Journal of Physics: Conference Series **513** (2014) 032039 doi:10.1088/1742-6596/513/3/032039

CMS computing model evolution

C Grandi[1], D Bonacorsi[2], D Colling[3], I Fisk[4] and M Girone[5]

[1]INFN Bologna, [2]Universita di Bologna e INFN, [3]Imperial College London, [4]FNAL, [5]CERN

E-mail: Claudio.Grandi@cern.ch

Abstract. The CMS Computing Model was developed and documented in 2004. Since then the model has evolved to be more flexible and to take advantage of new techniques, but many of the original concepts remain and are in active use. In this presentation we will discuss the changes planned for the restart of the LHC program in 2015. We will discuss the changes planning in the use and definition of the computing tiers that were defined with the MONARC project. We will present how we intend to use new services and infrastructure to provide more efficient and transparent access to the data. We will discuss the computing plans to make better use of the computing capacity by scheduling more of the processor nodes, making better use of the disk storage, and more intelligent use of the networking.

1. Introduction

In this work the modifications to the CMS computing model that are planned for the LHC Long Shutdown 1 (LS1) in 2013-2015 are presented. The original CMS Computing Model is described in [1] and in the Computing Technical Design Report [2] prepared in 2005. Some of the concepts discussed here have already been presented in [3]. We present in Section 2 the constraints to the model coming from the evolution of LHC and the CMS data taking system; in section 3 the modifications to the data model and the processing workflows; in section 4 the modifications to the offline framework relevant to the Computing Model evolution; in section 5 the services on which the Computing Model will be based and in section 6 the distributed computing system.

2. LHC and data taking constraints

The environmental conditions of the LHC during phase-1 (2015-2018) will be significantly different from those of 2012 implying an increased collision rate and complexity of the events. Besides the increase in the centre of mass energy, the intensity of the beams will be significantly higher, increasing, in the assumption of beam spacing at 25 ns, the number of pile-up events to 25 in 2015 and to 40 in 2016 and 2017. As a consequence the time to reconstruct an event will be larger by a factor of 2.5. We believe that the increase of reconstruction time due to the bigger number of out-of-time pile-up events, that would be a factor of 2 with current reconstruction code, will be re-absorbed by improvements in the code itself. In order to maintain the physics capability in all channels at the level of 2012 the trigger rate will need to be 0.8-1.2 kHz, implying a factor 2.5 more recorded events. In total, assuming no changes in the computing system, the increase in CPU resource need would be of a factor of 6.

In order to keep the resource request to a level compatible with flat funding, a number of changes will be implemented in the CMS Computing Model. The main points that will be addressed are:

- Reduction of persistent data formats;
- Reduction of processing needs;
- Reduction of data replicas;
- Exploitation of new hardware platforms;
- Flexible use of allocated resources;
- Exploitation of opportunistic resources;

3. Data model and workflow evolution

The CMS Data Model is schematically shown in figure 1. An important change that has been proposed is to make the ESD (Event Summary Data also known as RECO) transient for most datasets since the majority of the analysis is done from AOD (Analysis Object Data). The primary use of the larger ESD format is detector commissioning, which can be done with a transient disk-only copy with a limited life expectancy of 6 months to a year. This would save a significant amount of storage space.

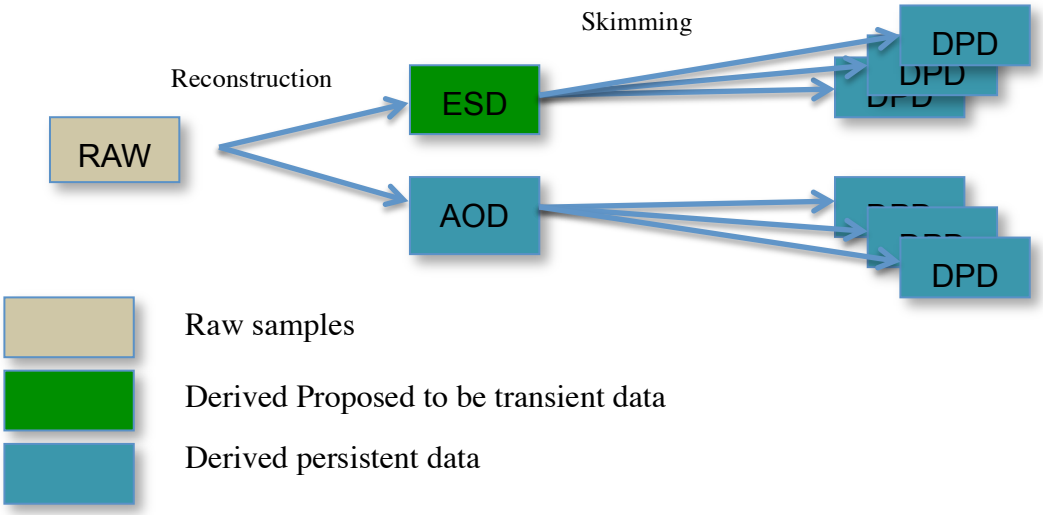

Figure 1. Schematic view of the CMS Data Model.

In 2012 there were 2.5 re-reconstruction passes. In Run2 there is only one envisaged full data re-processing from RAW per year on average at the end of the running year. In addition, targeted reprocessing passes of individual primary datasets are expected throughout the year, but in total adding to a fraction of the full reprocessing pass.

The RAW and AOD data formats have benefitted from programs to introduce efficient compression. The RAW data has historically been smaller than the original estimates. The AOD is compressed and an evaluation is ongoing to assess which objects are not heavily used and could be dropped. The target for reduction is 30%. At the same time CMS is also studying adding elements to the AOD to allow user level recalculation of particle flow elements. This would increase the size of the AOD by a similar factor of 30%, but would potentially eliminate some targeted centralized reprocessing passes. The overall estimated effect is shown in figure 3 (a).

The CMS simulation flow is schematically shown in figure 2. The kinematic step is either performed at run time directly or performed in advance with a dedicated program and written in specialized files (known in CMS as LHE). These events are then simulated to get the detector response. The underlying physics events are combined with in time pile-up from the same collision and out of time pile-up from a window of previous and following collisions. For Run2 this step can involve combining order of a thousand minimum bias pile-up events and requires enormous IO to maintain high CPU efficiency. To mitigate this CMS is preparing complete crossings in advance,

including all the minimum bias events needed, that will be reused with difference physics events. The complete event is reconstructed and AOD and ESD formats can be written.

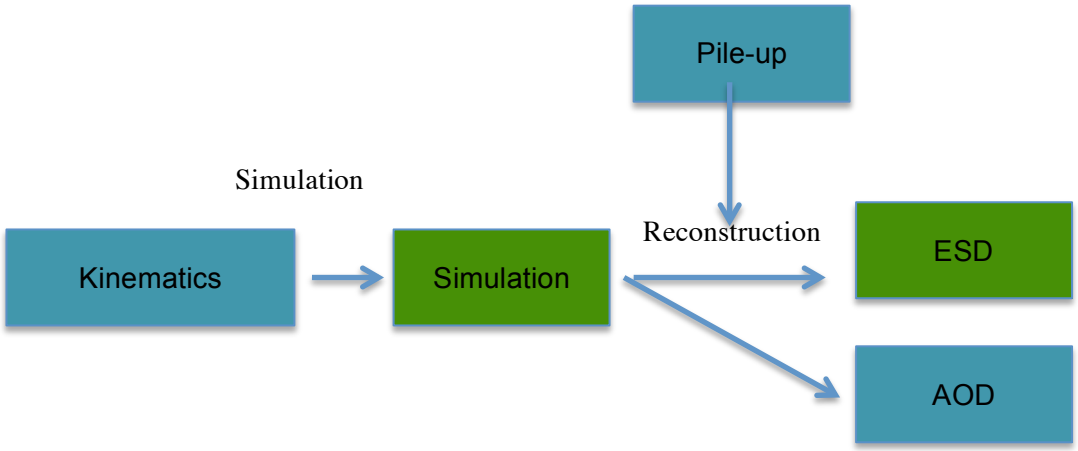

Figure 2. Schematic view of the CMS Simulation flow.

The amount of simulated data will be gradually reduced from 2 times the number of data events in 2012 to 1.5 times in 2015, 1.3 times in 2016 and 1.0 in 2017.

4. Performances of the offline code

CMS worked thoroughly on the performances of the reconstruction program (CMSSW) during Run1, reducing the per-event processing time by one order of magnitude for high luminosity events. The major improvements we expect during LS1 are related to the treatment of the pile-up. Figure 3 (b) shows that the CMSSW pre-releases (postLS1) save up to 30% in the processing time with respect to current versions due to a better treatment of the pile-up, including the out-of-time pile-up.

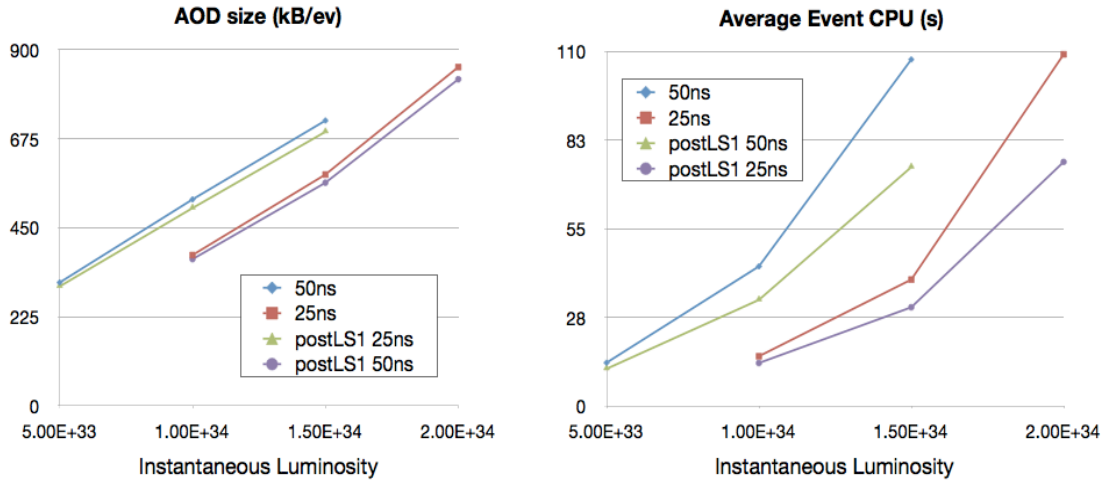

Figure 3. Size of the AOD (a) and processing time per event (b) in the 25 ns and 50 ns scenarios in Run1 and Run2.

CMS has invested heavily in I/O optimizations within the application to allow efficient reading of the data over the network. Over local I/O protocols the application does not need to stage the input file to the worker node disk. This work has also allowed reading the data over the wide area network using remote data access mechanisms while maintaining a high CPU efficiency.

The technology is evolving in the direction of an increased number of cores per machine. The new multi-threaded version of CMSSW will be available in fall 2013. In first approximation using a multi-threaded program will not improve the overall throughput, but will instead mostly stabilize the number of running jobs in the system, stabilize the number of file and squid connections and reduce the number of merge jobs needed. In addition the needed memory per core will decrease significantly making the requirements on the resources easier to fulfil.

In addition to moving to a multi-threaded software framework, CMS sees a benefit in increasing the number of architectures it is able to use. The commissioning period and Run1 were something of an artificial period with homogeneity centred on Scientific Linux (Enterprise Linux). The reconstruction program should be able to run on heterogeneous resources, e.g. ARM 64bit micro-server farms, Xeon Phi's or other X86 variants, GPGPU's, PowerPC farms (Blue Gene machines).

5. CMS services

5.1. Data management

CMS implemented a full mesh for data transfers among sites since years. With the implementation of a storage federation based on xrootd [4], CMS is now in the position to be able to relax the strong requirement of data locality that up to now obliged jobs to run where the data was. Currently the majority of the CMS sites provide access via xrootd to their storage and all of the CMS sites have a fall-back mechanism implemented so that if a file is not found locally it is accessed remotely. Besides for fall-back, remote data access has been used for remote data access on diskless Tier-3s, on public clouds where the storage was very expensive, on private clouds and on opportunistic resources. A set of geographically distributed xrootd redirectors properly interconnected allows optimizing the process of remote data access.

Currently the CMS storage federation is based on xrootd but it is not excluded that a move to an http-based federation is possible once the functionalities and the performances will be adequate.

Even though storage federations add an enormous flexibility to the data access model, the primary method to optimize data access will continue to be an appropriate data distribution via PhEDEx [5] and data-aware job scheduling. To help with data placement CMS is using a Data Popularity System [6] that provides information on the access patterns and is capable of deleting unused samples thus optimizing the use of storage resources.

PhEDEx is being modified in order to split the management of storage and transfers. This will allow using storage at unmanaged sites, thus adding the possibility to use opportunistic storage that can be made available to CMS for limited amounts of time.

5.2. Non-event data management

CMS expects only a few significant changes in the non-event data moving in 2015. Oracle will continue to be used for the master copy of conditions, but conditions will be served to applications through the use of Frontier [7] caches both on the online and offline farms. The current hierarchical structure of failover has served the experiment well.

No changes are expected in the logical to physical file name resolution. The Trivial File Catalogue (TFC) has good performance and scalability, and has facilitated the transition to storage federations for wide area access to data. No changes are foreseen for the second run.

The one significant change proposed is the full adoption of CVMFS [8] for software distribution. By the spring of 2014 CMS will expect that all sites use CVMFS to distribute the common software. Three modes of operations will be supported: native CVMFS clients; CVMFS over NFS; CVMFS in opportunistic mode through the use of Parrot. The third option can be used to enable CVMFS from a non-privileged application. The service is brought up dynamically in user space and can be used also on opportunistic resources.

5.3. Workload management

The transition of CMS to a pilot-based framework is almost completed. The separation of resource allocation and job management provided by the pilot model allows to easily adapting the CMS workload management system to mechanisms of resource allocation different from Grid, such as Clouds or even opportunistic resources via a gateway called BOSCO [9] that functionally replaces a Grid Computing Element and is accessible through the glidein-WMS [10].

Dynamic allocation of resources via the glidein-WMS using the EC2 interface on Amazon and OpenStack [11] based Clouds has been proved and used for real CMS work on the CMS *High Level Trigger Cloud* (see next section), on the CERN *Agile Infrastructure* (now *CERN Private Cloud*) and on test beds in UK and Italy, using the standard WMAgent [12] and CRAB2 [13] tools to manage jobs.

5.4. Security

Authentication and authorization continues to be based on X.509 certificates and VOMS [14] attribute certificates. CMS is open to evolutions going in the direction of identity federations with the aim of making easier for users to have access to resources and for CMS to be able to exploit opportunistic resources.

6. The CMS distributed computing system

6.1. Data flows

The main CMS data-flows are the following:

- The data is selected and reformatted into RAW data files using the Tier-0 resources at CERN. A copy of the data is archived at CERN, a second copy is sent to archive services away from CERN. A second copy is also transferred to disk resources at Tier-1 centres, either from CERN or from the remote archives based on PhEDEx routing decisions.
- The events in the ESD format are written to disk-based resources and maintained for several months for detector commissioning and specific analyses. The AOD formats are written to disk and sent to an archive service for permanent storage.
- AOD and ESD are served out to Tier-2 centres to provide more access and opportunities for processing. The replication and retention are based on the popularity of the dataset.
- Simulation samples from Tier-2s are continuously transferred to the Tier-1s to disk storage. Once the final processing steps are complete and the data is validated it is replicated to the archive.
- Group n-tuples are not normally replicated by CMS provided services unless the samples have to be elevated to official datasets through a validation and publication procedure.

6.2. Role of the computing centres

In Run2 the constraints in the definition of the roles of the different computing tiers will be greatly reduced. The Tier-0 facility is not expected to change in capability or expectation. To compensate the increased processing needs, Tier-1s will perform a significant fraction of the prompt reconstruction. Thanks to the separation of disk and tape storage resources, the Tier-1 centres will look much more like other sites. In addition to simulation production, data and simulation reprocessing and the above-mentioned fraction of prompt reconstruction, they will also participate in user analysis. Tier-2 sites will be involved in MC reconstruction, while continuing a central role in simulation production and user analysis.

Tier-1 sites in CMS will be classified according to their level of support and their operation of central services. Tier-1 sites are expected to operate central services like CVMFS Stratum-1 services and Frontier squids, FTS, and potentially pilot factories. Tier-1s are expected to operate 24/7 and respond to issues out of normal business hours. Tier-1 centres may be collocated with the archival services, but may not be. The archival services are tape-based systems located at a sufficient number

of locations to give 2 copies of the RAW data in separate facilities. We do not want to remove existing high availability centres, but new Tier-1s may be brought up with a higher share of CPU and disk and compensated with archival resources elsewhere. Tier-2 centres will have similar capabilities to Tier-1 centres, but are not expected to run services for other sites and are expected to operate with business hour support.

A significant addition for Run2 is the role of the Higher Level Trigger (HLT) farm. During 2013 CMS commissioned the HLT for use in organized reprocessing. The HLT farm has been demonstrated at 6000 cores running data reprocessing reading the samples over Xrootd from EOS at the CERN computing centre. The cores are made available through the OpenStack cloud interface provisioned through the glide-in WMS pilot system. From the production perspective the HLT is simply another large site, though no traditional Grid services are offered and the pilot system handles resource provisioning of virtual machines. Currently this farm in roughly a third of total Tier-1 processing capacity, and is expected to grow with farm upgrades needed for Run2. In order to fully process one year's data in Run2 the HLT will play an important role by allowing CMS to process the data for a year in 3 months during the regular winter shutdown. The HLT is only fully available for offline processing during shutdown periods, so CMS will need to be prepared when the farm is ready with new releases and new calibrations.

6.3. CMS replication policy

By the restart in 2015 CMS expects to automatically replicate the data produced and release caching space based on the level of access. The goal is that if a user requests any sample produced in the current run the system will be able to provide access to that dataset and there will never be an error thrown that the sample is not available on a site the user has access to. In order to meet this goal CMS will rely on automatic replication and transfer approval to a large pool of centrally controlled disk space at Tier-2s and the full disk at Tier-1 centres and on remote data access on the storage federation. The goal in CMS is that the vast majority of data will continue to be served from local copies but that during the completion of transfers, or to serve unpopular samples, data can be streamed to a running application.

6.4. CMS workflow management

CMS expects several improvements and consolidations in the computing services during 2013 and 2014 in preparation for the next run. The largest improvements expected are in the transition to multi-core workflows and the use of opportunistic computing and additional architectures. We expect to take advantage of the pilot system to facilitate the transition to multi-core scheduling. If the pilot system is given groups of processors or entire nodes, the system can schedule the resources as if they were batch nodes. It will take a mix of multi-core and single core jobs to make the most efficient use of the available resources. The plan is to go to multi-core scheduling only by March 2014.

Analysis will probably always be single core because it is not possible to trust the users to properly manage multiple threads. It is therefore important to move as many analysis activities as possible to organized production (e.g. group n-tuple productions). It is therefore proposed to treat the group production as just another step in the regular production or to develop something similar to the ALICE analysis train model where organized processing passes executing user code are launched with a predictable schedule. By 2015, CMS will have one of the two in operations, making more predictable use of the computing for production of analysis group data formats.

The duration of the allocation of the computing resource, being it the Grid batch slot or the lifetime of a Cloud virtual machine, needs to be increased when going to multi-core scheduling. We currently estimate that we would be losing 50% efficiency by scheduling single core jobs when the pilot lifetime is 48h, but this would go down to ~20% if the pilot lifetime went to 72 hours. These changes require negotiation with the sites. Another more ambitious possibility is extending the lease for a pilot. This is not possible with current batch systems but can be considered if abandoning them (e.g. on clouds). In any case we would benefit from knowing the duration of the resource allocation.

The use of services such as CVMFS, xrootd and BOSCO makes the system much more flexible and integration of opportunistic resources in the CMS computing system possible. The efficient use of opportunistic computing is an important goal for CMS because it can provide significant increases in capacity with low cost. CMS had good experience in the spring of 2013 with the use of the San Diego Supercomputer Centre for a month, where 8000 cores were provided to process a parked dataset, representing an increase of 40% of the processing capacity. One difficult aspect of using opportunistic resources is that CMS should be able to cope with pre-emption. Initially treat it as a failure and measure the impact on operations. In the long term CMS may develop check-pointing techniques. Note that this would not be full check-pointing, rather something to be implemented inside the application.

Another significant addition to CMS distributed computing resources is the ability to provision clouds. The work done to utilize the HLT resources and the CERN Agile Infrastructure is directly applicable to other resources provisioned with OpenStack. We expect that these techniques will be used to connect to academic cloud resources, other contributed clouds, and potentially even commercial clouds if the cost models become more competitive.

We will also consider volunteer computing. Small scale computing provided to use resources that would otherwise be lost. As an example, BOINC (Berkeley Open Infrastructure for Network Computing) [15] is being integrated into condor. This type of computing is probably only useful for specific workflows, but it has the potential for being large and free except for the obvious development and operations costs.

References

[1] Grandi C, Stickland D, Taylor L *et al*. 2004 The CMS Computing Model *Preprint* CERN-LHCC-2004-035/G-083
[2] The CMS Collaboration 2005 CMS Computing Technical Design Report *Preprint* CERN-LHCC-2005-023
[3] Grandi C *et al*. 2012 Evolution of the Distributed Computing Model of the CMS experiment at the LHC *J. Phys.: Conf. Ser*. **396** 032053
[4] http://xrootd.slac.stanford.edu/
[5] Wildish T *et al*. 2012 From toolkit to framework - the past and future evolution of PhEDEx *J. Phys.: Conf. Ser*. **396** 032118
[6] Giordano D *et al*. 2012 Implementing data placement strategies for the CMS experiment based on a popularity model *J. Phys.: Conf. Ser*. **396** 032047
[7] Dykstra D Lueking L 2010 Greatly improved cache update times for conditions data with Frontier/Squid *J. Phys.: Conf. Ser*. **219** 072034
[8] Blomer J Fuhrmann T 2010 A Fully Decentralized File System Cache for the CernVM-FS *19th IEEE International Conference on Computer Communications and Networks* doi: 10.1109/ICCCN.2010.5560054
[9] D J Weitzel *et al*. 2013 Accessing opportunistic resources with Bosco *Int. Conf. on Computing in High Energy and Nuclear Physics* Amsterdam
[10] Sfiligoi I *et al*. 2009 The Pilot Way to Grid Resources Using glideinWMS *WRI World Congress on Computer Science and Information Engineering* ISBN: 978-0-7695-3507-4
[11] http://www.openstack.org/
[12] Fajardo E *et al*. 2012 A new era for central processing and production in CMS *J. Phys.: Conf. Ser*. **396** 042018
[13] Codispoti G *et al*. 2009 CRAB: A CMS Application for Distributed Analysis *IEEE Trans Nucl Sci 5* **56** 2850-2858
[14] Ceccanti A *et al*. 2010 VOMS/VOMRS utilization patterns and convergence plan *J. Phys.: Conf. Ser*. **219** 062006
[15] http://boinc.berkeley.edu/