PROCEEDINGS OF SCIENCE

Grid Computing in CMS

José M. Hernández[*][†]
CIEMAT, Spain
E-mail: `jose.hernandez@ciemat.es`

CMS has chosen to adopt a distributed model for all computing in order to cope with the requirements on computing and storage resources needed for the processing and analysis of the huge amount of data the experiment will be providing from LHC startup. An overview of the architecture of the CMS Grid computing system will be given in this paper. The baseline system as well as possible extensions of the baseline capabilities and functionalities will be described. The architecture is based on a set of loosely coupled components that allow an iterative process of developing and integrating new components, increasing the scale and functionality of the system. The evolving computing system is tested in major data and service "challenges". Experience gained in past challenges and expectations from future challenges will be reported. The CMS computing environment is a distributed system of computing services and resources that interact with each other as Grid services. CMS-specific services are built on top of lowerlevel services provided by Grid projects. Emphasis will be placed on describing the CMS data placement and transfer system as well as the distributed Monte Carlo production on the Grid.

International Europhysics Conference on High Energy Physics
July 21st - 27th 2005
Lisboa, Portugal

[*] Speaker.
[†] On behalf of the CMS Collaboration

1. The CMS Computing Model

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 LHC will provide. Tens of thousands of today's PCs and Petabytes of disk and tape storage will be needed. In the CMS computing model, recently released [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. Data will be distributed to the next level, a small number of Tier-1 centres (around 10) where organized data processing will be performed. That includes calibration, re-processing, data skimming and other 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 sites, in addition to the production of simulated data, user data analysis of data imported from Tier-1 centres will take place.

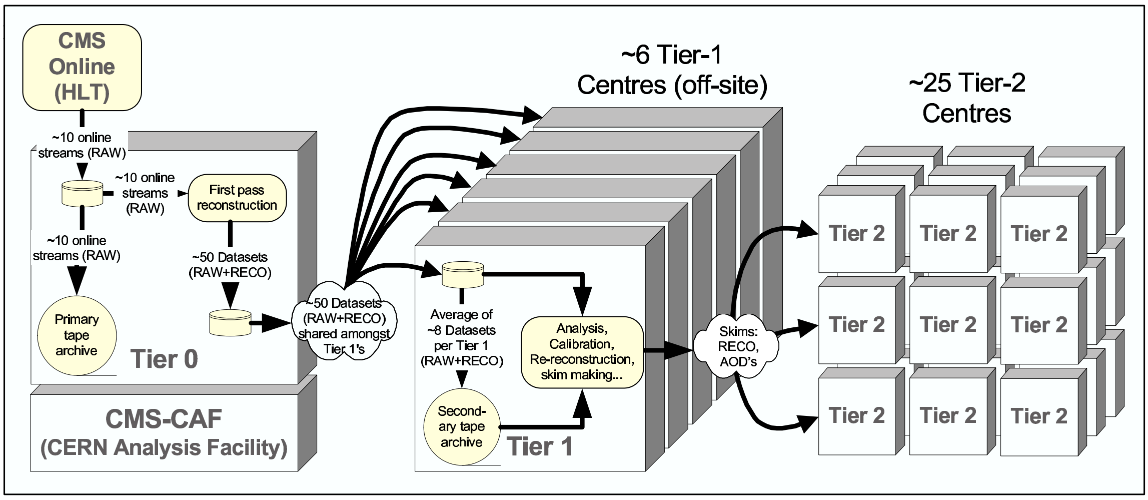

Figure 1: Tier architecture in the CMS Computing Model

2. Workload and Data Management Systems

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.

2.1 Data Management System

The Data Management System has been designed with no global file replica catalogue. Instead, the system has a global Data Bookkeeping System (DBS) to track what data exist, a distributed Data Location System (DLS) to track where data are located and a local File Catalogue at each site to provide Physical File Names to the data processing jobs (see fig 2). Data are tracked and replicated with a granularity of blocks made up of an arbitrary number 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).

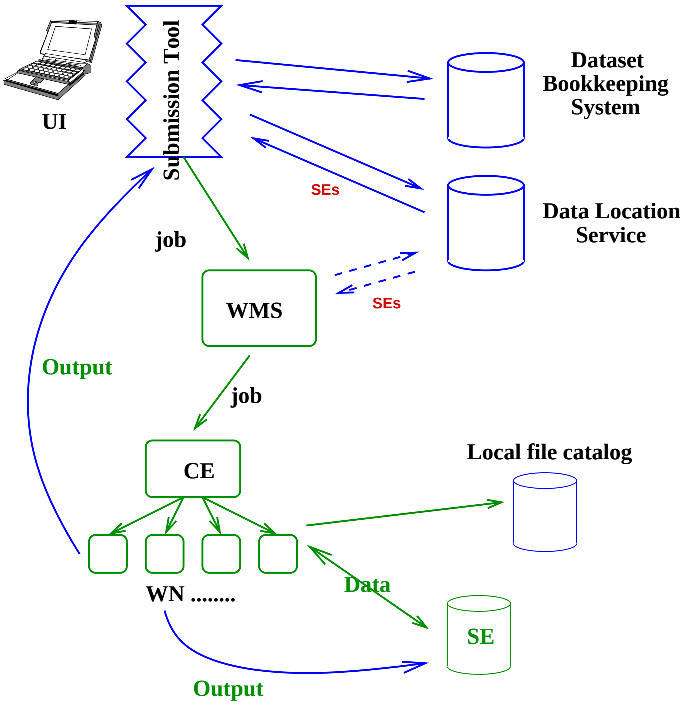

Figure 2: CMS Workload and Data Management Systems

2.1.1 Data Transfer and Placement System

CMS has developed a reliable point-to-point transfer system based on unreliable Grid transfers tools. PhEDEx [2] (Physics Experiment Data Export) is a large scale dataset replica management system which manages data flow following a certain transfer topology (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 in a central blackboard. 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 more than a year managing reliably and efficiently transfers of tens of Terabytes/day. It is running at CERN, at all the Tier-1's and at several Tier-2 centres.

2.2 Workload Management 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 WMS services a job submission (CRAB [3]) and monitoring (BOSS [4]) system. 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 sites where the requested data are located and submits the job to the Computing Element (CE) of the site which in turn schedules it in the local batch system. The Worker Node (WN) machines where jobs run have access to the local Storage Element (SE) where the data are located (see figure 2). Two important work flows will be described in the next two sections: Monte Carlo production and data analysis on the Grid.

2.2.1 Monte Carlo Production on the Grid

CMS has developed the McRunjob tool for running production jobs. McRunjob is highly configurable and flexible and it is interfaced to the different CMS Grid flavours (LHC Computing Grid -LCG- [5] and Open Science Grid -OSG- [6]). Different production steps (generation, simulation, digitization with pile-up mixing and reconstruction) are run separately. Several thousands of CPUs are available in both Grids. Few million events per month are produced with an efficiency ranging 70-90%. The main issue of production on the Grid is reliability and stability. There are frequent problems at the sites (hardware failures, site misconfiguration, etc) and on the Grid services (unreliable data transfer, intermitent problems in the Grid information system and the global Grid catalogue, etc).

2.2.2 Data Analysis on the Grid

Data samples for the CMS Physics Technical Design Report, currently in prepation, were distributed among the Tier-1 centres. In total 80 million events, corresponding to a data volume of around 80 TB) were distributed among 5 Tier-1's. A simple scenario where data are pre-located and Grid jobs are sent to the data is used. A CMS-specific tool (CRAB) is used for job preparation, submission, environment setting, monitoring and output retrieval. Around 300000 jobs have been run, O(10000) jobs been run per week and O(100) users using the distributed data analysis system.

3. Tests and Experience in CMS Grid Computing

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.

4. Summary

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. Basic Grid Infrastructure and Services are in place but reliability and stability should be improved.

References

[1] CMS Collaboration, *The Computing Project Technical*, June 2005.

[2] T. Barrass et al., *Software Agents in Data and Workload Management*, in Proceedings of the *CHEP 04 Conference*. Interlaken, Switzerland, September 27th - October 1st, 2004.

[3] C. Grandi et al., *Object Based System for Batch Job Submission and Monitoring (BOSS)*, CMS NOTE 2003-005 (2003).

[4] *CRAB project Web Site*, located at http://savannah.cern.ch/projects/crab/.

[5] LHC Computing Grid Technical Design Report, *CERN-LHCC-2005-024* Design Report, *CERN-LHCC-2005-023*, June 2005.

[6] *Open Science Grid Web Site*, located at http://www.opensciencegrid.org