Trying to predict the future – resource planning and allocation in CMS

Kenneth Bloom[1], Ian Fisk[2], Peter Kreuzer[3], Gonzalo Merino[4]

[1]Department of Physics and Astronomy, University Nebraska-Lincoln, USA

[2]Fermi National Accelerator Laboratory, USA

[3]RWTH Aachen University, III. Physikalisches Institut A, Germany (corresponding author)

[4]CIEMAT Madrid, Spain

E-mail: Peter.Kreuzer@cern.ch

Abstract. In the large LHC experiments the majority of computing resources are provided by the participating countries. These resource pledges account for more than three quarters of the total available computing. The experiments are asked to give indications of their requests three years in advance and to evolve these as the details and constraints become clearer. In this paper we will discuss the resource planning techniques used in CMS to predict the computing resources several years in advance. We will discuss how we attempt to implement the activities of the computing model in spread-sheets and formulas to calculate the needs. We will talk about how those needs are reflected in the 2012 running and how the planned long shutdown of the LHC in 2013 and 2014 impacts the planning process and the outcome. In the end we will speculate on the computing needs in the second major run of LHC.

1. Introduction

The computing facilities for storing, processing, and analyzing the LHC data are the final step in a long series to fully realize the value of the LHC program. The resources available to CMS [1] are critical to the exciting physics program; they have been requested and utilized in an efficient manner during the first two years of LHC collisions, and will also play a crucial role during the LHC shutdown period.

The yearly resource requests are typically submitted by the experiments in the first quarter to the Computing Resource Scrutiny Group (CRSG [2]) and endorsed at the bi-annual meetings of the LHC Computing Resource Review Board (C-RRB [3]). Upon endorsement, the participating countries and institutes place their resource pledges to the experiments and purchase the hardware, to be deployed in the following year.

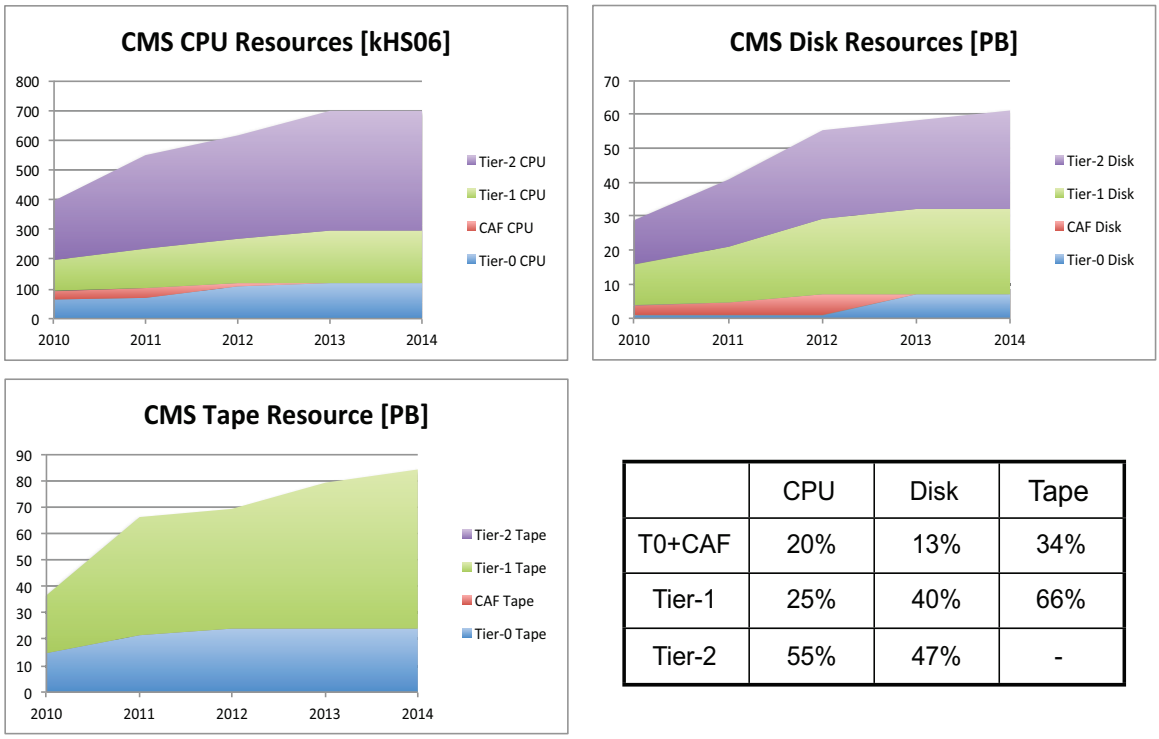
Sites operating under the WLCG Memorandum of Understanding [4] are typically accounted within a "Federation" of a participating country. Federation resource pledges and individual site capacities are tracked via the "REBUS" portal [5]. The site resource utilization level is published in monthly reports by the WLCG management. In parallel, CMS routinely tracks the site readiness and reliability, in particular that related to the CMS specific processing or storage workflows [6].

In this paper we present the techniques used by the CMS experiment to predict the computing resource needs, which is the basis of the resource request procedure to sites that take part in the distributed computing operations. While the official requests are based on a yearly process, the resource predictions are made typically two years ahead of time, to give sites enough lead time in their procurement cycles. We review the main input parameters and model assumptions that enter into the calculation. We also underline the computing activities planned during the upcoming long LHC shutdown period and the implication for the resource needs during that period. We conclude by summarizing the main ingredients that will affect the long-term resource needs after the shutdown.

2. CMS workflows and deployed resources on the Worldwide LHC grid (WLCG)

The baseline CMS Computing Model [7] is based on a tiered architecture: the Tier-0 center at CERN is dedicated to primary archiving, calibration and prompt reconstruction of the data; the Tier-1 centers host the secondary archiving, re-processing, skimming of the data, as well as data serving; the Tier-2 centers are dedicated to event simulation and data analysis. The model has evolved since the beginning of the LHC collisions, partially to optimize the resource utilization. For example, the Tier-1 centers are also used for event simulation during periods with low re-processing activity. The Tier-3 centers host a substantial fraction of the final data analysis activities by the majority of CMS physicists worldwide; their resources may also be utilized centrally by the experiment in opportunistic manner, but they are not included in the central resource accounting, hence Tier-3 centers are not covered in what follows.

In Fig. 1 the CMS CPU, disk and tape resource evolution between 2010 and 2014 are shown. The last two years are the result of the resource prediction calculation, as described in more detail below. The table in the bottom right corner shows the fraction of computing resources at various tiers; the resources at CERN are relatively modest compared to those at the distributed sites. Another particularity shown in Fig. 1 is that the dedicated CERN Analysis Facility (CAF, in red color), used for critical calibration of the data during collisions running, will be entirely moved to the Tier-0, which will become a general purpose analysis facility for all CMS collaborators, similar to any other Tier-2 site.



	CPU	Disk	Tape
T0+CAF	20%	13%	34%
Tier-1	25%	40%	66%
Tier-2	55%	47%	-

Figure 1: CMS computing resource evolution and fractions at various Tier levels

3. Resource Planning Techniques

The CMS Computing project has developed spreadsheets to model the activities of the experiment realistically and thereby estimate the future resource needs. Yearly comparisons between original resources predictions and their actual utilization have been very satisfactory so far, after the first two years of LHC operations. These prediction techniques are also regularly used to simulate variations to the default CMS Computing model and estimate their impact on the resource needs. The spreadsheets provide a monthly timeline of CMS processing and storage workflows at various tiers. Below we review the main input parameters and model assumptions that contribute to the resource needs calculations.

3.1. Main LHC and CMS parameters

The main input parameters to the computing needs are the integrated LHC beam time, the number of primary collisions per LHC bunch crossing or "Pile-Up" conditions, the CMS trigger rates, and the event sizes. While the latter two have been relatively stable since the beginning of collisions, the amount of Pile-Up has increased by a factor 15 (~30 in 2012) and the yearly integrated beam time has increased by a factor ~2, since 2010. These increases are reflected in the resource deployment gradients visible between 2010 and 2012 in Fig.1.

The resource needs scale differently at each tier. The Tier-0 needs scale with the incoming rates of data to be reconstructed. The Tier-1 needs scale with the total integrated data and with the length of time allocated to complete a reprocessing pass. The Tier-2 needs scale with the integrated data volume and with the corresponding simulated event sample.

The fact that the resource needs at Tier-1 and Tier-2 centers are not directly linked to the data taking activity, but rather to the delayed and repeated re-processing and simulation of the data, results in potentially increased resource requests also during non-data taking periods. Moreover, in 2012 CMS will acquire ~50% more events that the nominal 300Hz data taking rate, to be "parked" and reconstructed later during the long LHC shutdown, resulting in increased resource needs for 2013 and 2014. This is illustrated for example in Fig.2, which shows the output of the spreadsheet calculation for the Tier-1 CPU needs.

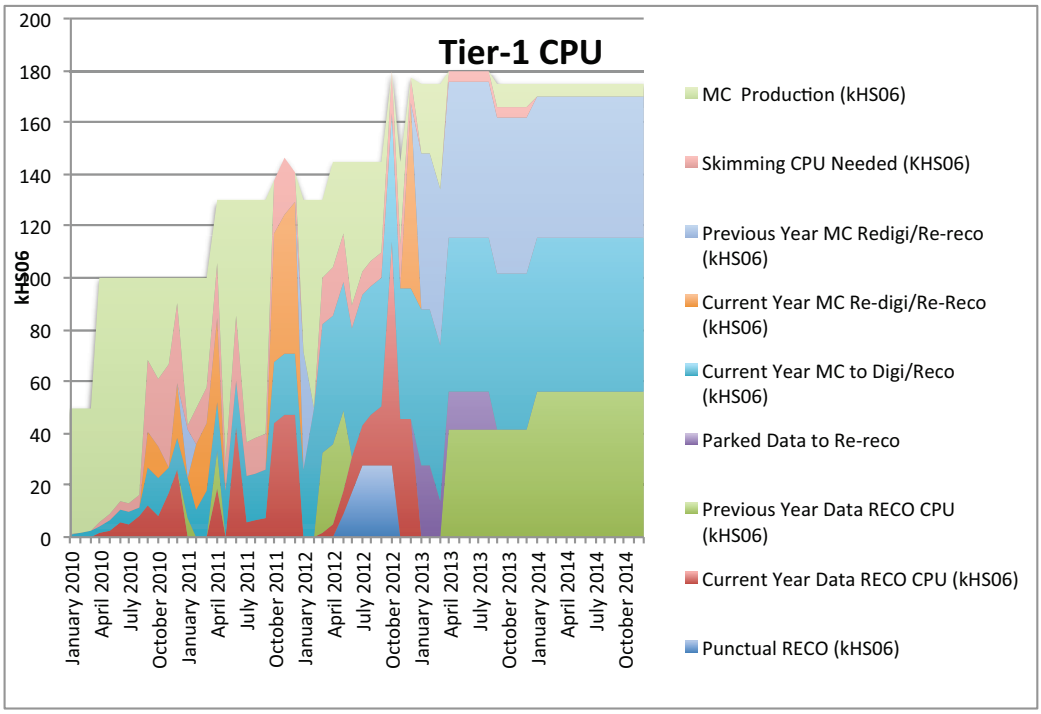


Figure 2: CMS CPU Needs at the Tier-1 centers, in kHS06 [8].

3.2. Folding in the CMS processing and analysis model

The resource calculations spreadsheets contain a number of CMS computing and analysis model assumptions as described below.

At Tier-1 centers, the most important assumption is the number of planned data reprocessing campaigns. As an example, in 2011 CMS went through 25 "re-reco" campaigns, mostly for only a fraction of selected events compared to the full sample, but with at least two reconstruction passes per event over the course of the year. In 2012, CMS is again planning at least two full re-processing passes per event.

At the Tier-2 level, one important ingredient is the amount of analysis based on summarized and smaller data types: with increasing confidence on the reconstructed objects and the quality of the data, a growing number of CMS physicists are basing their analysis on Analysis Object Data (AOD), which are smaller and hence easier to handle than the reconstructed (RECO) data. This is illustrated in Fig 3, where the CPU time spent on various data types since Fall 2011 is shown. These data were collected by the CMS data popularity service [9]. The assumption made for the 2012 and beyond resource calculation is that 95% of the CMS analysis are based on AOD.

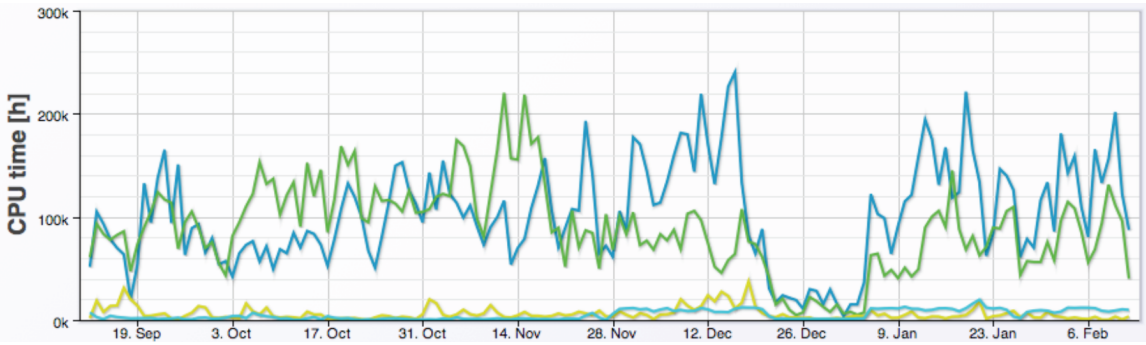


Figure 3: CPU time spent by CMS physicists on summarized data objects (AOD) in blue (simulation) and dark green (data), compared to larger data objects (RECO data and simulation) in other colors, between Sep. 2011 and Feb. 2012.

The resources needs are typically more challenging to predict for Tier-2 centers, given the less centralized and more "chaotic" nature of the resource utilization. As stated above, the main assumption for Tier-2 centers is that both disk and CPU needs scale with the accumulated and simulated data volume. More fine-grained assumptions such as the period of the year when data are analyzed, with peak activities prior to large conferences, or the type of analysts from participating institutes, like the number of students involved, have not been folded in so far. This has proven to be sufficient, at least when comparing the pledged resources with those actually used, as illustrated in Fig. 3 for the CPU hours at Tier-2 centers during 2011.

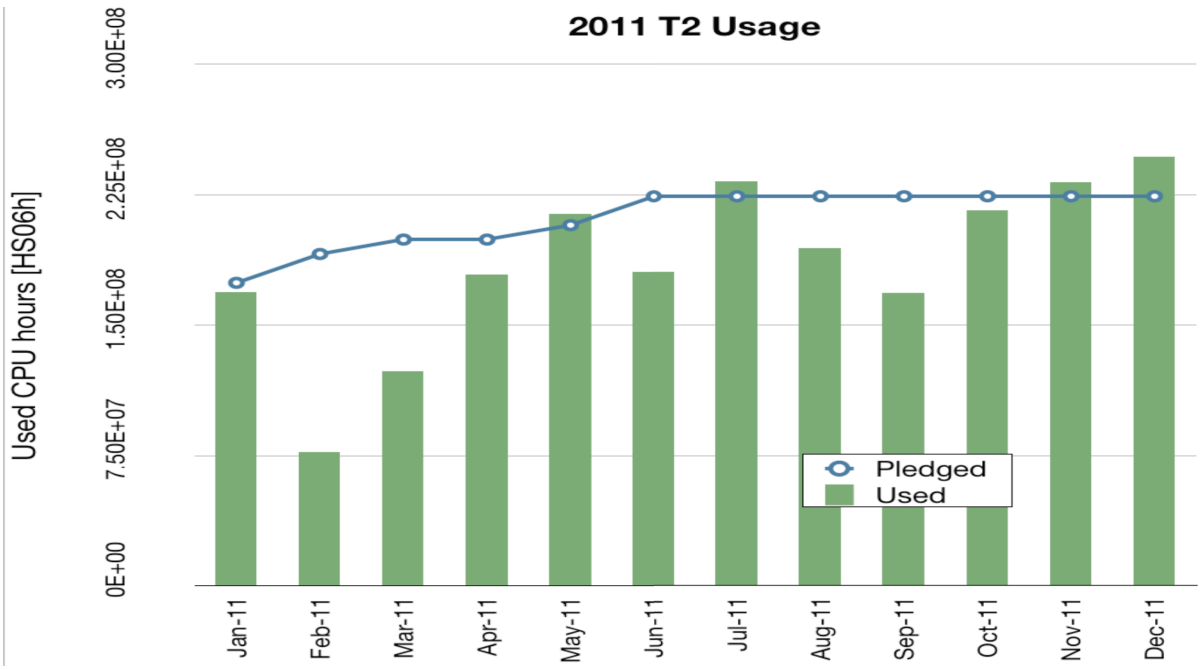


Figure 4: Pledged CPU hours (in blue), compared to the used resources (in green bars) at Tier-2 centers during 2011 (in units of HS06h).

4. Resource Needs: present and future

In this Section we show the outcome of the resource needs calculation for 2012 and the following 2 years.

4.1. Resource Request 2012, 2013, 2014

As shown in Table 1, the largest increase in 2013 needs compared to 2012 is for Tier-1 disk and tape resources, as a consequence of the accumulated data volume, while the increase in CPU needs for Tier-0 is mainly related to the growing amount of Pile-Up in the acquired data.

The resource request procedure is typically launched early spring every year and presented to participating institutes and countries for the upcoming accounting year, in order to give the latter enough time for their resource procurement and deployment.

	2012	2013	%increase over 2012	2014	% increase over 2013
T0 CPU [kHS06]	107	121	13%	121	0%
T0 Disk [TB]	1000	7000	-	7000	0%
T0 Tape [TB]	23000	23000	0%	23000	0%
CAF CPU [kHS06]	14	0	-	0	0%
CAF Disk [TB]	6100	0	-	0	0%
T1 CPU [kHS06]	145	145	0%	145	0%
T1 Disk [TB]	22000	26000	18%	26000	0%
T1 Tape [TB]	45000	55000	22%	60000	9%
T2 CPU [kHS06]	350	350	0%	350	0%
T2 Disk [TB]	26000	26000	0%	29000	12%

Table 1: CMS resource request for 2012, 2013 and 2014.

5. A look into the longer term future and conclusions

The long shutdown for the LHC upgrade will end by September 2014, with stable beams expected by Spring 2015, at larger beam energy and higher luminosity. Several unknowns are yet to be understood before being able to make a reliable computing resource prediction for 2015 and beyond, such as the expected beam time or the bunch spacing in the colliding beams. In case the latter stay at the current 50ns spacing (as opposed to the nominal design of 25ns), the resulting Pile-Up conditions could be as high as 60 at luminosities of $10^{34}cm^{-2}s^{-1}$, which would have a strong impact on the computing resource planning.

While waiting for a refinement in the knowledge of the 2015 data taking conditions, CMS is also refining the computing resource planning tools presented above, in order to make more reliable predictions for the future.

References

[1] CMS Collaboration, "The Computing Muon Solenoid Technical Proposal", CERN-LHCC 94-38, 1994.
[2] http://lcg.web.cern.ch/lcg/crsg.htm
[3] http://lcg.web.cern.ch/lcg/crrb.htm
[4] http://lcg.web.cern.ch/lcg/mou.htm
[5] http://wlcg-rebus.cern.ch/
[6] J.Flix Molina et al., "Towards higher reliability of CMS Computing Faciliites", in these proceedings.
[7] CMS Collaboration, "CMS Computing Project Technical Design Report", CERN-LHCC 2005-023, 2005.
[8] https://hepix.caspur.it/benchmarks/doku.php
[9] D. Giordano et al., "Implementing data placement strategies for the CMS experiment based on a popularity mode", in these proceedings.