CMS Computing Upgrade and Evolution

José M. Hernández, *CIEMAT, Madrid, Spain*

***Abstract*—The distributed Grid computing infrastructure has been instrumental in the successful exploitation of the Large Hadron Collider data leading to the discovery of the Higgs boson. The computing system of the CMS experiment will need to face new challenges from 2015 on when LHC restarts with an anticipated higher detector output rate and event complexity, but with only a limited increase in the computing resources. A more efficient use of the available resources will be mandatory. CMS is improving the data storage, distribution and access as well as the processing efficiency. Remote access to the data through the Wide Area Network, dynamic data replication and deletion based on the data access patterns, and separation of disk and tape storage are some of the areas being actively developed. Multi-core processing and scheduling is being pursued in order to make a better use of the multi-core nodes available at the sites. In addition, CMS is exploring new computing techniques, such as Cloud Computing, to get access to opportunistic resources, or as a means of using without interference dedicated computing resources, such as the high level trigger farm, while they are not utilized for their main purpose. We discuss in this paper the ongoing work and plans to upgrade and evolve the CMS computing system.**

I. Introduction

The Large Hadron Collider (LHC) [1] at CERN in its first physics Run (2010-2013) has delivered trillions of proton-proton collisions to the experiments. Few billions of these collisions were selected in real time and recorded on magnetic tape resulting in a data volume of few tens of terabytes. The raw data recorded by the detectors are processed to reconstruct physics objects. The reconstruction process generates a similar amount of data. In addition, large samples of simulated data need to be produced. All in all, the LHC experiments have generated during the LHC Run 1 about 200 PB of data.

To store, process and analyse such a huge amount of data, it has been necessary to develop and implement an innovative computing infrastructure, the Worldwide LHC Computing Grid (WLCG) [2]. The WLCG is a global collaboration that joins computing centres from around the world via Internet, using large bandwidth networks. With this Grid, computing capacity and data storage can be shared in hundreds of thousands of computers distributed around the world. WLCG establishes a worldwide collaborative computing environment that serves some 10000 member physicists of the LHC experiments, regardless of their location, and allows them to access data in real time. WLCG is formed by more than 150 centres spread around the world. It currently has a storage capacity of 400 PB (200 PB disk and 200 PB magnetic tape) and about 200000 processing cores.

The LHC computing challenge has resulted in a great success. An unprecedented amount of data have been analysed in record time delivering a wealth of scientific results, being the most prominent one the discovery of the Higgs boson.

The organizational model of the WLCG resources was initially based on a hierarchical structure of centres organized according to their functionality at several levels or tiers (see Fig. 1). At the Tier-0, located at CERN, the raw data with the readings of the detector sensors are stored on magnetic tape for long-term custody. The data are also copied onto magnetic disk for immediate processing, consisting on the identification of the particles produced in the collisions and the reconstruction of their properties, e.g. trajectory, energy and electric charge. The Tier-0 center has some 30000 processors and more than 100 PB of storage capacity. From the Tier-0, the data are distributed to some 10 major computing centres throughout the world, the Tier-1 centres. The Tier-0 is connected to these centres via a private optic communications network with a large bandwidth (dozens of gigabits per second). The Tier-1 centres, with a total capacity of about 70000 processors and 170 PB of storage, are in charge of saving a second copy of the data on tape, selecting the collisions that are considered to be interesting, reprocessing the data if necessary when the reconstruction techniques or detector calibrations are improved, and distributing the data to computing centres located in universities and laboratories, the Tier-2 centres. There are some 150 Tier-2 centres hosting around 100,000 processors, which analyse the data and produce simulated data that are used to study the measurements that can be made, compare to actual data, calibrate the detector response, etc. The disk space of Tier-2 centres, some 100 PB, is used as a sort of volatile data cache. Obsolete or little used data are deleted and replaced by reprocessed or frequently accessed data. The Tier-3 centres are located in research centres and complement the resources of the Tier-2 centres by contributing resources for the local scientific community without a formal agreement to supply a service to the global organization. Figure 2 depicts the computing workflows described above.

The Compact Muon Solenoid experiment(CMS) [3] is one of the 4 experiments at the LHC. The institutions participating in CMS contribute with about a third of the WLCG computing resources. This paper focuses on the evolution of the CMS distributed computing system during the LHC Run 1 and the planned evolution and upgrade in view of the Run 2 (2015-2017).

II. CMS Computing during LHC Run 1

During the Run 1 CMS used continuously all available resources. In Figure 3 we plot the average number of compute slots continuously occupied running jobs. The ramp-up in compute resources during Run 1 and the saturation of the approximately available 70k slots available by the end of the

On behalf of the CMS Collaboration.

978-1-4799-0534-8/13/$31.00 ©2013 IEEE

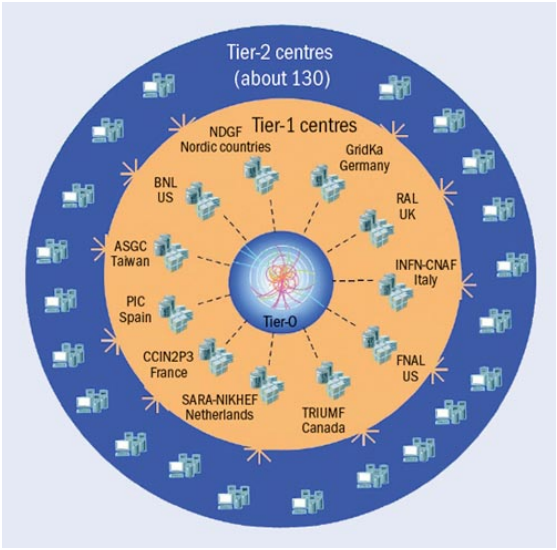

Fig. 1. Hierarchical structure of computer centres in WLCG.

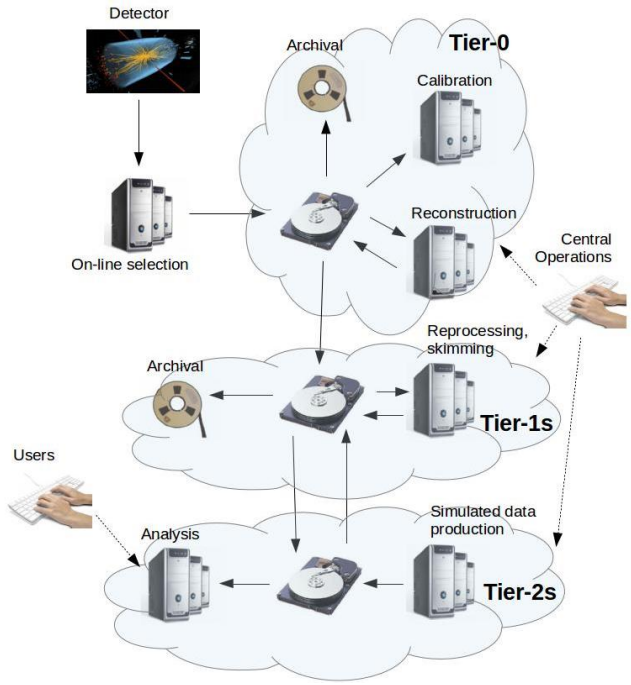

Fig. 2. CMS computing workflows.

run are clearly visible. About 500k jobs were completed daily by the end of Run 1 (see Figure 4).

The centres were continuously exchanging data at high speed. During Run 1 about 100 PB of data were moved worldwide between CMS sites (see Figure 5). About two thirds of the data transfers had a Tier-2 as destination. This flow corresponds to the data being distributed for analysis.

During Run 1 significant modifications and improvements were implemented in the CMS computing model based on the operational experience and the technological evolution. The main driver of the changes was the excellent reliability of the

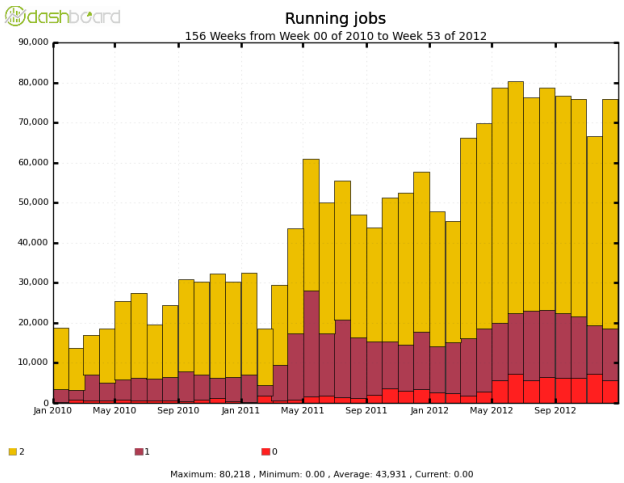

Fig. 3. Number of processing slots continuously occupied running CMS jobs during LHC Run 1.

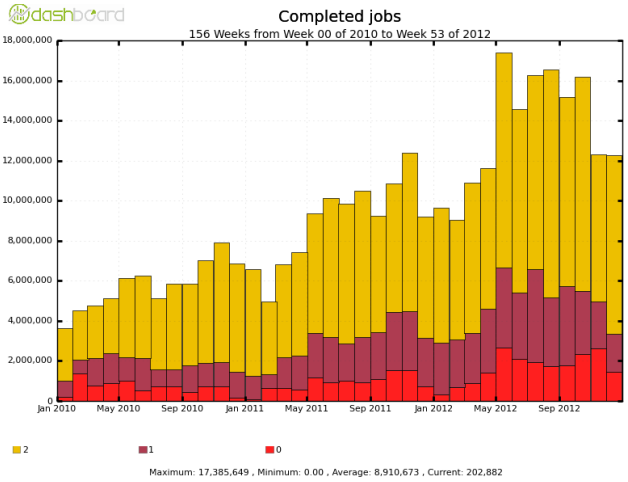

Fig. 4. Number of completed CMS jobs per week during LHC Run 1. .

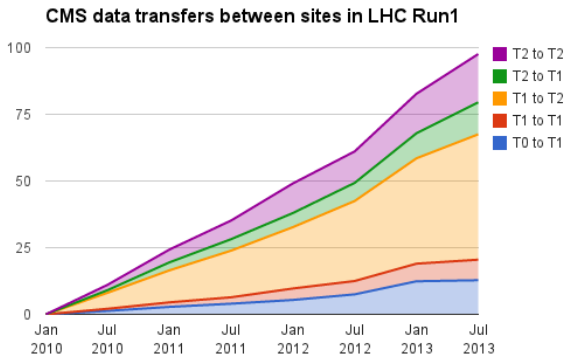

Fig. 5. Cumulative data transferred between CMS sites during LHC Run 1.

wide area network as well as its large increase in capacity, contrary to what had been assumed when the hierarchical

model of centres was designed. Data distribution and replication was optimized by moving away from the hierarchical data distribution model and allowing transfers between any two centres. Data access from the jobs was also optimized. The data access model, where data are pre-located at the sites and jobs are sent to those sites hosting the input data required by the jobs, was relaxed and remote data access (data read by a job running at a site through the WAN from a remote site) was allowed for certain cases such us fall-back of local access on failure, interactive browsing of data, disk-less sites and overflow of busy sites (allow running jobs on sites not hosting the requested input data when the sites hosting a replica of the data are busy running other jobs).

The access of the jobs to the experiment software was optimized as well. In the initial model the experiment software releases were pre-installed at all sites so that jobs had local access to them. The operational effort was large and the installation latency of new releases was significant in some cases. CMS moved to a model where software releases are installed centrally at CERN and worker nodes at sites mount the software repository by means of the CERN virtual machine file system (CVMFS) [4], a http caching file system optimized for efficient and scalable software delivery. Conditions data were provided to the processing applications by a similar mechanism. FroNTier [5] is a simple web service approach providing client HTTP access to a central database service. Because of the read only nature of the data, Squid proxy caching servers are maintained near clients and these caches provide high performance data access.

III. Computing evolution for LHC Run 2

WLCG computing during LHC Run 1 has been very successful but the second physics run from 2015 poses new challenges. The beam energy of the accelerator will almost double and the collision rate will be significantly increased. The events recorded by the experiments will be more complex, with higher particle multiplicity, and consequently, more CPU resources will be needed to reconstruct the data (about a factor of two according to simulations). At the same time, it is planned to increase the data recording rate from 400 Hz to 1000 Hz (from about 200 MB/s to 500 MB/s) in order to keep a similar physics sensitivity as in the Run 1.

Such a large increase in computing resources can probably not be accomplished by the start of Run 2. CMS will need to evolve its computing system towards a more efficient and flexible use of the available resources [6]. CPU and storage needs will have to be reduced by performing less reprocessing passes of the data (aiming at running just one reprocessing pass at the end of the year's data taking), reducing the amount of simulated events, trying to make the data format more compact and reducing the replication factor of the data.

Effort will be invested in automating data replication and deletion by making use of data access information. Unused data will be automatically purged from the Tier-2 caches and data frequently access will be automatically replicated as needed.

The rigid separation of tasks performed by the different Tier sites will be removed. Given the presumably insufficient

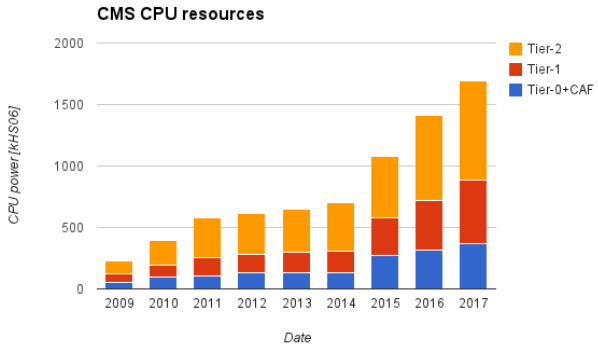

Fig. 6. CMS CPU power resources in kHS06 units. For reference, a modern CPU core has a power of about 10 HS06.

amount of resources at the Tier-0 to perform the prompt reconstruction of acquired data, approximately half of the data will be promptly reconstructed at the Tier-1 sites. Tier-1 centres will be used as well for production of simulated data and data analysis, tasks reserved so far to Tier-2 centres. In turn, Tier-2 sites will be used for simulation reconstruction, a task that had been performed exclusively at the Tier-1 centres.

CMS will investigate the possibility of producing group analysis datasets in a centralized way. By converting the first analysis step into an organized workflow, the goal is removing redundancies in analysis processing and storage, reducing operational workloads while improving the turnaround for users. Chaotic analysis should be limited to what really is user specific.

A. CMS Computing resources

In figures 6, 7 and 8 we show the CPU, disk and tape resources available to CMS until now and the requests for Run 2 (2015-2017). The ramp-up of resources during Run 1 and the plateau during the ongoing shutdown are clearly visible. About a 25% increase in resources is requested yearly for Run 2 according to the expected available funding. From the evolution of technology, it is expected to have a similar lowering in price for the hardware so that a flat funding could cope with the requested increase of resources.

CMS will try to access resources not owned by the collaboration but that could be used to execute CMS workflows. These so-called opportunistic resources, High Performance Computing clusters, unused capacities at Grid sites and academic or commercial Clouds that allow opportunistic usage, even volunteer computing, could result in a significant increase in capacity at a low cost to satisfy capacity peaks. The strategy to incorporate these resources is to try to be as less invasive as possible. Interfaces to these resources have been included in the CMS workload management system. Jobs running at the opportunistic nodes read input data through the WAN, copy output data to a remote CMS site, and access the experiment software remotely via CVMFS so that no CMS specific configuration is required.

CMS has also implemented interfaces to Cloud resources. Clouds complement and extend the Grid. They allow the

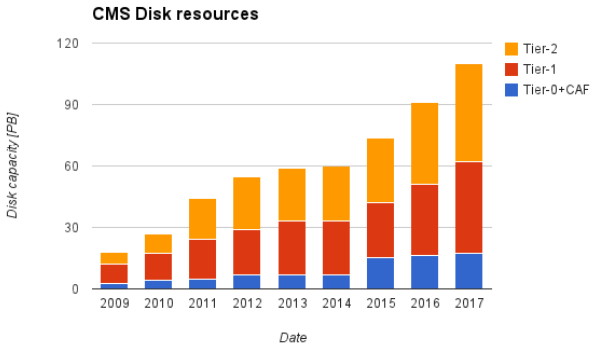

Fig. 7. CMS disk resources.

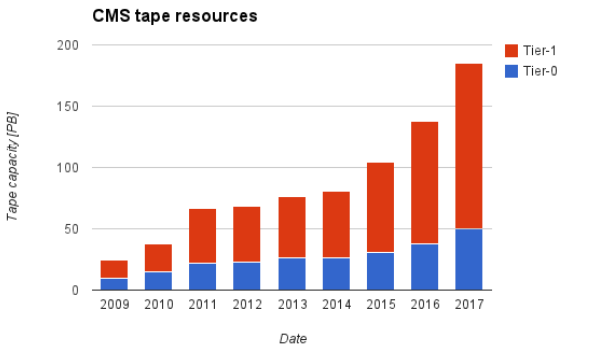

Fig. 8. CMS tape resources

sharing of resources to achieve coherence and economies of scale and turn computing into a utility providing infrastructure as a service. Hardware virtualization help to decrease the heterogeneity seen by the users. Virtual machines provide a uniform user interface to the resources integrating diverse resources manageably and isolating software from physical hardware. Some Grids sites are offering their resources through Cloud interfaces and commercial Clouds are proliferating. They might become an alternative to resource provisioning.

In order to utilize all available CPU capacity, CMS has incorporated its High Level Trigger farm for offline usage. The trigger farm, a significant resource with a CPU power equivalent to about 50% of the Tier-1 centres capacity, can be used during extended periods with no data taking.

B. Evolution of processing workflows

In Run 1 the multi-core architecture of the processing nodes was not efficiently exploited. Single-core jobs were executed in multi-core CPUs. An ever increasing number of independent and incoherent jobs running on the same physical hardware not sharing resources can significantly affect processing performance. Multi-core aware applications can improve memory sharing and processing performance. A multi-core data processing application using in parallel N cores in the same node significantly consumes less RAM than N independent single-core applications because part of the memory can be shared between the sub-processes or threads [7]. In addition to the gain in memory usage, the load on the workload management system is greatly reduced since the number of multi-core jobs the system has to handle is much smaller. CMS runs in all processing workflows a merging step that produces larger files from the output files of the jobs. Special merge jobs read the unmerged files from mass storage, generate larger files, copy those files to mass storage and delete the unmerged files. This step causes a significant load on the storage system. The merging step is present because it is more efficient to handle large files in data transfers and data storage on tape. Multi-core jobs produce larger output files and therefore the merging step is largely reduced.

Scheduling multi-core jobs poses some challenges though. It is likely that not all workflows can be run in multi-core mode since the software needs to be parallelized. Single and multi-core jobs will have to run on the same resources since partitioning resources according to the job type is not efficient. Therefore, sites will need to efficiently provision multi-core slots at the local batch system, implementing proper scheduling strategies to drain nodes minimizing the idle time. In order to handle single- and multi-core jobs simultaneously, CMS is following the strategy of submitting multi-core pilots. These special jobs request for a number of cores in the same node and once they start running contact a central task queue and pull real jobs to be executed. Multi-core pilots will internally scheduling single- and multi-core jobs trying to minimize scheduling inefficiencies. Finally, multi-core jobs need to efficiently use the allocated cores.

Multi-core scheduling is being commissioned in the CMS workload management system and multi-threaded processing is being developed in the application framework. The deployment will start early 2014 with the aim of running most of the workflows in multi-core mode by the start of Run 2.

The data access model evolution that started in Run 1, where jobs were allowed to access data over the WAN for certain cases, as described in section II, will be extended in Run 2. CMS wants to evolve towards a distributed data federation, a collection of disparate storage resources transparently accessible across a wide area via a common namespace [8]. Jobs running at any of the sites should be able to transparently access data hosted at any site. The aim is to have a more efficient distributed data handling, lower disk storage demands and better use of available CPU resources. An efficient remote I/O over the WAN is key. CMS has invested heavily in I/O optimizations within the application to allow efficient reading of the data over the (long latency) network using the xrootd technology while maintaining a high CPU efficiency.

IV. Conclusions

CMS computing performed extremely well at all levels in Run 1. Excellent networks, flexible and adaptable computing models and software systems paid off in exploiting resources. CMS computing needs to face new challenges for LHC Run 2, where a large increase of computing resources will be required in a context of constrained budgets. Access to opportunistic resources will become important to complement own

resources. CMS needs to use the available resources as fully and efficiently as possible. A major development program is in place to enable a more dynamic data access and distributed parallel computing exploiting the multi-core architecture of the compute nodes and enabling remote data access through the network. The explosive growth in data and highly granular processors in the wider world gives us a powerful ground for success in our evolution path.

REFERENCES

[1] The Large Hadron Collider, http://home.web.cern.ch/about/accelerators/large-hadron-collider
[2] The Worldwide LHC Computing Grid, http://wlcg.web.cern.ch/
[3] The Compact Muon Solenoid experiment, http://cms.web.cern.ch/
[4] http://cernvm.cern.ch/portal/filesystem
[5] CMS conditions data access using FroNTier, B Blumenfeld et al 2008 J. Phys.: Conf. Ser. 119 072007.
[6] Evolution of the Distributed Computing Model of the CMS experiment at the LHC, C. Grandi et al., J.Phys.Conf.Ser. 396 (2012) 032053.
[7] Multi-core processing and scheduling performance in CMS , J.M.Hernández et al., J.Phys.Conf.Ser. 396 (2012) 032055.
[8] Using Xrootd to Federate Regional Storage, L Bauerdick et al., J. Phys.: Conf. Ser. 396 042009.