CMS resource utilization and limitations on the grid after the first two years of LHC collisions

Giuseppe Bagliesi¹, Kenneth Bloom², Daniele Bonacorsi³, Chris Brew⁴, Ian Fisk⁵, Jose Flix⁶, Peter Kreuzer⁷ and Andrea Sciaba⁸

¹ Università di Pisa, Sezione dell’ INFN, I-56100 Pisa, Italy
² Department of Physics and Astronomy, University of Nebraska-Lincoln, Lincoln, NE 68588-0299, USA (corresponding author)
³ Università di Bologna, Sezione dell’ INFN, I-40127 Bologna, Italy
⁴ Rutherford Appleton Laboratory, Chilton, Didcot OX11 0QX, United Kingdom
⁵ Fermi National Accelerator Laboratory, Batavia, IL 60510-5011, USA
⁶ Port d’Informaci Científica (PIC), Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), E-28040 Madrid, Spain
⁷ III. Physikalisches Institut A, RWTH Aachen University, D-52056 Aachen, Germany
⁸ European Organization for Nuclear Research, CH-1211 Genève 23, Switzerland

E-mail: kenbloom@unl.edu

Abstract. After years of development, the CMS distributed computing system is now in full operation. The LHC continues to set records for operational performance, and CMS records data at more than 300 Hz. Because of the intensity of the beams, there are multiple proton-proton interactions per beam crossing, leading to ever-larger event sizes and processing times. The CMS computing system has responded admirably to these challenges, but some reoptimization of the computing model has been required to maximize the efficient delivery of data analysis results by the collaboration in the face of increasingly constrained computing resources. We present the current status of the system, describe the recent performance, and discuss the challenges ahead and how CMS intends to meet them.

1. Introduction

It is perhaps part of human nature for people to try to achieve their goals in life under the assumption that unlimited resources are available to do so. Indeed, the United States, the home of CHEP 2012, is a nation that is quite rich in natural and human resources, and thus it has always seemed possible to imagine endless frontiers for our quality of life, industrial production, and even scientific discovery [1]. But ultimately all kinds of resources are of finite extent, and we all must learn to live within our means.

This is true of computing for the Large Hadron Collider (LHC) too. The LHC experiments and their collaborating institutes have made significant monetary investments in computing hardware, in the forms of processing, storage, and wide-area network bandwidth. But due to the multi-petabyte data samples produced by the experiments each year that must be stored, transferred, and analyzed, the computing needs of the experiments are in a position to saturate the available resources. This has implications for how computing tasks must be performed, and ultimately for how the physics measurements are going to be produced. It is important to be...
able to plan for this, and then to make adjustments to the computing models to optimize the scientific output of the experiments.

In this paper, we discuss the use of computing resources by the Compact Muon Solenoid (CMS) experiment during the 2011 data run of the LHC, and what sort of limitations were encountered. Even last year, CMS had to make adjustments to its original computing model to handle the amount of data that the experiment produced. The 2011 experience has implications for the 2012 run, in which the collision events will be more complex (as discussed in Section 3) and thus require more computing resources. We also discuss some of the expected resource constraints for this year, and how CMS intends to live within them.

2. Review of the CMS computing model
The CMS computing model has been discussed extensively elsewhere [2]. Here we briefly summarize the important features. Computing facilities are organized into tiers, with particular specified roles for facilities in each tier and configurations that reflect those roles. Data from the experiment is immediately sent to the Tier-0 facility at CERN, where events are reconstructed and then archived to tape. Additional resources at CERN are modest, and are devoted predominantly to calibration and alignment tasks, rather than to simulation or user analysis. The total resources available at CERN in 2011 were 106 kHS06 [3] of CPU, 4.5 PB of disk and 21.6 PB of tape.

A copy of each event is then transferred over the wide-area network to at least one of the seven CMS Tier-1 centers, so that a backup copy can be written to tape. The Tier-1 centers are also responsible for re-processing the raw data with new calibrations, alignments and algorithms as needed; for making enriched subsamples (skims) of the data for further analysis; for some amount of the simulated event production; and for archiving simulated samples that are created at the site and elsewhere. A total of 131 kHS06, 16.1 TB of disk and 44.1 PB of tape were available across all of the Tier-1 centers.

All of the activities at Tier 0 and Tier 1 are centrally controlled by the CMS computing organization; physicist users do not have access to those facilities. Instead, users have access to approximately 50 CMS Tier-2 centers that are spread around the world. Processing resources are used for user-controlled physics data analysis and centrally-controlled simulation production. At the start of the LHC run, it was expected that the processing time would be split about evenly between those two activities, and that all simulation production would be done at Tier 2. However, it has turned out to be more optimal to also use the Tier-1 centers for simulation during periods when those facilities are not busy with re-processing or skimming functions. Now, event generation and detector simulation, resulting in simulated events that are in the same format as the real events produced by the detector, is split between the Tier 1 and Tier 2 centers, while all reconstruction of the simulation output into physics objects that can be used for data analysis is performed at Tier 1. This leaves more processing resources available for user analysis at Tier 2. Disk resources at the site are mostly reserved for hosting of real and simulated data samples, with smaller amounts reserved for user files and scratch space for simulation production. Tier-2 sites do not run tape facilities. In 2011, a total of 312 kHS06 and 20.4 PB of disk were available at Tier 2.

3. Resource usage in 2011
The primary driver of computing resource usage is the amount of LHC data that is available. The machine was originally expected to run pp collisions at 7 TeV center of mass energy for over 5.2M seconds, followed by an additional 0.7M seconds of heavy-ion running. The actual LHC operational performance agreed with this within about 5%. The expected CMS trigger rate for pp running was 300 Hz; as a single event can be placed in multiple datasets, the total event
output rate was expected to be 375 Hz. The actual average trigger rate fluctuated between 350 and 430 Hz over the course of the pp run.

During that time 1.5 billion events were recorded. As the year progressed, the number of “pile-up” interactions, in which multiple protons collide in a given beam crossing, increased. By the end of the run, there were typically 16 to 17 pp interactions per beam crossing. This was also consistent with the original expectations. The larger the number of pile-up interactions, the greater the size of each data event on disk or tape and more time is required to reconstruct the event. Table 1 shows the expected and observed event sizes for different pile-up scenarios and different data formats. The “reconstruction” (RECO) format includes the full output of the reconstruction, which is often needed for detailed studies of detector or algorithm performance, while the “analysis object data” (AOD) format keeps only summary information that is expected to be sufficient for most physics measurements.

The computing facilities at sites available to CMS host and process these data. A breakdown of the use of processing and storage resources can be obtained from the Worldwide LHC Computing Grid (WLCG) accounting reports [4].

3.1. Tier 0
The Tier-0 cluster is provisioned to handle peak demand during data collection. There were a number of periods during 2011 when the facility was full, even during times between LHC fills, when CMS was not taking data. CMS also made use of an “overspill” scheme that allowed jobs to use shared CERN CPU resources, such as the public batch queues. Figure 1 shows the number of running and pending Tier-0 jobs during a busy period in October 2011, when both the dedicated and shared resources were in use. Even with the extra resources, there was still a significant number of pending jobs at some times.

However, even if all of the job slots were full, the CPU utilization in the cluster only reached about 70%. The memory footprint of the reconstruction executable was larger than expected, and not all cores in each compute node could be used. Since 2011, CMS has successfully reduced the memory consumption, but some of those gains will be offset by the increased pile-up expected in 2012.

3.2. Tier 1
Figure 2 shows the pledged and utilized CPU resources across all Tier-1 centers for CMS in each month of 2011. The amount of pledged CPU rises in April, which is when the pledges for the year take effect. Averaged over the whole year, CMS used 87% of the available Tier-1 processing resources. In 2010, the usage never exceeded 60%. This improvement in usage was a result in moving some of the simulation production from Tier 2 to Tier 1.

Table 1. Comparison of actual and expected event size in kilobytes for different pile-up scenarios and different data formats.

| 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                     |
Figure 1. Number of Tier-0 jobs during five days of October 2011. Both dedicated (cmst0) and shared CPU resources are shown.

Figure 2. Utilization of CMS Tier-1 processing resources in 2011. The blue line indicates the pledged resources, and the green bars the used resources.

However, not all sites were used equally. Figure 3 again shows the pledged and used CPU resources, broken out by site for each of the seven CMS Tier-1 facilities. The most active site provided 113% of its pledged resources, while the least active provided only 34%. CMS hopes to improve the performance of the lagging sites in the coming year.

The disk and tape usage at Tier 1 can also be obtained from the WLCG accounting portal.
Figure 3. Utilization of CMS Tier-1 processing resources in 2011, separated out by Tier-1 site. The blue line indicates the pledged resources, and the green bars the used resources.

The storage usage for all four LHC experiments is shown in Figure 4. At the end of 2011, CMS was using 24.6 PB of tape at Tier-1 sites, consistent with the planning expectations, with a
total of 45 PB available. 17 PB of disk was in use at the centers, slightly more than the pledged amount.

Figure 4. Use of Tier-1 disk (left) and tape (right) resources for all four LHC experiments. CMS usage is shown in green. The total installed capacity is indicated by the solid line, and the pledged resources by the dashed line.

3.3. Tier 2
The CMS Tier-2 centers are much more diverse in both implementation and usage than their Tier-1 counterparts. The size of Tier-2 sites varies by almost an order of magnitude, with some having as much CPU and disk as some of the Tier-1 sites, while others are only a handful of nodes and disks. Some sites are dedicated resources for CMS, while others are shared across multiple experiments. The centers are used for both centrally-managed processing and on-demand activity that is driven by individual users. How the latter manifests itself at sites is affected by factors as diverse as the particular datasets hosted by each site, the reliability of the site services, and the preferences of individual users (including their perception of the reliability of site services). Thus it is difficult to make comparisons between individual sites, and more reasonable to consider the aggregated performance of Tier-2 centers.

Figure 5 compares the CPU usage at all CMS Tier-2 sites to the pledged amount available during 2011. Averaged over the entire year, the fraction of pledged CPU that is actually used is 88%. Most of the deficit occurred in the early months of the year; at the end of the year, when the full 2011 LHC dataset was available, usage rates were close to or exceeding the pledge. This demonstrates that the pledged resources are indeed needed by the collaboration during peak demand. There is also a noticeable variation of CPU pledge and usage among various sites and various nations hosting the sites. In the United States, home to about 17% of the CMS Tier-2 processing resources, 126% of the pledge was used during 2011, i.e. there was a significant overflow into opportunistically available CPU.

Figure 6 shows the number of running and pending jobs at CMS Tier-2 sites during 2011. The numbers track well with the CPU consumption over time; when the CPU usage was close to the pledge, the number of pending jobs grew. There are a significant number of pending jobs when the full CPU pledge is not being used, suggesting that further optimizations can be made in the assignment of jobs to sites.

The usage of disk at the sites is harder to estimate, as some fraction of the space goes to user files that are not tracked by the data management system. At the end of 2011, 13 PB of data tracked by the management system was on disk at Tier-2 centers. Of that, more than 3 PB was
in disk space controlled by the Analysis Operations group to host datasets of common interest to a large number of users. We estimate that another 4-5 PB of untracked data was also in place. The total usage is about 70% of the pledge for 2011. Disk usage does not increase steadily over time; the amount of usage can be significantly increased through occasional deletion campaigns to remove older datasets that are no longer of interest.

It is interesting to note that CMS physicists are making more efficient use of disk space over time. In 2010, the RECO data format was widely used, but in 2011, most analyses had migrated to the smaller AOD format. This is demonstrated in Figure 7, which shows the number of file accesses by data tier as recorded by the CMS data popularity service. The switch to the smaller format means that more datasets can be hosted on disk at any one time.

4. Anticipated resource constraints in 2012
The CMS computing model allows us to make predictions of resource usage in the future [5]. While the model predicts that there will be some headroom in the usage of processing and storage resources at the Tier-1 centers, the Tier-2 situation is much more constrained. This is demonstrated in Figure 8, which shows the model predictions for Tier-2 CPU and disk usage over time, separated out by different functions. For storage, pressure should be relieved as sites deploy their 2012 pledges. This has been achieved in part by the migration to the smaller AOD format. However, CPU is expected to be constrained throughout the year. Given that we know there are further optimizations to be made in matching jobs to sites, we recognize that we face challenges this year in delivering the maximum available processing power to users.

Figure 5. Utilization of CMS Tier-2 processing resources in 2011. The blue line indicates the pledged resources, and the green bars the used resources.
5. Conclusions
Is CMS living within its computing means? Overall, the evidence suggests yes, at least when considering computing resources as a whole. In general, the use of processing and storage resources is slightly below the amounts that have been pledged by the participating sites. This suggests that the computing models that drive the annual requests are valid within their expected uncertainties, and that the collaboration is making good use of the deployed CPU and disk. However, CMS has encountered resource limitations that are localized in space and time. Some sites are routinely saturated, providing opportunistic resources beyond those pledged, or having 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 expanding the total resources available, but making sure that resources that are already available are 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.

References
[1] V. Bush, “Science, The Endless Frontier,” United States Printing Office, Washington (1945).
[2] C. Grandi, D. Stickland, L. Taylor et al., “The CMS Computing Model”, CERN-LHCC-2004-035/G-083 (2004); D. Bonacorsi, J. Phys.: Conf. Ser. 331, 072005 (2011).
[3] http://w3.hepix.org/benchmarks/doku.php/
[4] http://www4.egee.cesga.es/accounting/tier1_view.html and http://www4.egee.cesga.es/accounting/
tier2_view.html
[5] P. Kreuzer, “Trying to Predict the Future – Resource Planning and Allocation in CMS,” this conference.
Figure 6. Number of running (top) and pending (bottom) jobs at CMS Tier-2 sites during 2011.
Figure 7. CPU time spent processing different CMS data formats. The “SIM” designation represents simulation files that include generator-level information.

Figure 8. CMS computing model predictions for the use of Tier-2 processing (left) and storage (right) resources, broken down into different functions.