

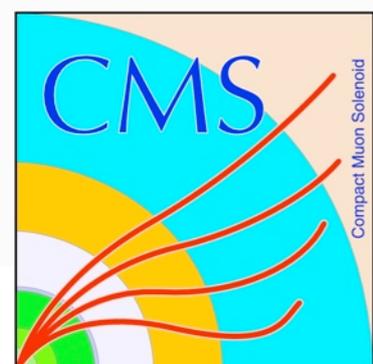
Computing Operations activities

Jacob Linacre (FNAL)
for CompOps

CMS Week

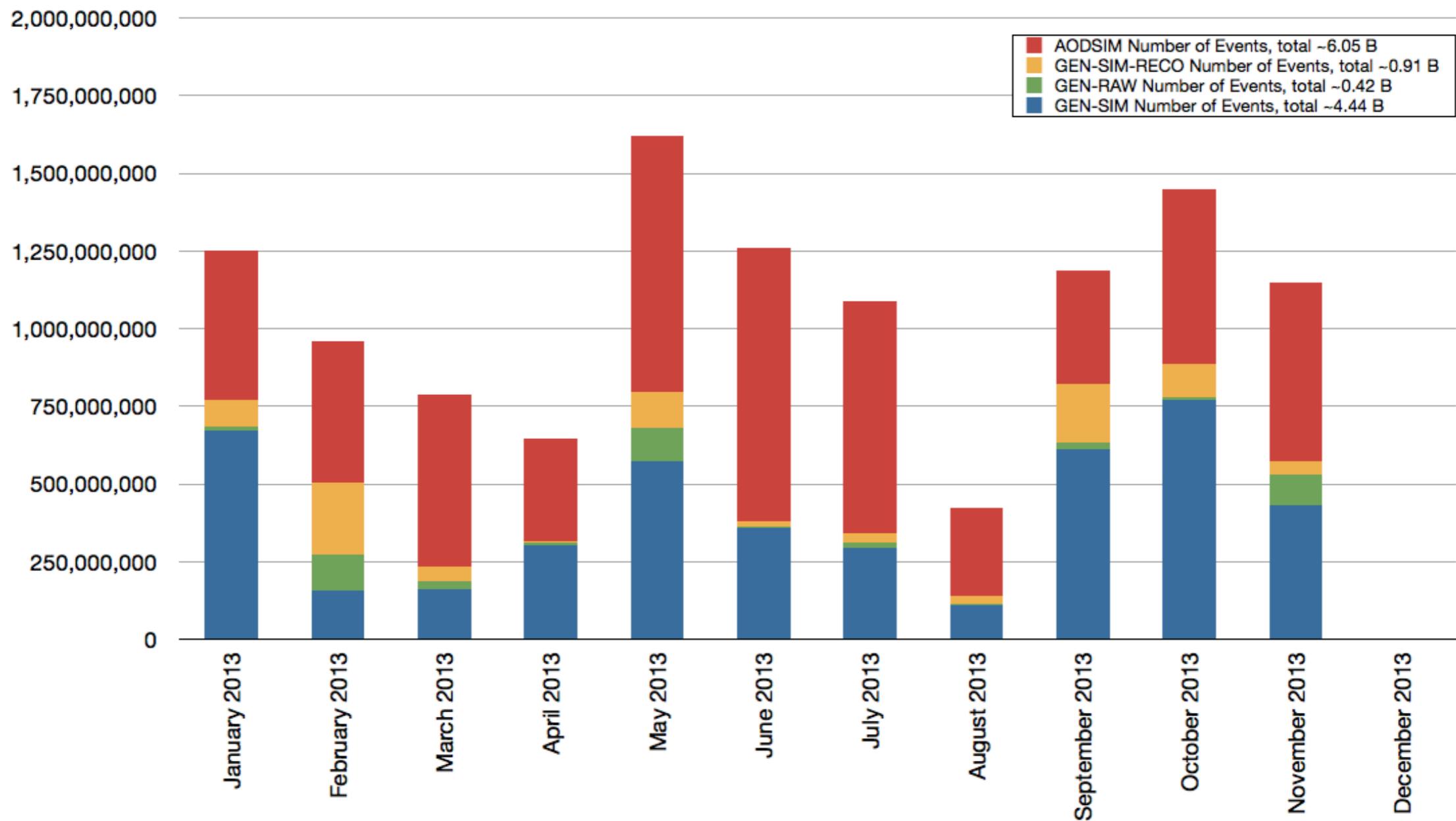
Joint Physics/Offline/PPD/Computing/Trigger Plenary

11th December 2013



- ▶ We processed ~1B MC events/month in 2013
- ▶ each event is counted twice here, once for **simulation** and once for **DR**

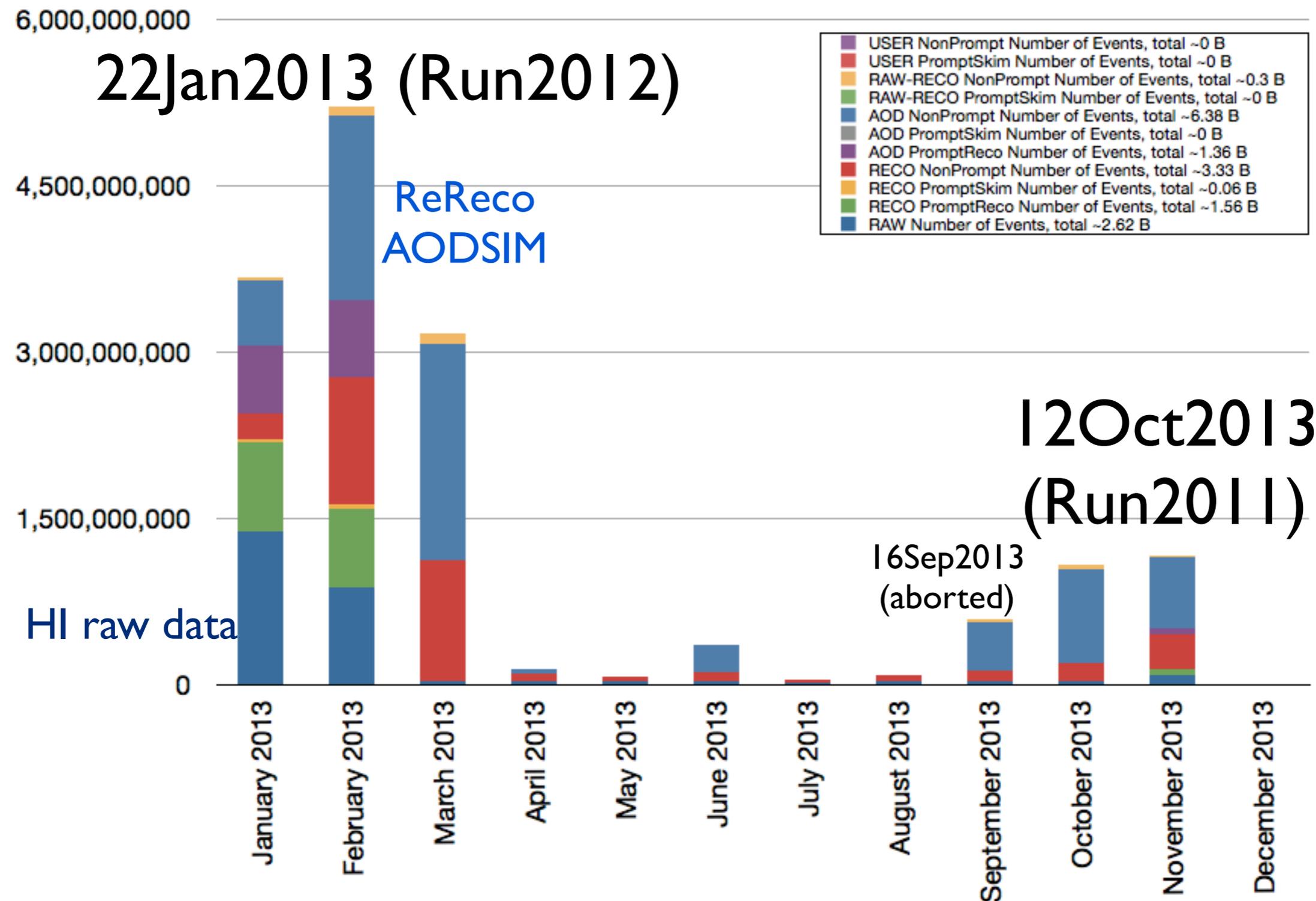
MC in 2013: Number of Events per Month



- ▶ Can sustain > 1B events/month (dips were when all work was finished)

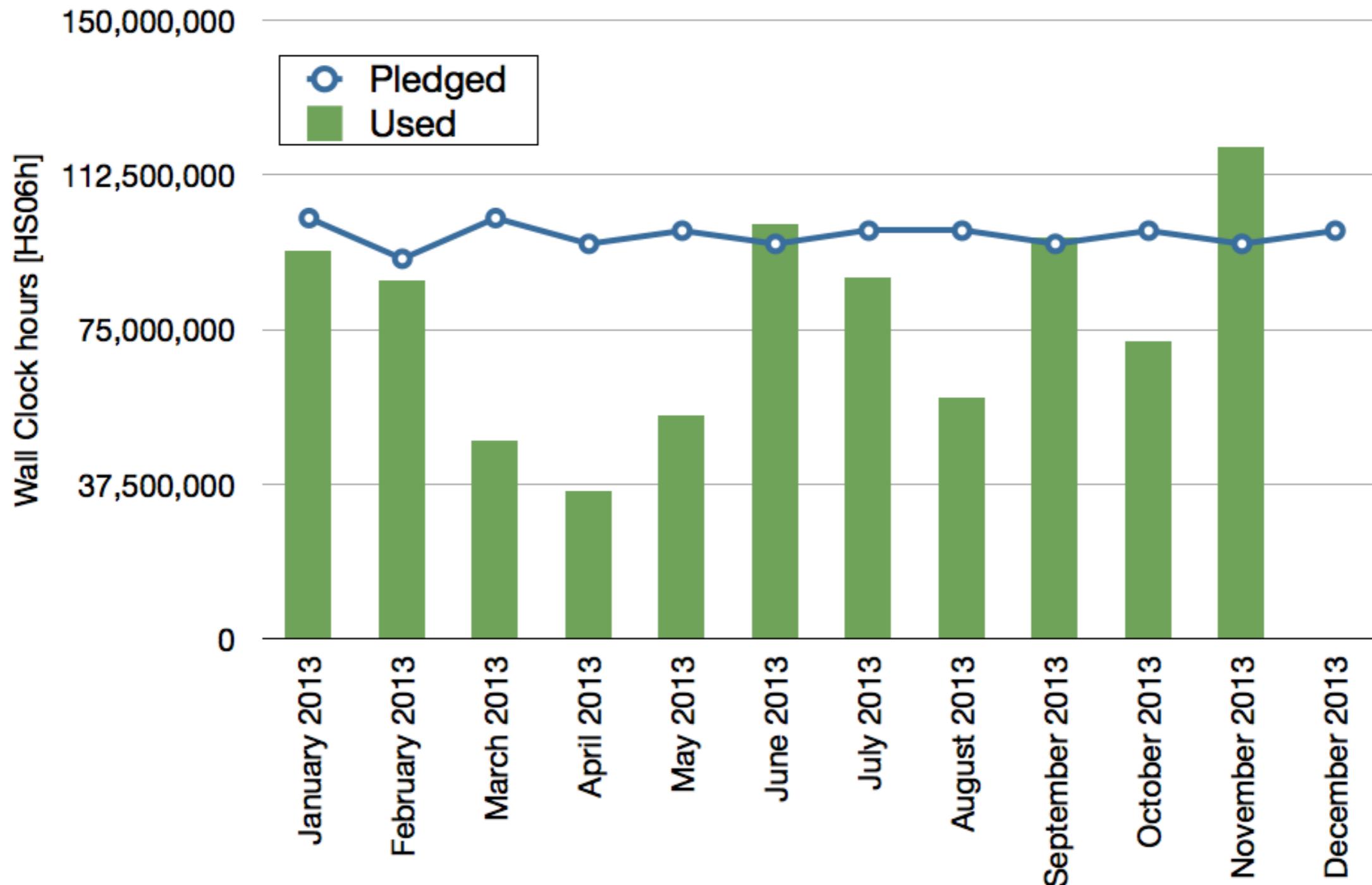
▶ Two large ReReco campaigns in 2013

Data in 2013: Number of Events per Month

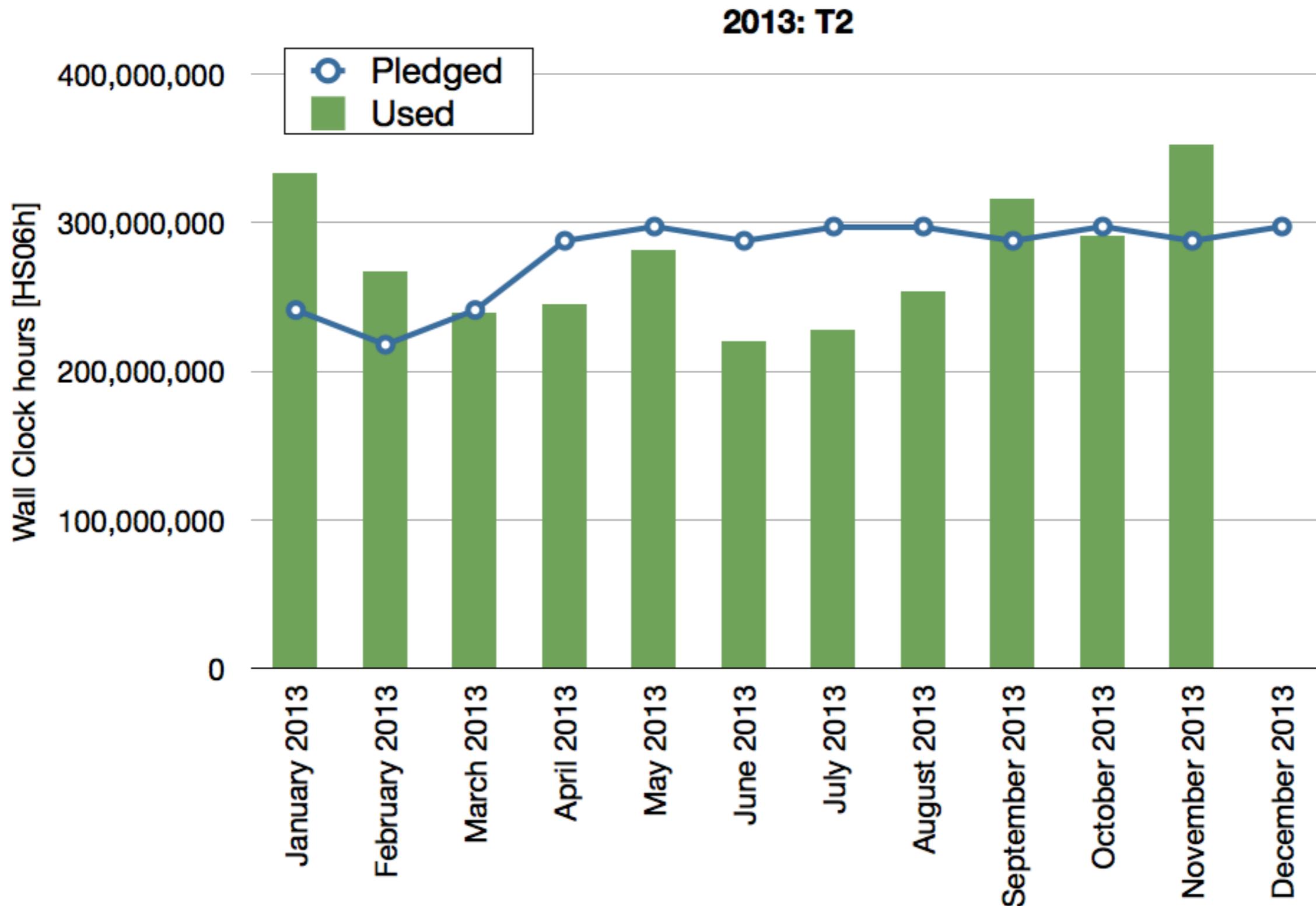


- ▶ Mostly using all our T1 pledged resources in 2013
- ▶ under-utilised in spring when there was little work requested

2013: T1



► Consistent usage of pledged T2 resources throughout the year



► Event counts for currently active MC campaigns

	PRODUCTION	VALID
FSIM	79.7M	3361.9M
HiWinter13	0.0M	458.8M
LowPU2010	0.0M	405.7M
Summer11 GEN	236.0M	5778.2M
Summer11 GEN-SIM	8.4M	218.4M
Summer12 GEN	651.3M	26015.0M
Summer12 GEN-SIM	226.4M	7152.3M
Upgrade(S13)	1.1M	246.4M

=> Fall 11 AODSIM

=> Summer12_DR53X
AODSIM

- ▶ 2011 Legacy ReReco (12Oct2013) completed this week, including missing lumi recovery
 - ▶ missing lumis in the initial processing are an unavoidable consequence of temporary site problems and problematic lumi sections
 - ▶ this is the first time we have guaranteed 100% of processable lumis
- ▶ As expected, the missing lumi recovery took longer (4 weeks) than the bulk of the processing (3 weeks for ~99%)
 - ▶ feature is new, still quite labour intensive, and does not yet support much parallelisation
- ▶ Unfortunately many (18) workflows that ran at T2_CH_CERN have a small fraction (mostly $<0.1\%$) of missing events despite no missing lumis (/MinimumBias/Run2011A has $\sim 0.15\%$, and /Jet/Run2011A $\sim 0.5\%$)
- ▶ 2011 Legacy ReReco was previously run as 16Sep2013
 - ▶ would have been complete much sooner, but had to be restarted due to a new bug causing missing parentage information (fixed for 12Oct)

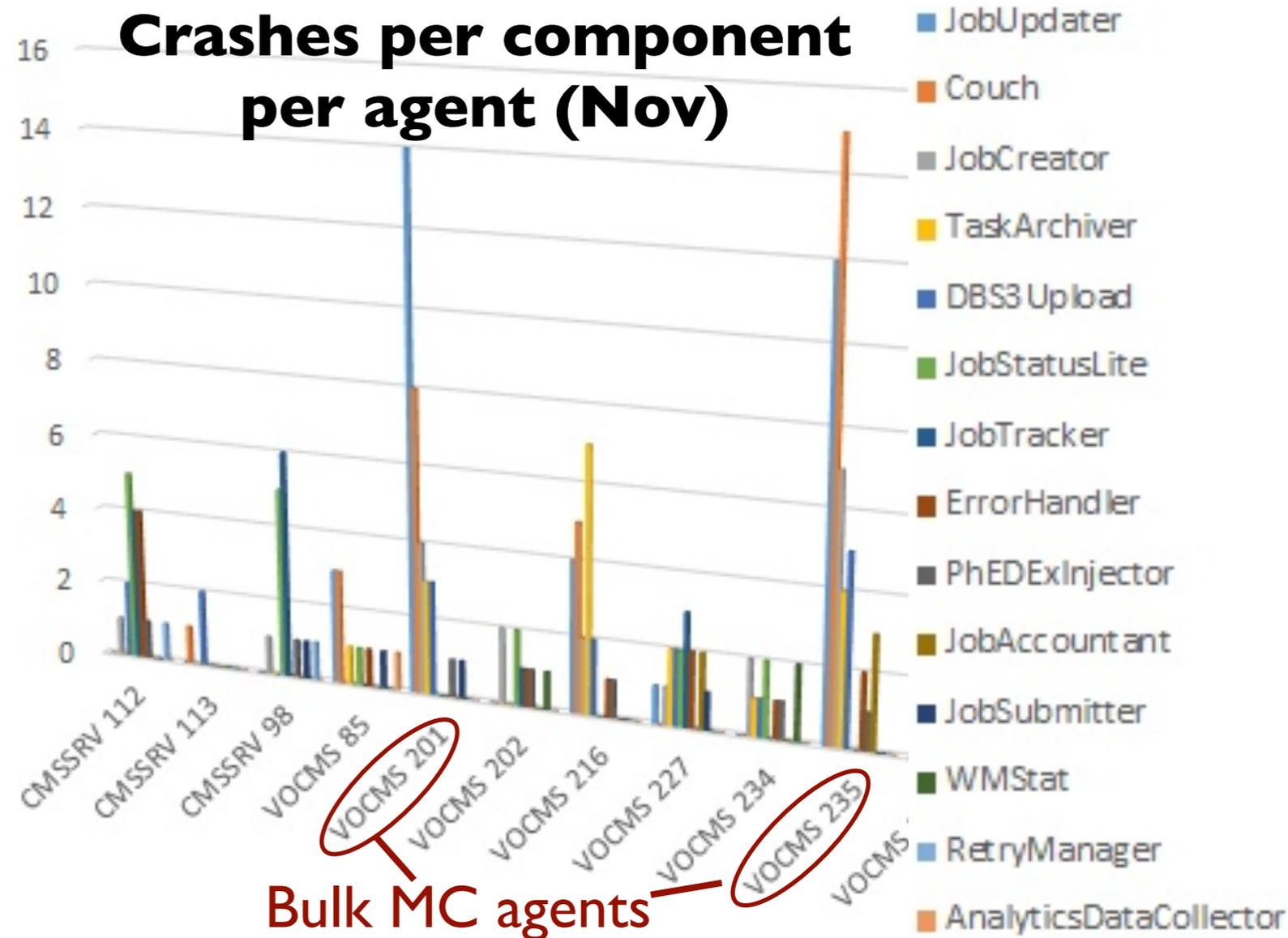
- ▶ Missing events in the 2011 Legacy ReReco led us to identify a feature of CMSSW that can lead to invisible loss of events
- ▶ Signals SIGUSR2 and SIGINT cause CMSSW to exit gracefully after processing the current event, with exit code 0
 - ▶ all remaining events in the lumisection are invisibly lost (since the job is marked as successful, and accounting is done at the lumi level)
- ▶ CMSSW_5_3_X patched to fix this (exit code 9000)
 - ▶ <https://github.com/cms-sw/cmssw/pull/1542>
- ▶ We only observed this at T2_CH_CERN
 - ▶ legacy feature of batch system, e.g. SIGUSR2 for RUNtime limit reached
 - ▶ <http://information-technology.web.cern.ch/services/fe/lxbatch/howto/how-interpet-batch-job-return-codes>
 - ▶ occurs rarely, and hadn't noticed it before because this was the first time we ran a significant amount of ReReco at T2_CH_CERN

- ▶ Some requests have been affected by long latencies in the WMAgent system
- ▶ Data and MC processing during LHC Run I identified several development issues
 - ▶ Some already resolved
 - ▶ smart lumi splitting (more uniform job length reduces tails in ReReco)
 - ▶ WMAgent kills for timeout and memory (prevents repeated condor resubmission)
 - ▶ split workflows across agents (system more scalable for large workflows)
 - ▶ missing lumi recovery (resolves problem of lost unmerged files at a site)
 - ▶ Work ongoing on other issues
 - ▶ stuck workflows
 - ▶ WMAgent components going down
 - ▶ not all problems are automatically flagged
 - ▶ in general, lots of hand-holding to keep the system running (lack of automation)
- ▶ Lots to do, but all the original WMAgent lead developers have moved on
 - ▶ slowly ramping up expertise
 - ▶ lots of time spent resolving immediate problems with workflows rather than long-term improvements (can't just halt production)

- ▶ Stuck workflows have been a big problem over the last months
- ▶ This is a symptom with many causes!
 - ▶ Slow DB query in task archiver that takes too long to complete (several hours) can prevent workflows from moving to "running" to "complete"
 - ▶ Bug that prevents correct Phedex subscription prevents workflows that are "complete" from moving to "closed-out"
 - ▶ Suboptimal prioritisation of merge, logCollect and clean-up jobs means they can be left waiting when higher priority work is injected
 - ▶ Site issues, e.g. can't merge up files at a site that is down!
 - ▶ General agent instability under high load (more later)
 - ▶ Many more!

- ▶ Not all these problems are automatically flagged
- ▶ All require manual intervention to be fixed
 - ▶ several scripts have been developed that usually get the workflow moving again
 - ▶ sometimes it's more complicated and support from the developers is necessary
- ▶ It's often difficult to identify exactly which problem is affecting a given workflow, particularly when several problems at once or compounded by technical problems with the request
 - ▶ issues with filter efficiencies, job splitting, input dataset, and config problems
- ▶ At the start of each week we often have ~50 stuck workflows
 - ▶ usually down to ~0 by the end of the week
 - ▶ but this takes a lot of manual work

- ▶ Many WMAgent components go down (crash) on a daily basis
 - ▶ plot shows the occurrences per agent over a 1 month period
 - ▶ more frequent in the MC agents with much higher load (# jobs)
- ▶ When a component goes down, all workflows running in the agent are affected
- ▶ Manual intervention always required to restart components
 - ▶ sometime developer assistance also necessary
- ▶ Lots of hand-holding to keep everything running

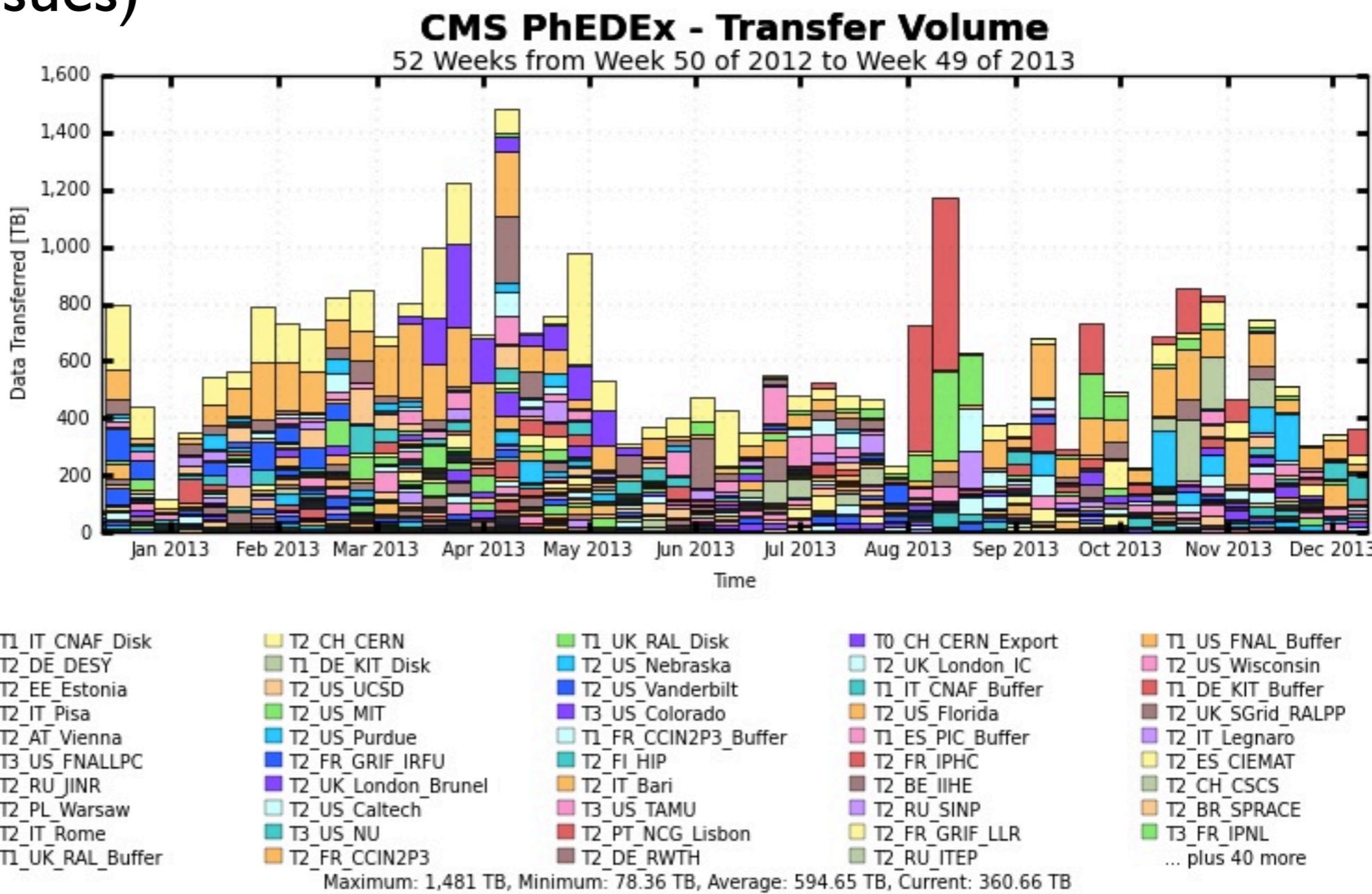


- ▶ Identify stuck workflows as soon as possible
 - ▶ too many workflows to rely solely on human observation
 - ▶ until recently we have been using external scripts
 - ▶ not an ideal solution
 - ▶ “request alert” system is built in to WWMStats
 - ▶ still under development, but now in a useful state. Workflow team iterating with developers to improve it
- ▶ Fix the underlying issues with the code
 - ▶ prioritise the known issues that cause components/workflows to get stuck
- ▶ Job prioritisation
 - ▶ last 5% of a wf should have priority boost, so it completes ahead of new work
 - ▶ merge jobs and cleanup/logCollect jobs should have top priority
- ▶ Smart error handler (decision on how often to retry failed jobs)
 - ▶ segfault, timeout, stage-out error etc all have different requirements
- ▶ site readiness (“waiting room” for sites with problems)
 - ▶ agent flexibility to resubmit work run at a site that has gone down

- ▶ Run low priority testing workflows filling up unused production resources to harden system and identify and remove remaining problems
- ▶ Permanent high-scale running means we don't run in to problems just at the point that we need 100% resources

- ▶ Transfers performing very well, up to 1.4 PB/week during evacuation of ASGC
- ▶ A lot of work goes into the last percent of transfers (many site infrastructure issues)

▶ Establishing PerfSonar to tackle the low level network monitoring (network fabric monitoring)





Status	Site Name	Status	Site Name
✓	T0_CH_CERN	✓	T2_AT_Vienna
✓	T1_CH_CERN	✓	T2_BE_IHE
⚠	T1_DE_KIT	✓	T2_BE_UCL
✓	T1_DE_KIT_Disk	✓	T2_BR_SPRACE
✓	T1_ES_PIC	✓	T2_CH_CERN
⚠	T1_FR_CCIN2P3	✓	T2_CH_CERN_AI
⊛	T1_FR_CCIN2P3_Disk	✓	T2_CH_CERN_HLT
✓	T1_IT_CNAF	✓	T2_CH_CSCS
✓	T1_RU_JINR	✓	T2_CN_Beijing
✓	T1_RU_JINR_Disk	✓	T2_DE_DESY
✓	T1_TW_ASGC	✓	T2_DE_RWTH
✓	T1_UK_RAL	✓	T2_EE_Estonia

- ▶ In 2013 we ramped up the site readiness effort
 - ▶ going into Run 2 ~100% readiness very important
- ▶ More automated systems to optimise resource usage efficiency
- ▶ T2 waiting room for problematic sites

T2 waiting room

Status	Site Name	Status	Site Name	Status	Site Name	Status	Site Name
⊛	T2_BR_UERJ	⊛	T2_PK_NCP	✓	T2_RU_ITEP	✓	T2_TH_CUNSTDA
⊛	T2_IN_TIFR	✓	T2_PL_Warsaw	⚠	T2_RU_PNPI	✓	T2_TR_METU
⚠	T2_MY_UPM_BIRUNI	✓	T2_RU_INR	✓	T2_RU_RRC_KI		

- ▶ Generated ~5B events in both GEN-SIM and AODSIM
- ▶ Good usage of pledged resources
- ▶ Working on improving latencies
- ▶ Improvements in transfers and site readiness

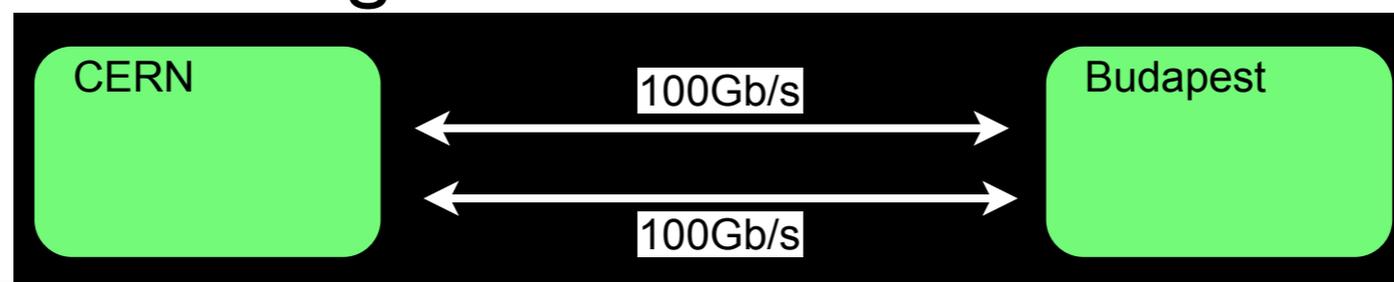
- ▶ Excellent progress in many other areas I didn't have time to mention
 - ▶ glideinWMS including opportunistic resources and clouds
 - ▶ cvmfs
 - ▶ glexec
 - ▶ others ...



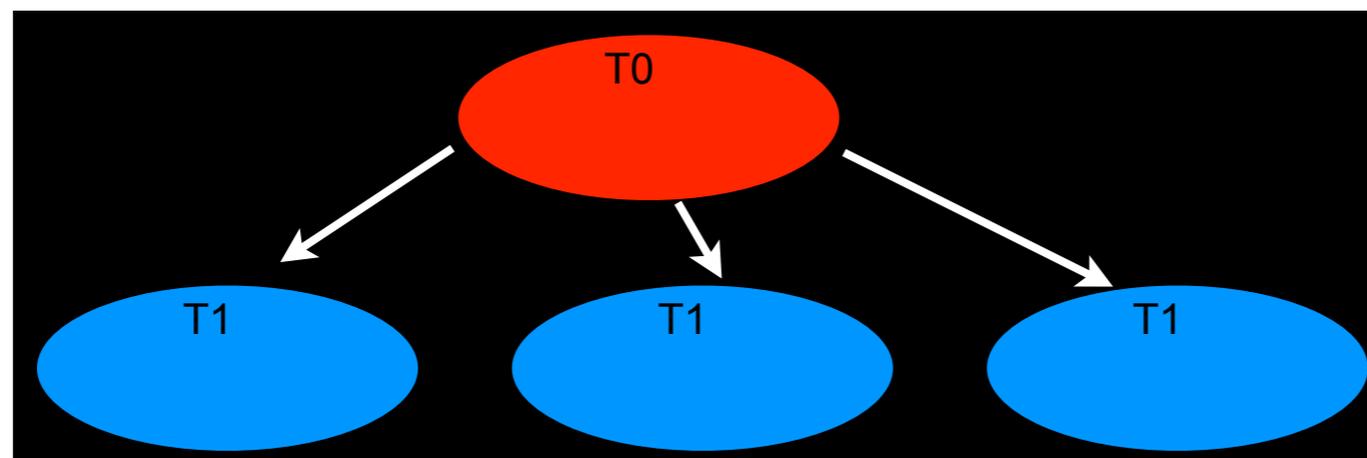
Backup



- ▶ Demands on computing will go up by ~factor of 6, while our capacity will only roughly double
 - ▶ CSA14 discussion yesterday , need big efficiency gains
- ▶ cloud and HLT commissioning work ongoing to increase capacity
 - ▶ recently tested by running the 12Oct2013 Tau datasets on the HLT
- ▶ Opportunistic computing increases capacity for processing tasks
 - ▶ recently used 8000 cores at the San Diego Super Computing Center
- ▶ CERN has deployed a remote computing facility in Budapest
 - ▶ 200Gb/s of network link to CERN at 35ms latency
 - ▶ To users this is indistinguishable from local CERN resources!
 - ▶ More such links are being commissioned



- ▶ Increase efficiency by restructuring the processing infrastructure
 - ▶ decrease distinction between T1 and T2 sites, so we can benefit from the combined total of slots for processing
 - ▶ disk/tape separation at T1s (already at RAL, CNAF and KIT)
 - ▶ read data from remote storage for CPU intensive tasks
 - ▶ only functional difference between T1s and T2s will be that T1s are used for archival (tape storage)
- ▶ use T1 sites for prompt reco (previously restricted to T0)



- ▶ AAA
- ▶ Lots of potential for improvement

- ▶ we need multicore by 2015 especially when trigger rates per PD get so high that we cannot process a single lumi section in 48 hours on one core
- ▶ we can already run in forked mode but this does not help (1 lumi still the smallest unit)
- ▶ the multi-threaded framework will enable us to do this
- ▶ or change the lumi section length

- ▶ finalize switch to EOS and new StorageManager
- ▶ continue implement Prompt Calibration Loop
- ▶ monthly replays and continue to follow global runs
- ▶ we need to test running 50% of prompt reco at T1 sites

▶ DBS3

- ▶ we need to update DBS to a more modern code base and remove functionality which is not needed anymore (data certification) but slows down the system considerably
- ▶ planned switch is Mid January 2014 (DBS2 will be read only)
- ▶ CRAB2 is able to talk to DBS3

- ▶ resource utilization efficiency
 - ▶ dynamic data placement and automatic cache release
 - ▶ proposal is: 60% manager, 40% unmanaged (no physics group space anymore)

- ▶ analysis tools