THE CMS COMPUTING MODEL

J.M. Hernández
CIEMAT, Madrid
On behalf of the CMS Collaboration

Abstract

Since CHEP04 in Interlaken, the CMS experiment has developed a baseline Computing Model and a Technical Design for the computing system it expects to need in the first years of LHC running. Significant attention was focused on the development of a data model with heavy streaming at the level of the RAW data based on trigger physics selections. We expect that this will allow maximum flexibility in the use of distributed computing resources. The CMS distributed Computing Model includes a Tier-0 centre at CERN, a CMS Analysis Facility at CERN, several Tier-1 centres located at large regional computing centres, as well as many Tier-2 centres. The workflows involving these centres have been identified, along with baseline architectures for the data management. This paper will describe the computing and data model, give an overview of the technical design and describe the current status of the CMS computing system.

COMPUTING ARQUITECTURE

CMS has adopted a distributed computing model in order to cope with the requirements for storage, processing and analysis of the huge amount of data the experiment will provide. Tens of thousands of today's PCs and Petabytes of disk and tape storage will be needed. In the CMS computing model, released last year [1], resources are geographically distributed, interconnected via high throughput networks and operated by means of Grid software.

The computing resources are structured in a tiered architecture (see fig. 1) with specific functionality at different levels. CERN, where data will be taken and where the first processing and storage of the data will take place, constitutes the so-called Tier-0 centre. In addition to the Tier-0 centre, CERN will host the CMS Analysis Facility (CAF). The CAF will have access to the full raw dataset and will be focused on latency-critical deterctor, trigger and calibration activities. It will also provide some CMS central services like the storage of conditions data and calibrations. Reconstructed data at the Tier-0 together with the corresponding raw data will be distributed to the next level in the tiered structure, a small number of Tier-1 centres (around 7) where organized mass data processing will be performed. That includes calibration, re-processing, data skimming and other organized intensive analysis tasks. The Tier-1 centres will archive the fraction of data distributed to them as well as the simulated data produced at the Tier-2 centres. In these latter sites, in addition to the production of simulated data, user data analysis of data imported from Tier-1 centres will take place.

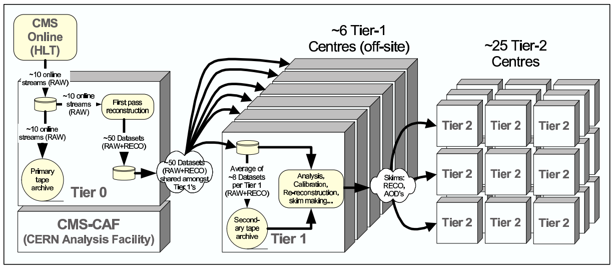

Figure 1: CMS computing architecture

In the CMS computing model, site activities and functionality are largely predictable. Activities are driven by data location, organized mass data processing and custudial storage is performed at the Tier-1 centres while 'chaotic' computing is essentially restricted to end-user data analysis at the Tier-2 sites.

Figure 2 shows the profile expected for the evolution of the computing resources at the different Tier levels. Already for 2008 a significant amount of resources will be required.

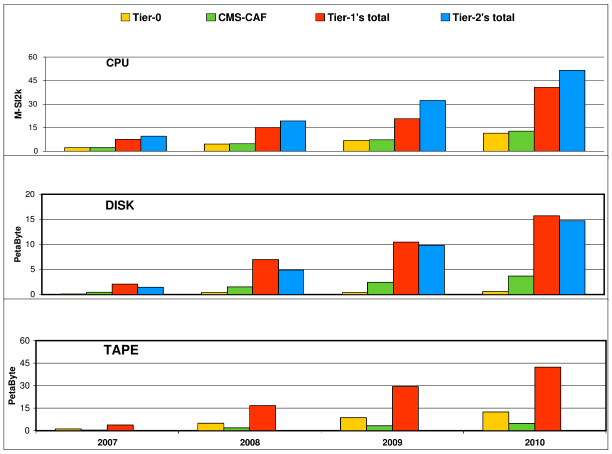

Figure 2: Profile for the ramp up of CMS computing resources

In the computing model, prioritization is one of the guiding principles. At the early stage, in 2007/8, the computing system efficiency might not be 100% so that flexibility to set priorities will be important. Examples are coping with potential reconstruction backlogs without delaying the processing of critical data, having the possibility of performing prompt calibration using low-latency data, running prompt analyses, flexibility in the data distribution, etc.

Another important concept is data streaming. Classifying events early allows prioritization and data access optimization. We forsee to have O(10) online data streams coming from the High Level Trigger (HLT) farm into the reconstruction farm. For example a express stream of hot/calibration events. After the reconstruction, those raw data streams will be split into O(50) trigger-determined primary datasets that will be distributed to the Tier-1 centres for analysis skimming. The analysis datasets resulting from the data skimming will be transported to the Tier-2 centers for end-user data analysis.

Event Model

CMS will use a number of event data formats with varying degrees of detail, size and refinement. Starting from the raw data produced in the online system, successive degrees of processing refine this data, apply calibrations and create high level physics objects.

The 'RAW' data will be created by the HLT system containing the detector data together with the trigger results and high-level trigger objects created during the HLT processing. The raw event data size for 2008 is expected to be about 1.5 MB including compression, poor understanding of the detector, etc. Given the expected HLT output rate of approximately 150 Hz, the expected raw data volume amounts to 4.5 PB/year including two copies, one at the Tier-0 and the other distributed among the Tier-1 centres.

After reconstruction, the 'RECO' data format will contain reconstructed objects with their associated hits. The reconstructed event size is expected to be around 250 kB, yielding a total reconstructed data volume of about 2 PB/year, including 3 reprocessings of the data.

A more compact format, the Analysis Object Data (AOD), will be produced from the RECO event. This will be the main analysis format containing physics objects. The expected AOD event size is 50 kB producing a total data volume of about 2.6 PB/year. Each Tier-1 centre will host a whole copy of the AOD data and the Tier-2 centres will import a large fraction as the result of organized skimmings run at the Tier-1 sites.

Data Flows

Figure 3 shows the expected average data flows at the Tier-0 and at a nominal Tier-1 and Tier-2 centres. It should be noted that the numbers correspond to average sustained throughputs. Network bandwidth and processing capacity at the sites should be much larger in order to cope with backlogs and to be able to perform bursty data transfers when required. The CPU, disk, tape capacities and network bandwidths are also included in the picture inside the boxes representing the Tier centres.

The Tier-0 will be accepting 225 MB/s of raw data coming from the CMS detector. After the reconstruction, these data will be archived on tape at CERN and distributed to the Tier-1 sites. Each Tier-1 (7 in total) will receive about 40 MB/s of RAW+RECO+AOD data. A similar amount of Monte Carlo data, 48 MB/s, will arrive from the associated Tier-2 (4 in average) centres. Each Tier-2 will produce an amount of MC data of about 1 TB/day (12 MB/s).

The data skimming process taking place at the Tier-1 centres will generate a traffic from storage to CPU of about 800 MB/s. This organized activity will be triggered by physics groups. The result of the skimming, 240 MB/s, will be transported to the associated Tier-2's (60 MB/s or 5 TB/day in average to each Tier-2). The processing of the skimmed analysis datasets at each Tier-2 centre will generate a data throughput from storage to CPU up to 1 GB/s.

During data reprocessing, each Tier-1 site will re-reconstruct its share of raw data archived locally. The resulting AODs will have to be distributed to all Tier-1 centres.

COMPUTING SERVICES AND WORKFLOWS

The Workload and Data Management Systems have been designed following the philosophy of using existing Grid Services as much as possible, building on top of them CMS-specific services. We intend to deliver a working baseline system with minimal functionality by the time the first experiment data are taken. The driving principles for the baseline system are: i) Keep the system as simple as possible. ii) Optimize for the common case: optimize for read access (most data is write-once, read-many) and for organized bulk processing. iii) Decouple parts of the system: minimize job dependencies, site-local information should remain local. iv) Use explicit data placement: data does not move around in response to job submission but data is placed at a site through explicit CMS policy. v) Grid interoperability: different Grid flavours should be supported and be transparently interoperable.

Figure 4 shows the CMS services of the present Data and Workload Management systems. The Data Management System has been designed with no global file replica catalogue. Instead, the system has a global Data Bookkeeping System to track what data exist, a distributed Data Location System to track where data are located and a local File Catalogue at each site to provide Physical File Names to the data processing jobs. The data management services and catalogues are currently being refactored in order to track and replicate data with a granularity of file blocks, logical groups of files. Data storage management is done through the Storage Resource Manager (SRM). Files are read from the storage system using the file access protocols RFIO (for the CASTOR storage system) and DCAP (for the dCache storage system).

The CMS Workload Management System (WMS) relies on the Grid WMS provided by the Worldwide LHC Computing Grid project for job submision and scheduling onto resources according to the CMS Virtual Organization (VO) policy and priorities. CMS has built on top of the Grid

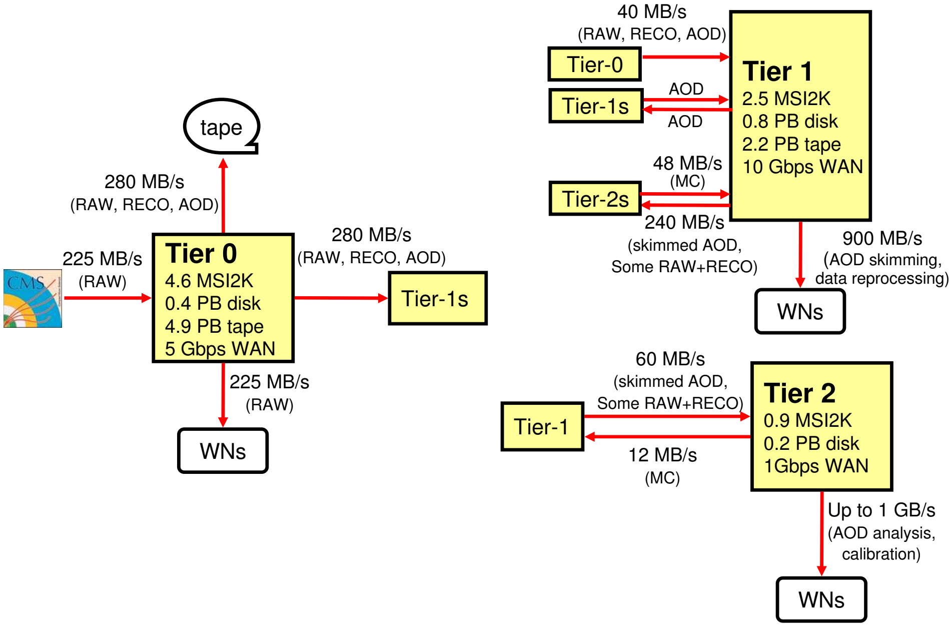

Figure 3: CMS data flows

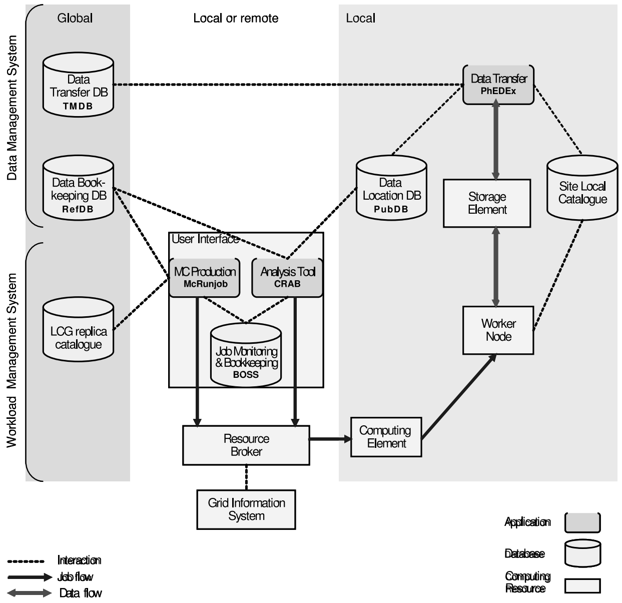

Figure 4: CMS computing services in the current Data and Workflow Management Systems

WMS services analysis (CRAB [3]) and Monte Carlo Production (McRunjob [4]) job submission services as well as a monitoring and bookkeeping sytem (BOSS [5]). Jobs are submitted to a Grid Resource Broker (RB) from an User Interface (UI) machine. The RB using the Grid Information System knows the available resources and their usage. It performs matchmaking to determine the best site to run the job and submits it to the Computing Element (CE) of the selected site which in turn schedules it in the local batch system. The Worker Node (WN) machines where jobs run have POSIX-IO-like access to the data stored in the local Storage Element (SE).

CURRENT DISTRIBUTED COMPUTING SYSTEMS

The current CMS distributed computing system has been used in the last couple of years for running mainly three production activities: distributed Monte Carlo production on the Grid, data distribution and distributed data analysis.

Distributed Monte Carlo Production

More than 150 million simulated events have been produced for elaborating the CMS Physics Technical Design Report (PTDR) due to middle 2006. About one third of the events have been produced running production on the Grid, with the latest productions almost completely carried out on the Grid. The current CMS MC production system (McRunjob) was originally designed for local farm production where one has local access to the resources and the data. It was later ported to the Grid. However, it suffers from severe inneficiencies due to its monolithic design and the constraints imposed by the current CMS event data model. Both the event data model and the MC production system are being overhauled based on the invaluable experience gained running the old system. The new production system will hopefully gain in automation and efficiency making better use of the resources. It will be better coupled to the data management system, better handle errors and will incorporate new features like job chaining and data merging.

Data Transfer and Placement System

CMS has developed a reliable point-to-point transfer system [2] based on unreliable Grid transfers tools. PhEDEx (Physics Experiment Data Export) is a large scale dataset replica management system which manages data flow following a specified transfer topology (e.g. Tier-0 → Tier-1's ↔ Tier-2's) performing multi-hop routed transfers. PhEDEx is built as a set of quasi-independent, asynchronous software agents running at each transfer node, posting messages into a central blackboard for synchronization. Transfer nodes subscribe for data allocatd in other nodes. PhEDEx enables distribution management at dataset level, implements experiment's policy on data placement and allows prioritization and scheduling. It is in production since almost two years managing reliably and efficiently transfers of tens of Terabytes/day. A total of 150 TB of data are known to PhEDEx, with 350 TB of data replicated in different sites. It is running at CERN, at all the Tier-1's and at most of the Tier-2 centres.

Distributed Data Analysis

Distributed data analysis is done at CMS using the CMS Remote Analysis Builder (CRAB) tool. Intensive data analysis has been carried out during the last couple of years in preparation of the CMS PTDR. Around half a million analysis jobs have been run in the last 8 months, with an average of 60000 jobs/month,analysing the data distributed in all Tier-1 and some Tier-2 centres.

COMPUTING CHALLENGES

CMS has chosen to build its computing system in an iterative way testing prototypes of Grid resources and services of increasing scale and complexity. This way, problems are found and addressed, and missing components are identified. For this purpose CMS undertakes periodic computing challenges to test its computing model and Grid computing systems. These tests have shown that the basic Grid infrastructure and services are in place but their stability and reliability should be greatly improved. In addition, important features like implementation of VO policies and priorities and dynamic behaviour in the WMS and DMS systems like re-scheduling are still missing. Grid services like job monitoring and accounting are still quite primitive and suffer from high latency. Investing efforts in integrating Grid services with sites has been found to be of great importance.

The latest big computing challenge, the LCG Service Challenge, took place during the last half of 2005. It was regarded by CMS as an integration test exercising the bulk data processing part of the CMS computing model under realistic conditions. It focused on validation of data storage, transfer and data serving infrastructure in addition to the required workload components for job submission. In a first phase, the throughput phase on July 2005, CMS distributed 280 TB of data from CERN to the Tier-1 sites reaching aggregate transfer rates of 200 MB/s sustained during days. In the second phase of the challenge, the service phase between September and November 2005, data transfer at a lower rate concurrently run with automatic data publication at the sites and data analysis. About 70000 jobs run at the sites reaching data throughputs of 200 MB/s from storage to CPU at some sites.

CONCLUSIONS

CMS has adopted a distributed computing model which makes use of Grid technologies. Production CMS services on the Grid such as the data transfer and placement system, the Monte Carlo production system and the data anlysis are in place. Scale and complexity are steadily increased. During 2006 major changes in the computing systems will be done, specifically in the data management system, data processing framework and event data model, and Monte Carlo production system. Major computing challenges are scheduled for 2006 as well. By the end of 2006 we hope to have validated our computing model and have commissioned the CMS computing systems to be operational for the first data taking scheduled by the end of 2007.

REFERENCES

[1] The CMS Computing Technical Design Report, CERN-LHCC-2005-023.

[2] J. Rehns et al., PhEDEx high-throughput data transfer management system. These proceedings.

[3] M. Corvo et al., CRAB, a tool to enable CMS Distributed Analysis. These proceedings.

[4] Runjob Project, *http://projects.fnal.gov/runjob*

[5] S. Wakefield et al., BOSS, a tool for job submission and tracking. These proceedings.

[6] P. Garcia-Abia et al., CMS Monte Carlo production on the LHC Computing Grid. These proceedings.