CMS GRID COMPUTING AT THE SPANISH TIER-1 AND TIER-2 SITES

P. Garcia-Abia[a], J.M. Hernández[a], F. Martínez[b], G. Merino[b], M. Rodríguez[b], F.J. Rodríguez-Calonge[a]
[a]CIEMAT, Madrid, Spain
[b]PIC, Barcelona, Spain

Abstract

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.

The architecture is based on a tier-organised structure of computing resources, based on a Tier-0 centre at CERN, a small number of Tier-1 centres for mass data processing, and a relatively large number of Tier-2 centres where physics analysis will be performed. The distributed resources are connected using high-speed networks and are operated by means of Grid toolkits and services.

We present in this paper, using the Spanish Tier-1 (PIC) and Tier-2 (federated CIEMAT-IFCA) centres as examples, the organization of the computing resources together with the CMS Grid Services, built on top of generic Grid Services, required to operate the resources and carry out the CMS workflows. The Spanish sites contribute with 5% of the CMS computing resources.

We also present the current Grid-related computing activities performed at the CMS computing sites, like high-throughput and reliable data distribution, distributed Monte Carlo production and distributed data analysis, where the spanish Sites have traditionally played a leading role in development, integration and operation.

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 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. Data will be distributed to the next level, a small number of Tier-1 centres (around 7) where organized 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 associated 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.

THE SPANISH SITES

Spain is providing one of the seven Tier-1 centers as well as one of the about 25 Tier-2 sites. The Tier-1 center is based at the *Puerto de Información Científica* (PIC), Barcelona. The Tier-2 is a federated center combining in a transparent way computing resources located at CIEMAT, Madrid, and IFCA, Santander.

Spanish sites contribute with about 5% of the computing resources required by CMS. Tables 1 and 2 show the ramp-up of CPU, storage and network resources forseen at the Tier-1 and Tier-2 centers up to the end of 2007.

Table 1: PIC contribution to CMS hardware resources

	End 2005	End 2006	End 2007
CPU (kSI2k)	100	300	800
Disk (TB)	50	200	400
Tape (TB)	100	400	800
WAN (Gbps)	1	10	10

Table 2: T2-Spain contribution to CMS hardware resources

	End 2005	End 2006	End 2007
CPU (kSI2k)	150	400	800
Disk (TB)	20	100	200
Tape (TB)	-	-	-
WAN (Gbps)	1	1	10

CMS 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

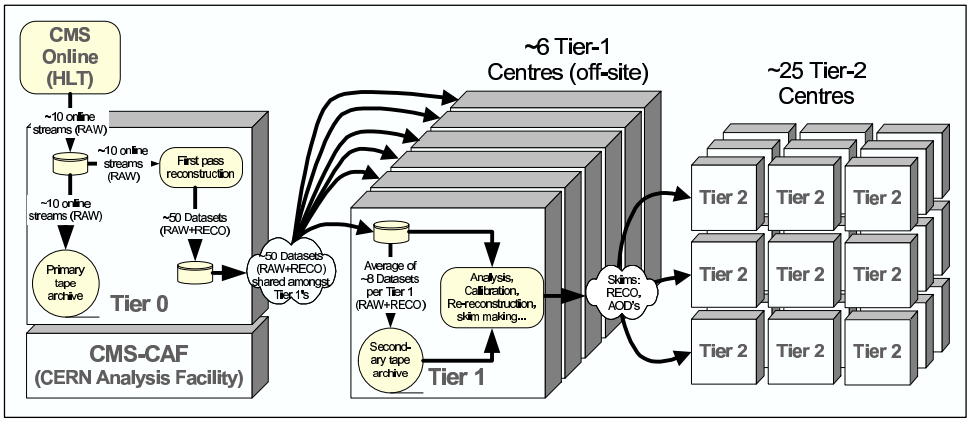

Figure 1: CMS computing architecture

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 transparently interoperable.

Figure 2 shows the CMS services of the 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 (RefDB [2]) to track what data exist, a distributed Data Location System (PubDB [3]) to track where data are located and a local File Catalogue at each site to provide Physical File Names to the data processing jobs. Data storage management is done through the Storage Resource Manager (SRM [4]). Files are read from the storage system using the file access protocols RFIO (for the CASTOR [5] storage system) and DCAP (for the dCache [6] storage system).

CMS has developed a reliable point-to-point transfer system [7] 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 $\rightarrow$ Tier-1's $\leftrightarrow$ 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 most of the Tier-2 centres.

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 analysis (CRAB [8]) and Monte Carlo (MC) Production (McRunjob [9]) job submission services as well as a monitoring and bookkeeping system (BOSS [10]). 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.

COMPUTING GRID ACTIVITIES

Spanish sites have traditionally played a leading role in CMS computing activities. They are involved at all levels, participating in development, integration and operations.

Development Activities

The CMS MC production tool has been ported to LCG [11] as a contribution to the development activities of the Spanish Sites. An end-to-end implementation was done, from the generation of events to the publication of data for analysis, through all intermediate steps. A complete performance analysis has been performed extracting important conclusions that will serve as guidelines for the new MC production system being currently developed. The new production system will 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.

Production Activities

At present, we can distinguish three main areas of production activities in CMS computing: data distribution, distributed data analysis and distributed Monte Carlo Production.

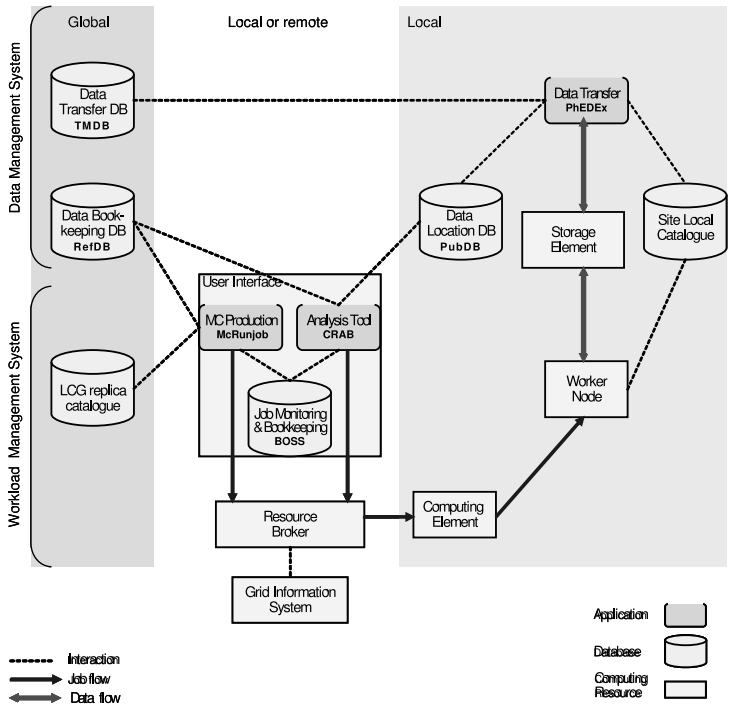

Figure 2: CMS computing services

The CMS data transfer and placement system, PhEDEx, is fully installed at the Spanish sites. Sustained transfer rates of about 3 TB/day are routinely achieved.

Spanish sites host about 15% of the CMS Monte Carlo data being used for physics analyses carried out in preparation of the CMS Physics Technical Design Report. The data volume amounts 15 TB corresponding to 15 million events. All services required for supporting distributed data analysis are in place. Of the order of 10000 analysis jobs have run so far at the Spanish sites.

Spanish sites are playing a leading role in Monte Carlo production on the LHC Computing Grid (LCG). As a contribution to this area, a large fraction of the MC production on LCG operations is being carried out by a Spanish team. Around 15 million events have been produced so far on LCG during the past year, representing about 10% of the total MC production in CMS. Most of the latest MC production is now being done on the Grid.

Integration Activities

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.

The last computing challenge undertaken by CMS was the LCG Service Challenge 3 (SC3). Is was regarded by CMS as a computing integration test exercising the bulk data processing part of the CMS computing model under realistic conditions. It focused on validating the data storage, transfer and data serving infrastructure together with the required workload components for job submission. During a throughput phase on July 2005, 280 TB were transferred from CERN to the 7 CMS Tier-1 centers, reaching an aggregate throughput of 200 MB/s sustained for days. The Spanish Tier-1, PIC, sustained a throughput of 50 MB/s (about 4 TB/day) for two weeks (see figure 3). During the SC3 service phase, between September and November 2005, the data transfer activity at a reduced rate was run concurrently with automatic data publishing at the sites and analysis job execution at the Tier-1 and Tier-2 sites. About 70000 jobs were run in total, 10000 of which run at the Spanish sites (see figure 4). PIC reached an aggregate data throughput of about 200 MB/s from storage to CPU and the Spanish Tier-2 attained 100 MB/s, surpassing CMS expectations for this challenge.

Computing challenges 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 at this stage.

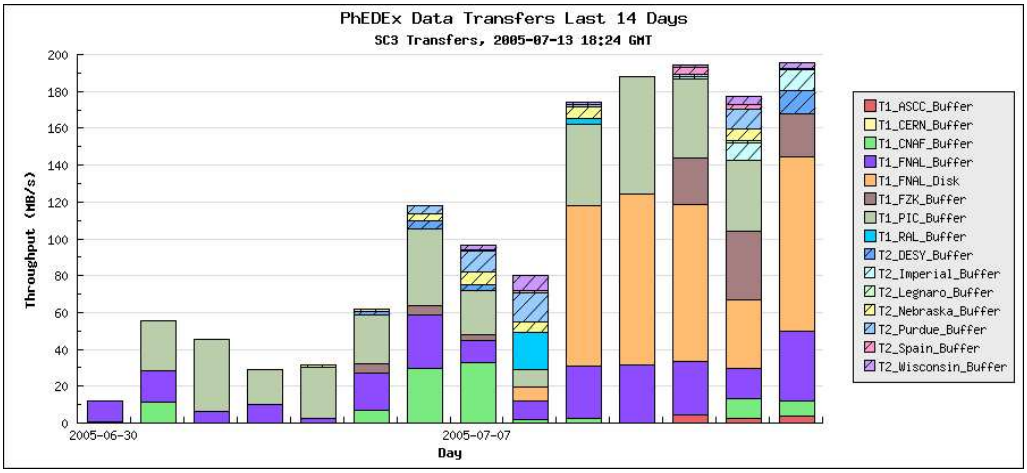

Figure 3: Transfer data throughput during SC3 throughput phase. PIC sustained a transfer rate of about 50 MB/s for almost two weeks.

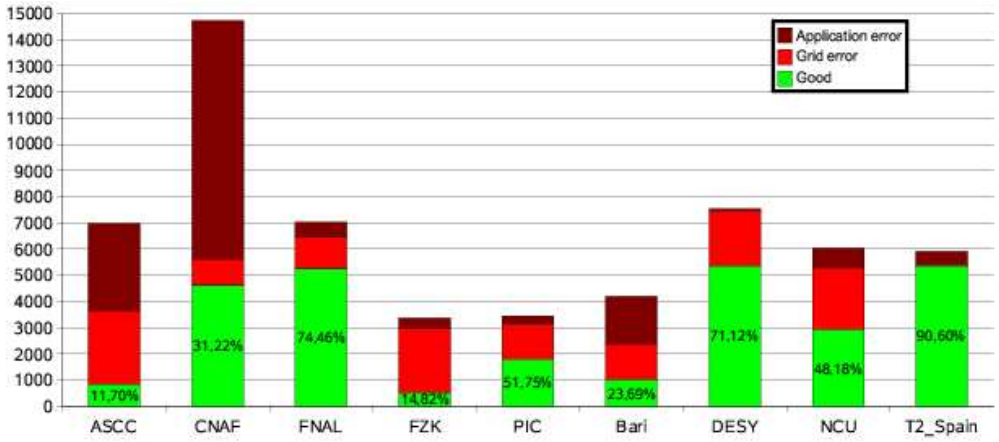

Figure 4: Data analysis jobs run during the SC3 service phase. The Spanish Tier-2 showed a remarkable reliability compared with other sites.

REFERENCES

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

[2] *http://cmsdoc.cern.ch/cms/cpt/Computing/Technical/subproj/RefDB.html*.

[3] *http://cmsdoc.cern.ch/cms/cpt/Computing/Technical/subproj/PubDB.html*.

[4] *http://sdm.lbl.gov/srm-wg/*

[5] *http://castor.web.cern.ch/castor/*

[6] *http://www.dcache.org*

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

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

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

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

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