CMS Resource Utilization and Limitations on the Grid after the First Two Years of LHC Collisions

Giuseppe Bagliesi, Kenneth Bloom*, Daniele Bonacorsi, Chris Brew, Ian Fisk,
Jose Flix, Peter Kreuzer and Andrea Sciaba

**Corresponding Author, University of Nebraska-Lincoln*

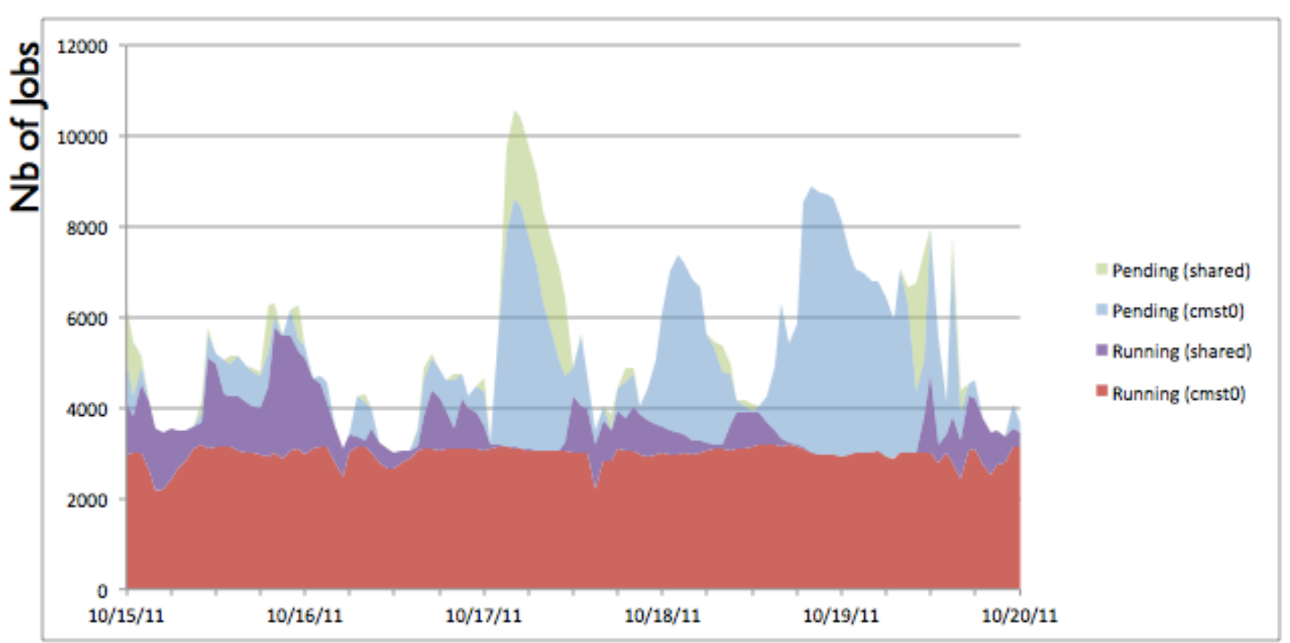
Even after multimillion dollar investments, the LHC experiments ultimately have limited resources for processing and storing data. This has implications for how computing tasks are performed and how physics measurements get done. Computing models must be adjusted and optimized to make the best use of these limited resources. How has the CMS experiment used the available resources in 2011, and what adjustments have been made and need to be made to the computing model?

Drivers: The primary driver of resource usage is the amount of available LHC data. The LHC was expected to run pp collisions for 5.2M seconds and heavy-ion collisions for 0.7M seconds in 2011, and the actual live-time agreed within 5%. The trigger rate fluctuated between 350 and 430 Hz during the pp run, and 1.5 billion events were recorded. During the run the number of "pile-up" interactions increased to 16-17 per beam crossing. Event sizes (in KB below) for different data formats were largely in line with expectations, even as the pile-up increased.

Format	Observed (8 PU events)	Expected (8 PU events)	Observed (30 PU events)	Expected (30 PU events)
Data RAW	230	390	356	800
Data RECO	590	530	1316	900
Data AOD	165	200	327	250
MC RECO	970	600	-	1100
MC AOD	250	265	-	300

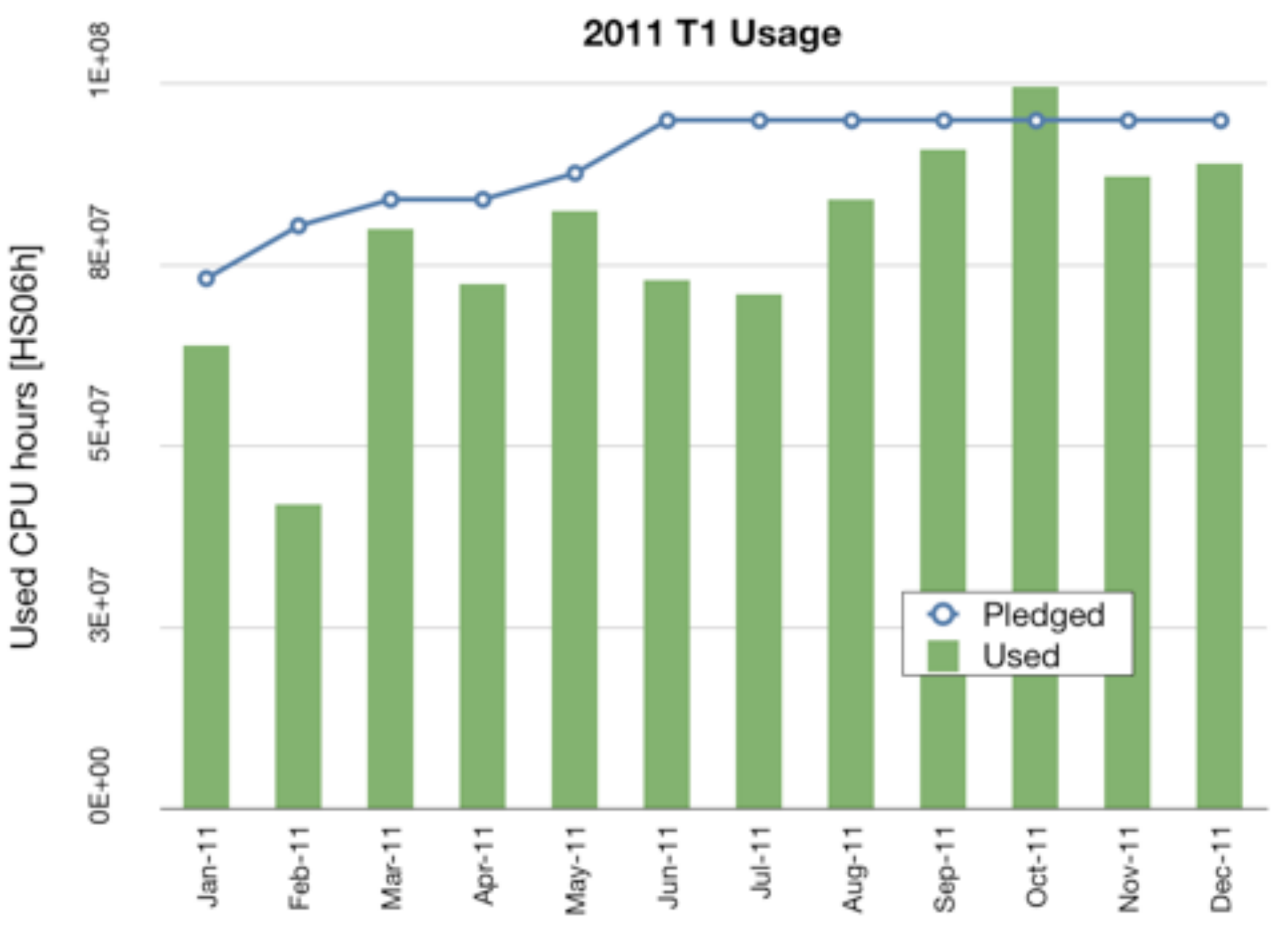
The data are hosted and processed at the facilities of the three-tiered distributed computing system. The use of processing and storage resources can be obtained from the WLCG accounting reports.

Tier 0: Data are first reconstructed at the Tier-0 cluster at CERN, which is meant to handle peak demand. CMS also made use of an "overspill" scheme into shared CERN CPU resources. Even so, there was still a large number of pending jobs at some times.



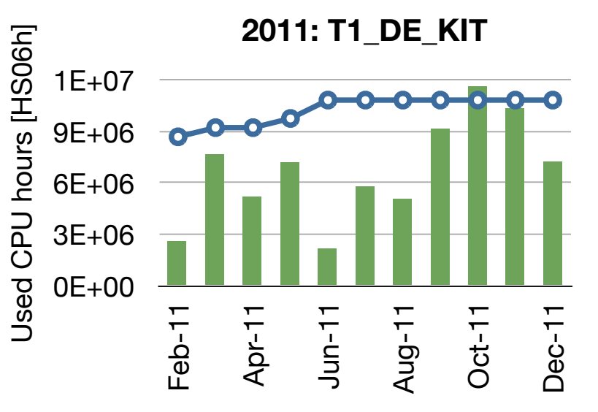
Even when all job slots were full, CPU utilization was only 70%. The memory footprint of the reconstruction executable was larger than expected and not all cores in each compute node could be used. CMS has since made improvements in memory consumption.

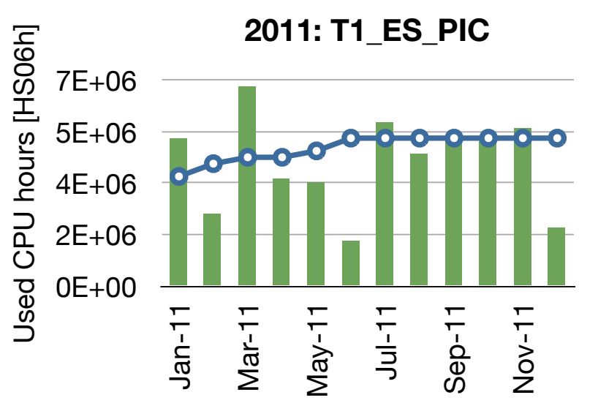
Tier 1: The seven Tier-1 sites are used for archiving data and simulation on tape, re-processing and skimming data, and simulation production. Averaged over 2011, CMS used 87% of pledged Tier-1 processing resources. In 2010, the usage never exceeded 60%. The resource usage was increased in 2011 by moving some simulation production from Tier 2 to Tier 1.

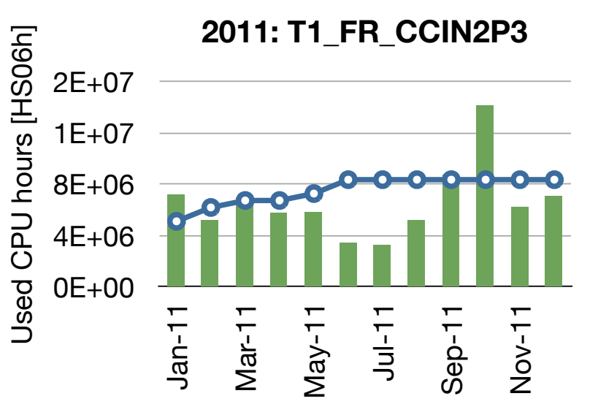


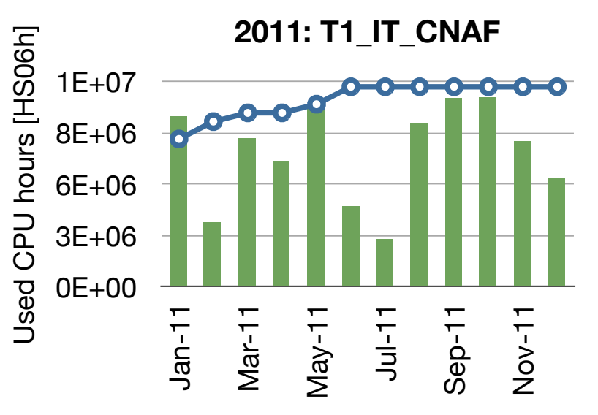
However, not all sites were used equally. The most active site provided 113% of pledged resources, while the least active site provided only 34%. CMS hopes to improve site performance this year.

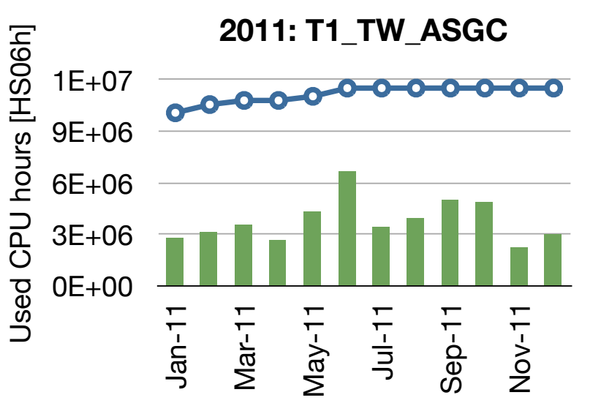
Related CMS posters:
- ★ Towards higher reliability of CMS Computing Facilities (#260)
- ★ Trying to Predict the Future -- Resource Planning and Allocation in CMS (#244)

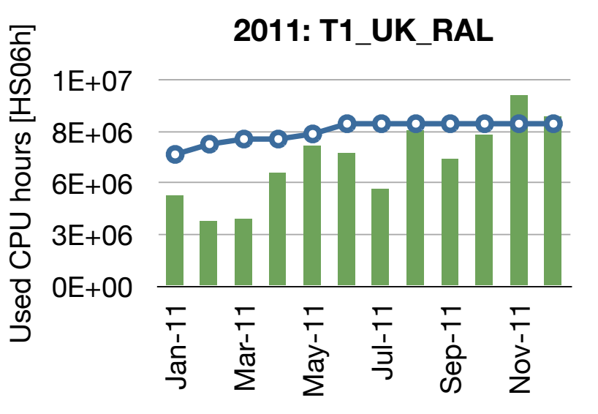


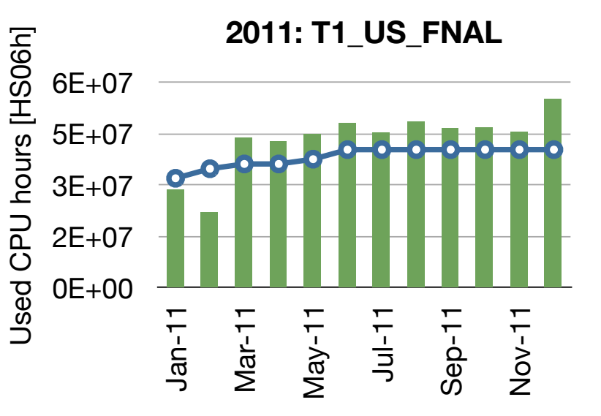




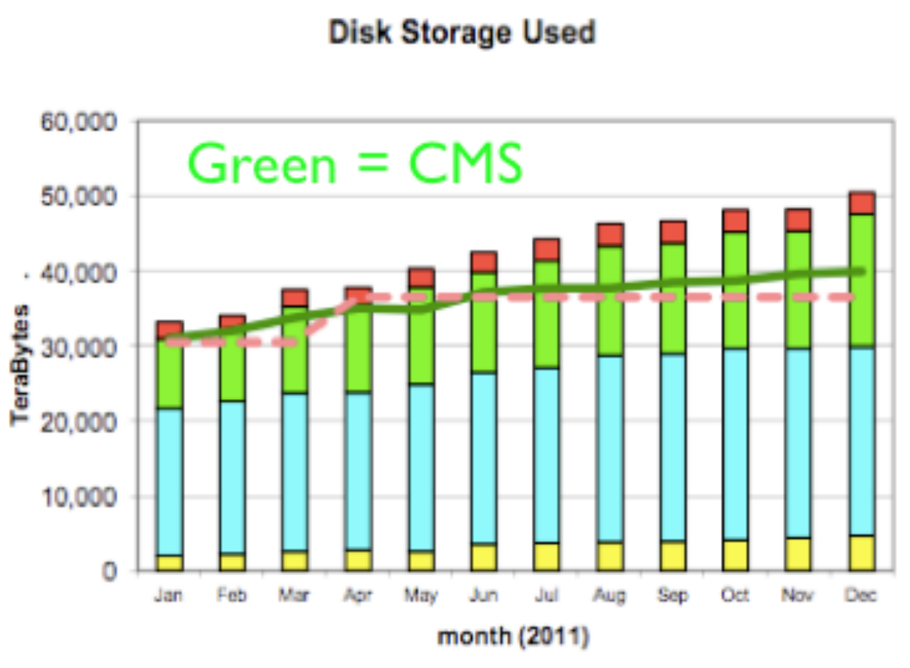


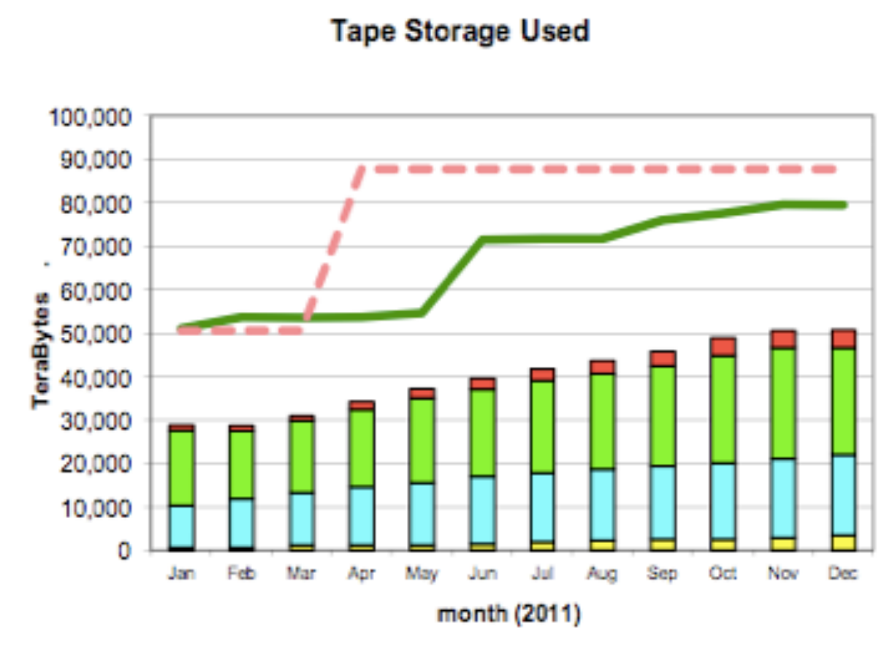




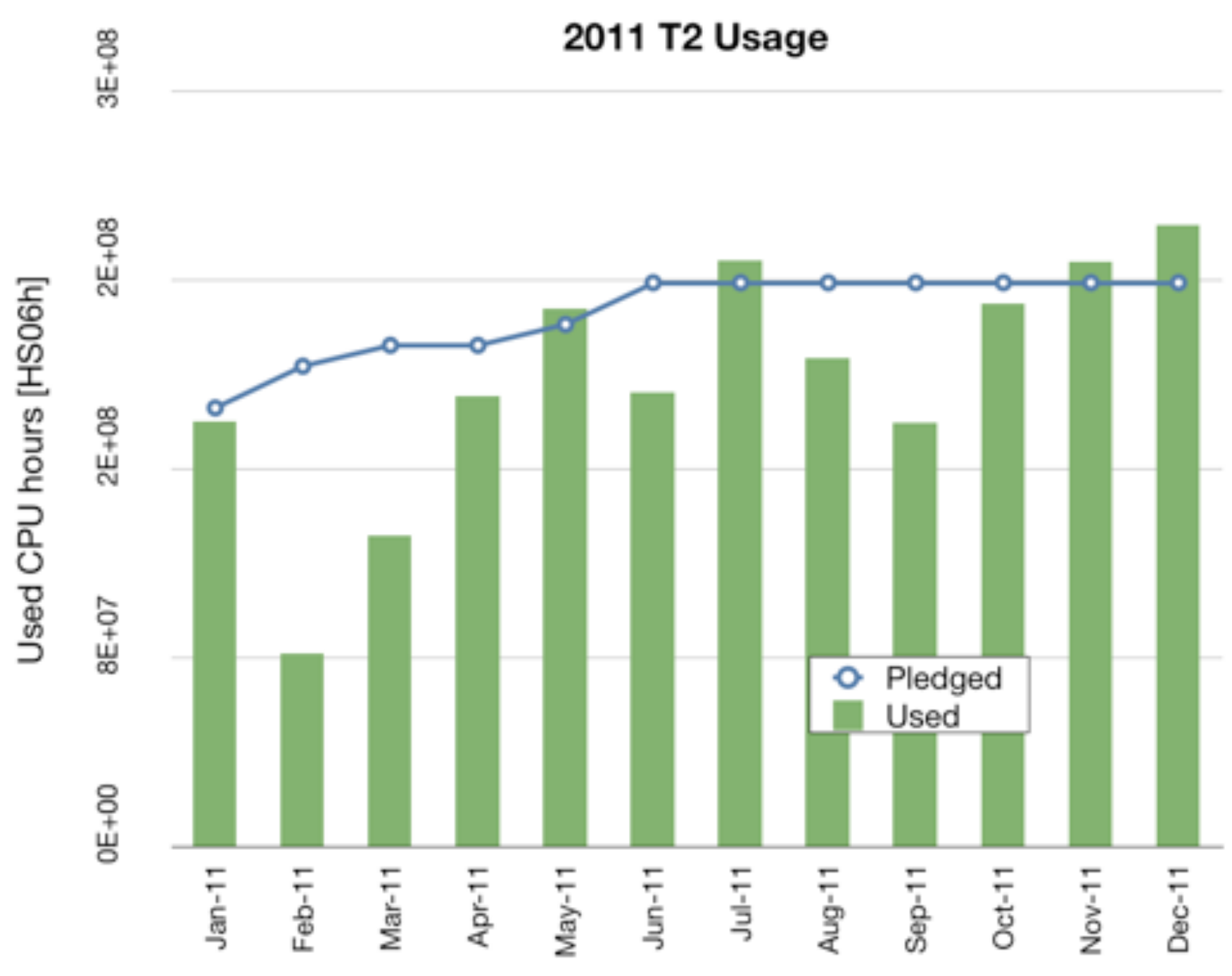


Tier-1 disk and tape use was within expectations. At the end of 2011, CMS was using 24.6 PB of tape, with 45 PB available, and 17 PB of disk, slightly more than the pledged amount.

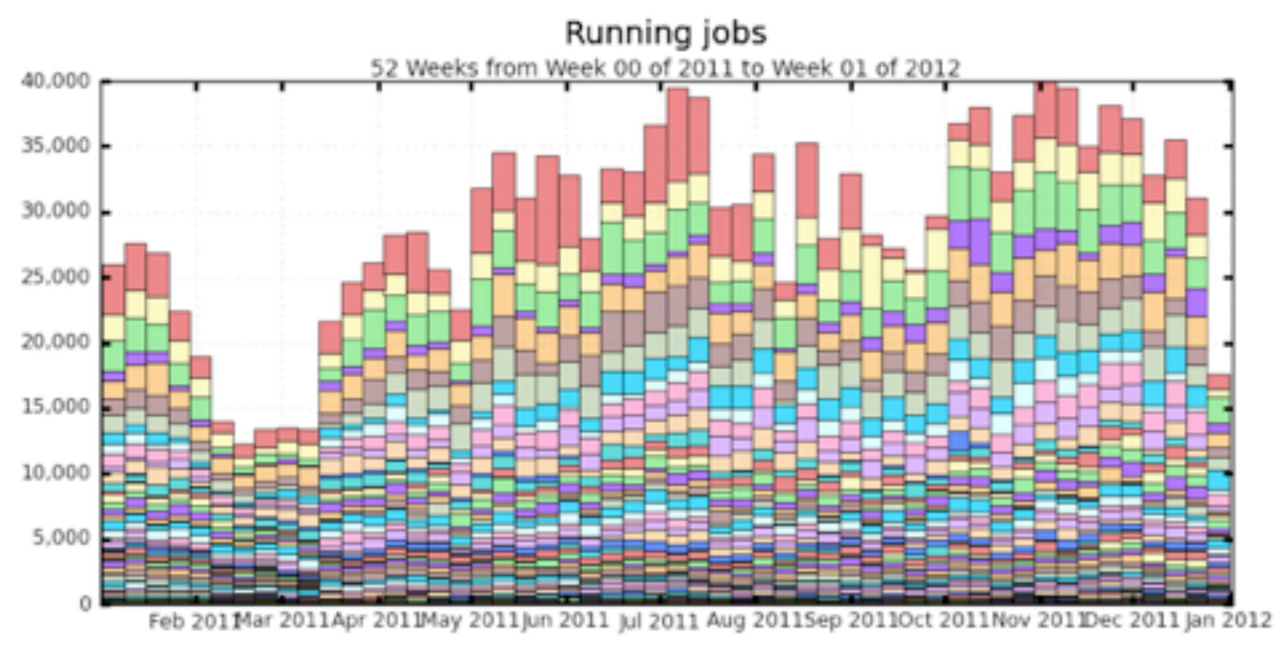


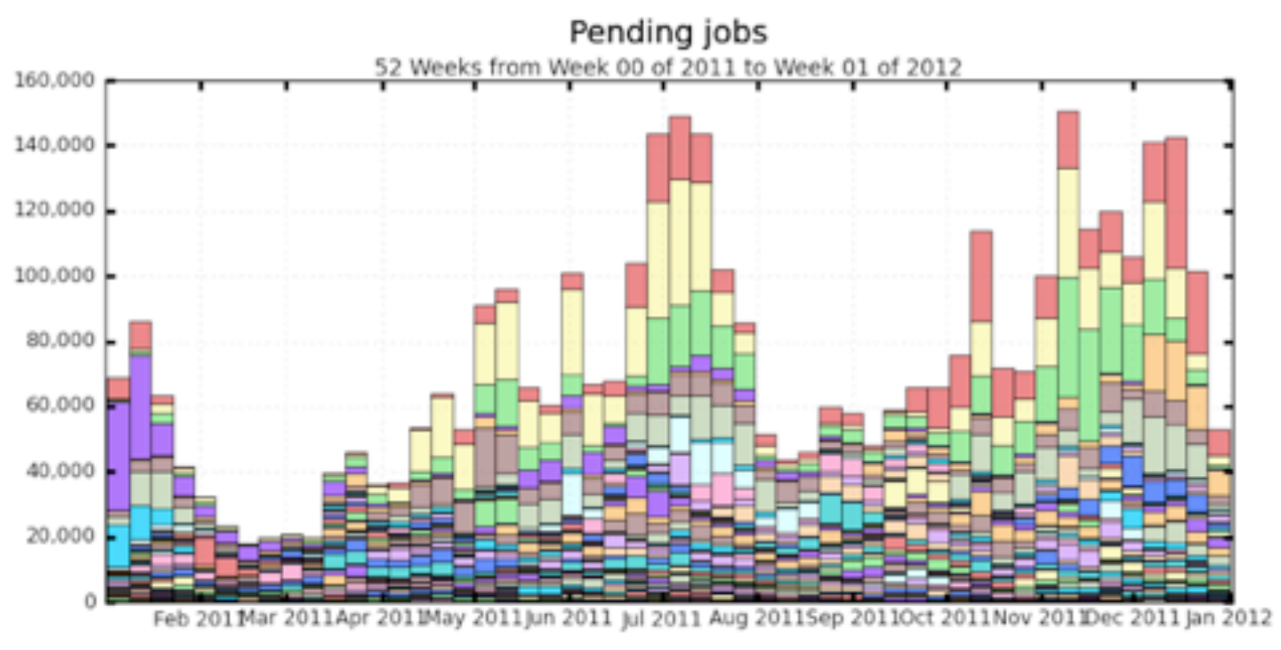


Tier 2: The ~50 CMS Tier-2 sites are used for both centrally-controlled simulation production and user-controlled physics analysis. Disk storage is mostly devoted to analysis samples, with some space reserved for user files and production scratch space. Average CPU usage during 2011 was 88% of the pledged amount. Most of the deficit was incurred early in the year; when the full LHC dataset was available, usage rates were close to or exceeding the pledge. Because of the shift of some simulation production to Tier 1, the CPU usage at the Tier-2 sites tends to follow the patterns of user analysis.



The number of running and pending jobs at the Tier-2 sites tracks well with the CPU consumption over time; when the CPU consumption was close to the pledge, the number of pending jobs grew.

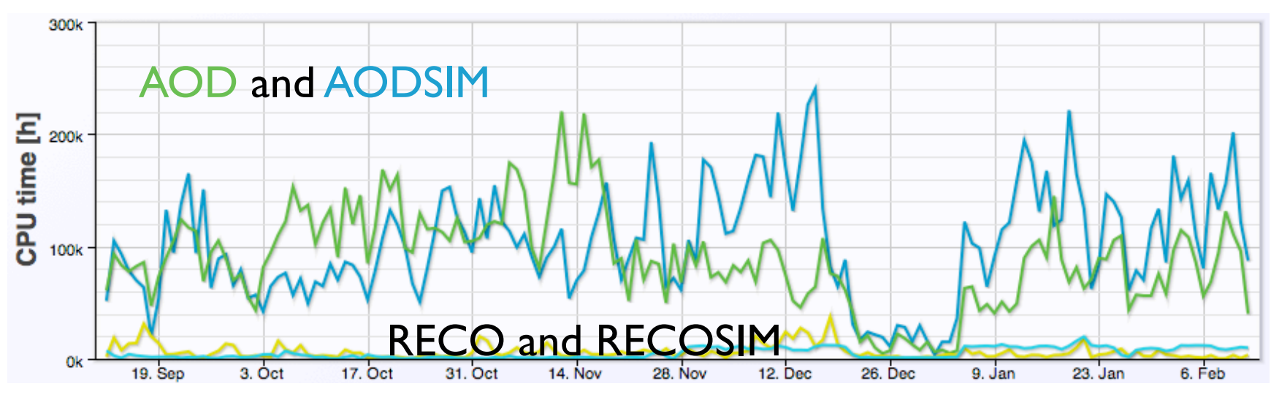




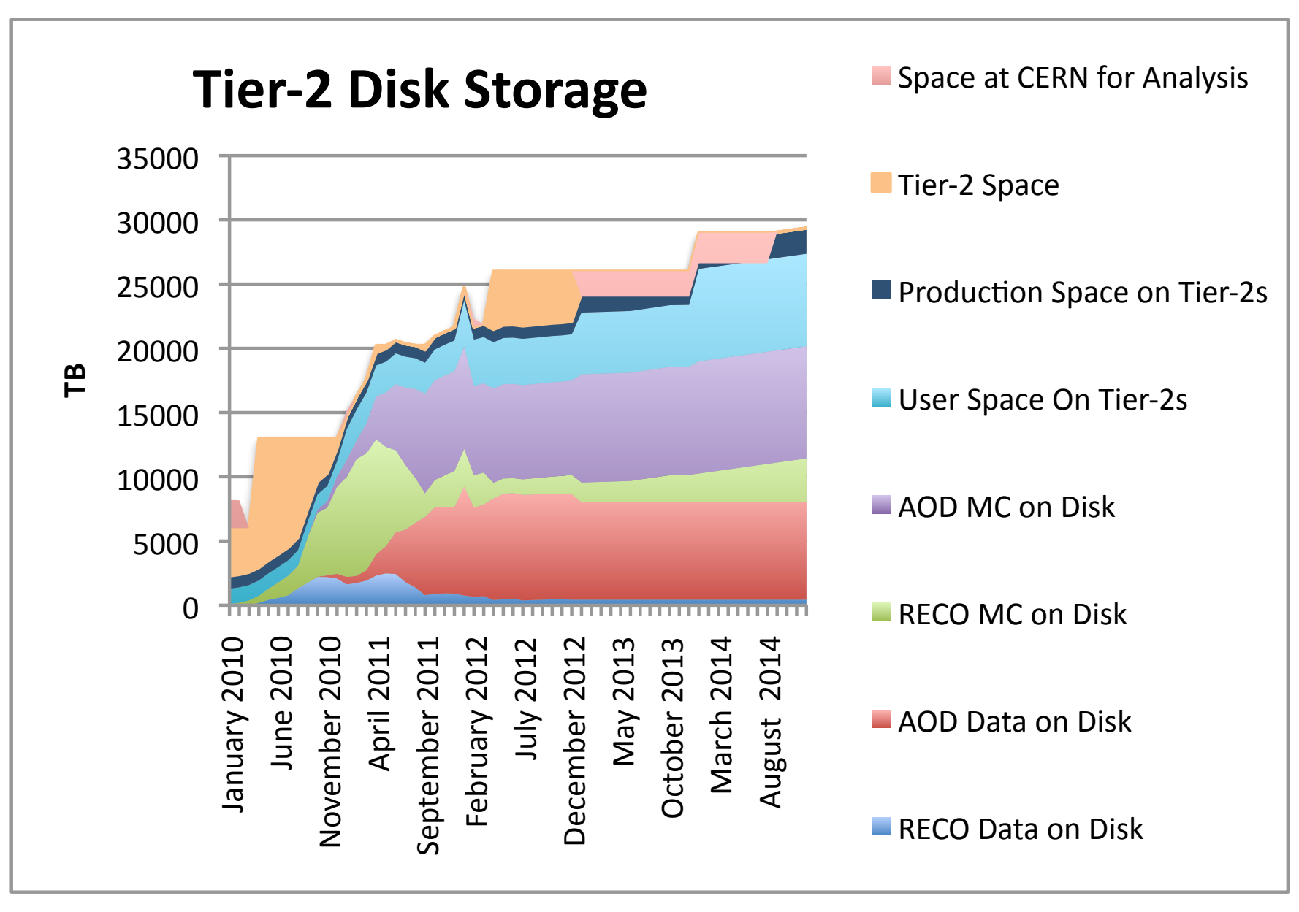
However, there was still a significant number of pending jobs even when the full CPU pledge is not being used. This suggests that further optimizations can be made in the assignment of jobs to computing sites.

Disk usage at Tier-2 centers was estimated to be 17-18 PB, about 70% of the pledged resources, at the end of 2011. The data management system tracked 13 PB of files, of which about 3 PB were very popular samples in disk space controlled by the Analysis Operations group. Another 4-5 PB of untracked data was dominated by user-owned files.

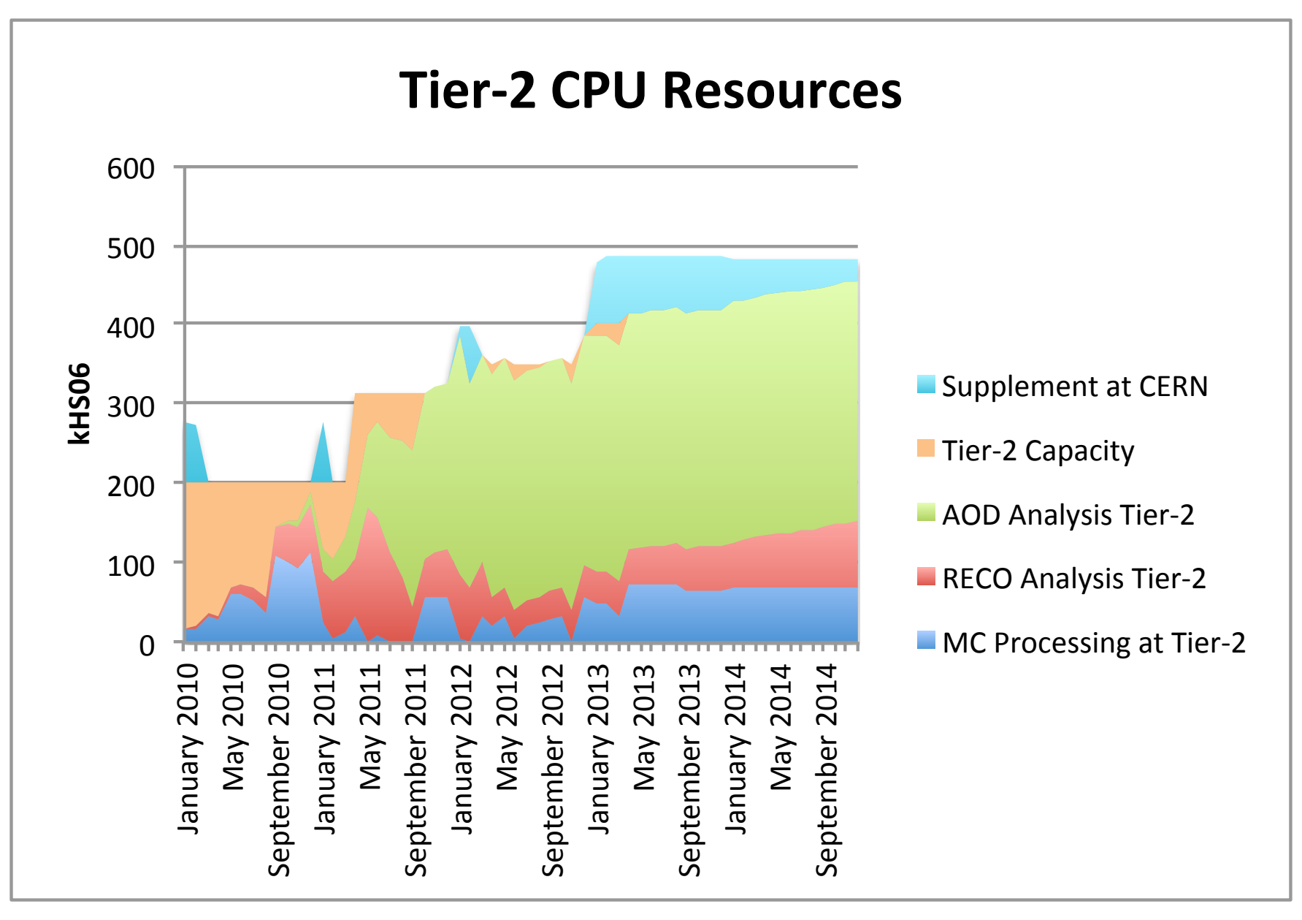
Physicists are making more efficient use of disk space, thanks to a wide-spread transition from the complete RECO data format to the smaller AOD format that is sufficient for most physics analyses. This allows for more datasets to be hosted at the Tier-2 sites.



2012: The CMS computing model allows us to make predictions of resource usage in the future. The model predicts that there will be some headroom in processing and storage resources at Tier-1 centers, but resources are more constrained at Tier-2.



Pressures on disk space at Tier 2 are expected to be relieved over the course of 2012 as sites deploy their pledged resources for this year. The observed migration of analyses to the AOD format is what makes this possible.



On the other hand, CPU resources at Tier 2 are expected to be heavily used throughout the year. Given the evidence that the assignment of jobs to sites is not optimal, we recognize that we face challenges in delivering the maximum amount of processing power to users.

Outlook: Is CMS living within its resources? The answer is yes, at least in the aggregate. In general the use of processing and storage resources is slightly below the amounts that have been pledged by the participating sites. This tells us that the computing models are valid, and that CMS is making good use of the deployed CPU and disk. But CMS has observed limitations that are localized in space and time. Some sites are routinely saturated, and are providing opportunistic resources beyond those pledged, or have large queues of pending jobs. At some times of the year, the total CPU pledges are fully utilized, whereas at other times there are cores going idle.

Thus, the challenge for the future is perhaps not in expanding the total resources available, but in making sure that the available resources are being used optimally. As the LHC continues to perform well and the experiments seek to extend their physics reach through more inclusive datasets, this optimization will become all the more important. The success that CMS has had so far in adapting its computing model to improve resource use suggests that these efforts will be successful in the future too.