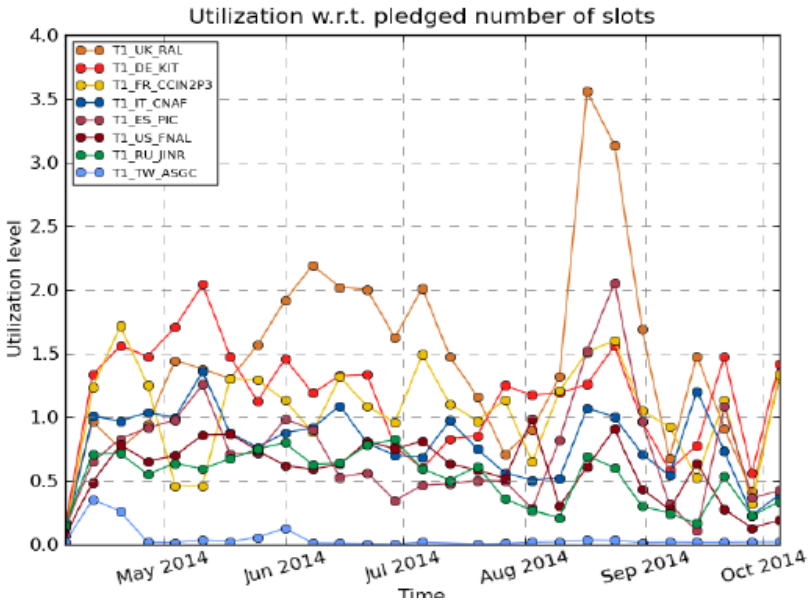
- Prompt reco requires nearly all of CPU resources and a good fraction of T1 CPU
- Move to a Openstack based Cloud-like virtualized resources located at CERN (1/3 CPUs, 2/3 disk storage) and Wigner (2/3 CPUs, 1/3 disk storage)
- Based on the Agile Infrastructure
- New method of resource allocation: GlideinWMS
- Work underway to commission this system

Towards physics

- 1B events simulated by June 1st
- Digitized with both 25ns and 50ns bunch spacing
- Reprocessing during technical stops, e.g., to update
  - Machine parameters
  - Alignment calibrations
  - Reconstruction developments
- 4B MC events by end of 2015

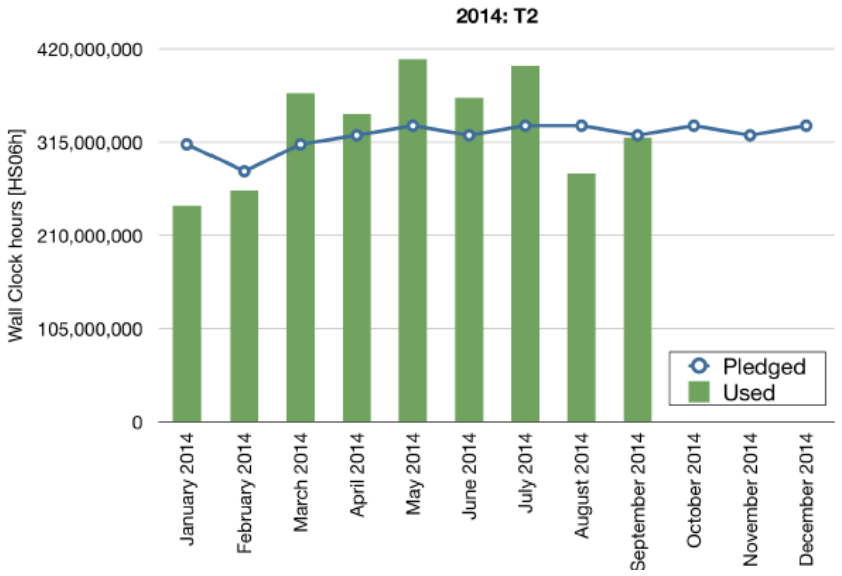
Resource Utilization

Tier 1



- Utilization beyond WLCG pledges
- Main usage: Mostly reprocessing
- Now open for user analysis jobs

Tier 2



- Mostly also beyond WLCG pledges
- Main usage: Analysis and MC production

Resource demands

Country	Federation	Pledge Type	CMS	% of Req.
France	FR-CCIN2P3	CPU (HEP-SP	22600	8%
France	FR-CCIN2P3	Disk (Tbytes)	1960	8%
France	FR-CCIN2P3	Tape (Tbytes	5580	8%
Germany	DE-KIT	CPU (HEP-SP	26850	9%
Germany	DE-KIT	Disk (Tbytes)	2600	10%
Germany	DE-KIT	Tape (Tbytes	7400	10%
Italy	IT-INFN-CNA	CPU (HEP-SP	39000	13%
Italy	IT-INFN-CNA	Disk (Tbytes)	3380	13%
Italy	IT-INFN-CNA	Tape (Tbytes	9620	13%
Russian Fede	RU-JINR-T1	CPU (HEP-SP	28800	10%
Russian Fede	RU-JINR-T1	Disk (Tbytes)	2400	9%
Russian Fede	RU-JINR-T1	Tape (Tbytes	5000	7%
Spain	ES-PIC	CPU (HEP-SP	15300	5%
Spain	ES-PIC	Disk (Tbytes)	1326	5%
Spain	ES-PIC	Tape (Tbytes	3774	5%
Taiwan	TW-ASGC	CPU (HEP-SPEC06)		
Taiwan	TW-ASGC	Disk (Tbytes)		
Taiwan	TW-ASGC	Tape (Tbytes)		
UK	UK-T1-RAL	CPU (HEP-SP	24000	8%
UK	UK-T1-RAL	Disk (Tbytes)	2080	8%
UK	UK-T1-RAL	Tape (Tbytes	5920	8%
USA	US-FNAL-CM	CPU (HEP-SP	120000	40%
USA	US-FNAL-CM	Disk (Tbytes)	10400	40%
USA	US-FNAL-CM	Tape (Tbytes	29600	40%

2015 T1 pledge breakdown

	2014	Increase from 2013	2015	Increase from 2014	2016	Increase from 2015	2017	Increase from 2016
Tier-0 CPU (kHS06)	121	0%	256 (256)	111%	302	18%	350	18%
Tier-0 Disk (TB)	7000	0%	3250 (3000)	Reallocated to CAF	3250	0%	3250	0%
Tier-0 Tape (TB)	26000	0%	31000 (31000)	31%	38000	23%	50000	31%
CAF CPU (kHS06)	0	0%	15 (15)	-	15	17%	17	21%
CAF Disk (TB)	0	0%	12100 (12000)	-	13100	8%	14000	7%
CAF Tape (TB)	0	0%	4000 (4000)	-	6000	50%	8000	33%
T1 CPU (kHS06)	175	0%	300 (300)	71%	400	33%	525	31%
T1 Disk (TB)	26000	0%	27000 (26000)	4%	35000	30%	45000	28%
T1 Tape (TB)	55000	11%	73500 (74000)	34%	100000	36%	135000	35%
T2 CPU (kHS06)	390	14%	500 (500)	25%	700	40%	800	14%
T2 Disk (TB)	27000	4%	31400 (29000)	16%	40000	27%	48000	20%

Table 8: Processing, disk, and tape resources requested by CMS for all centrally controlled computing tiers. The column named "2015" shows within parentheses the resources as scrutinized by C-RSG in April 2014.

Evolution of resources demands

Conclusions

- Run 2 is around the corner → Presents new challenges for computing
- Lots of effort went into preparing for this challenge during LS1
  - Data federation
  - Multicore processing
  - New grid submission tools
  - Use of cloud resources
- Large scale production of simulation is underway
- Plenty of work left to in terms of commissioning