CMS Conference Report

6 March 2004

CMS Distributed Data Analysis Challenges

C. Grandi

INFN Bologna, Viale Berti-Pichat 6/2, 40127 Bologna, Italy

on behalf of the CMS collaboration

Abstract

In Spring 2004 CMS will undertake a 100 TeraByte-scale Data Challenge (DC04) as part of a series of challenges in preparation for running at CERN's Large Hadron Collider. During 1 month, DC04 must demonstrate the ability of the computing and software to cope with a sustained event data-taking rate of 25 Hz, for a total of 50 million events. The emphasis of DC04 is on the validation of the first pass reconstruction and storage systems at CERN and the streaming of events to a distributed system of Tier-1, and Tier-2 sites worldwide where typical analysis tasks will be performed. It is expected that the LHC Computing Grid project will provide a set of grid services suitable for use in a real production environment, as part of this data challenge. The results of this challenge will be used to define the CMS software and computing systems in their Technical Design Report.

Presented at *IX International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT)*, High Energy Accelerator Research Organization (KEK), Japan, December 1-5, 2003

1 Introduction

The CMS Computing data challenges are constructed to prepare for LHC running and include the definition of the computing infrastructure, the definition and set-up of analysis infrastructure, the validation of computing model. By design they entail each year a factor 2 increase in complexity over the previous year, leading to a full scale test in 2006.

Even though their primary goal is to gradually build the CMS computing system in time for the start of LHC, they are tightly linked to other CMS activities and provide computing support for production and analysis of the simulated data needed for studies on detector, trigger and DAQ design and validation, and for physics system setup.

CMS data challenges are distributed over many sites in the collaborating countries.

The main CMS Data Challenges are:

- Spring 2002 Data Production for DAQ TDR (Technical Design Report)
- 2003 Data Production (PCP04) and 2004 Data Challenge (DC04) for Computing & Core Software TDR
- 2004/05 Data Production (DC05) for Physics TDR. Includes data analysis for physics TDR
- 2005 Data Production (PCP06) and 2006 Data Challenge (DC06) for the *Readiness Review*
- In 2006, until the beginning of 2007, there will be the commissioning for LHC startup.

After a brief description of the CMS Computing Environment and the collection of tools for managing the CMS Production System, OCTOPUS, we will describe the past data challenges, with some emphasis on the prototypes that used grid tools. We will then describe the activities related to the coming Data Challenge 2004.

2 CMS Computing Environment

LHC will produce 40 million bunch crossing per second in the CMS detector, which correspond to a data rate of about 1000 TB/s.

The on-line system will reduce the rate to 100 events per second, which correspond to a data rate of about 100 MB/s. These data are saved on permanent store as *raw data* and are used as input for the off-line processing. The on-line system has two components:

- the *Level-1 trigger* which is implemented as ad-hoc designed hardware;
- the High Level Trigger which is implemented as software running on a computing (on-line) farm.

The raw data event size will approximately be 1 MB and will be archived on persistent storage (~ 1 PB/year). Raw data will be reconstructed and *Data Summary Tapes* (DST) will be produced. A reconstructed event size is approximately 0.5 MB. A further reconstruction step will produce *Analysis Object Data* (AOD) which will be about 20 KB per event.

The DST and AOD data (and part of raw data) will be distributed to computing centers of collaborating institutes to be analyzed by physicists at their own institutes.

CMS software has been produced both for simulating the events as they will be produced by the CMS detector and data acquisition system and for processing them. Figure 1 describes the programs and data formats.

Event Generation Pythia and other generators. They are generally Fortran programs. They produce N-tuple files in the HEPEVT format;

Detector simulation Two programs are used: CMSIM that uses GEANT-3 and OSCAR that uses Geant-4. CMSIM is a fortran program and produces Formatted Zebra (FZ) files starting from the HEPEVT N-tuples. Requires an additional step (*ooHitFormatting*) to translate the FZ files into the POOL format. OSCAR is a C++ program that uses the CMS COBRA framework. It produces POOL files (the so-called *hits*) starting from the HEPEVT N-tuples.

Figure 1: CMS data formats and software used for data simulation and analysis. Description of components is in the text

Digitization This is the simulation of the DAQ process. It is done by ORCA, that uses the CMS COBRA framework. It is a C++ program that produces POOL files containing the so-called *digis* starting from hits POOL files or FZ (via the *ooHitFormatting* process).

Trigger simulation The simulation of the Level-1 trigger and of the High Level Trigger is done by ORCA, reading the *digis* POOL files. Trigger simulation is normally run as part of the reconstruction phase.

Reconstruction Production of DST's and AOD's is done by ORCA starting from the *digi* POOL files. Output files are also in the POOL format.

Analysis Both the Physics group and the end-user analysis is done using ORCA. It is possible to read any kind of POOL files produced in the previous steps. Visual analysis is done using IGUANA, a program that uses ORCA and OSCAR as back-end. It provides the functionalities needed for event displaying and statistical analysis.

3 OCTOPUS: CMS Production System

Figure 2 describes the components of the CMS data handling system. RefDB is a central database located at CERN where all information needed to produce and analyze data are kept. It has a web interface with kerberos authentication that allows the submission of processing requests, the assignment of work to distributed production centers and the browsing of the status of the requests. McRunjob and CMSProd are tools that allow the creation of the actual jobs to produce or analyze the data following the directives stored in RefDB or provided directly by a user. Jobs are prepared and submitted to local or distributed resources. Each job is instrumented to send to a dedicated database (BOSS) information about the running status of the job and to update the RefDB.

3.1 Job Production

McRunJob is a modular tool that creates jobs to be submitted to a variety of environments. At the time being CMS is using several modules that allow to read information on how to create the jobs from the CMS reference database

4 2002 Data productions

In spring 2002 a data production challenge provided the simulated data for the Trigger and Data Acquisition TDR analysis. About 6 million events in 150 physics channels have been simulated and then more than 13 million events have been digitized with different LHC luminosity configurations. This corresponded to about $200 KSI2000 \times$ months of CPU and more than 20 TB of data (in the form of Objectivity/DB databases). Productions were performed by 11 Regional Centers (corresponding to more than 20 sites). About 30 people (mainly the site managers) were involved.

In fall 2002 more than 10 million events were produced and analyzed to support studies for CMS calorimetry. They corresponded to about $300 KSI2000 \times$ months. Two grid prototypes were deployed to produce data for this challenge: the *CMS/EDG Stress Test* on the European DataGrid testbed and the *USCMS IGT Production* in the US. The former followed a *top-down* approach providing more functionality but less robustness, the latter followed a *bottom-up* approach providing less functionality but more stability and little manpower was needed to manage the production.

5 2004 Data Challenge (DC04)

The purpose of the 2004 Data Challenge is to demonstrate the validity of the software baseline to be used for the Physics TDR (due fall 2005) and in the preparation of the Computing TDR (due fall 2004).

The challenge comprises the completion of a *5% data challenge*, which successfully copes with a sustained data-taking rate equivalent to 25Hz at a luminosity of $0.2 \times 10^{34} cm^{-2} s^{-1}$ for a period of 1 month (approximately 5×10^7 events that have to be simulated up to the *digi* beforehand).

The emphasis of the challenge is on the validation of the deployed grid model on a sufficient number of Tier-0, Tier-1, and Tier-2 sites. Six of the Tier-1 centers and about 8-10 of the Tier-2 centers intending to supply computing to CMS in the 2007 first LHC run will participate to this challenge.

The DC04 will consist of the following phases:

- Reconstruction at Tier-0 (CERN) at 25 Hz
- Distribution of DST to Tier-1's
- Re-calibration at selected Tier-1's
- Physics-groups analysis at the Tier-1's
- User analysis from the Tier-2's

5.1 2003 Pre-Challenge production (PCP04)

Table 1: Summary of Pre-Challenge Production 2004 statistics (numbers for digitization report the situation on Feb 9th 2004 and the estimate to the end of PCP).

	n. events ($\times 10^6$)	CPU/event $KSI2000 \cdot s$	size/event MB	total CPU $KSI2000 \cdot \text{month}$	total size TB
Pythia	78	~ 0.15	0.02	4.6	1.6
CMSIM	50	$50 \rightarrow 150$	1	2000	50
OSCAR	17	$100 \rightarrow 200$	0.5	950	9
ORCA (digitization)	2/50	~ 30	~ 1	23/600	2/50

The Pre-Challenge Production is the preliminary phase that is preparing the about 50 million simulated events to be used during DC04 itself. It started in July 2003 and is now running the last step of the simulation chain (digitization). The summary of PCP04 statistics is in table 1

Part of the 50 TB of raw data are being produced outside CERN and need to be transferred to CERN before the start of DC04. We estimate more than 1 TB/day transfer rate in this phase. SRB (*Storage Resource Broker*) is used for data movement during all PCP04.

5.2 PCP04 on grid

Two grid-based prototypes have been used in PCP04: the *USMOP Regional Center* in the US and the *CMS-LCG Regional Center* mainly in Europe.

5.2.1 USMOP Regional Center

The USMOP Regional Center is using resources of the Grid3 System, consisting of about 2000 CPU's shared among US HENP experiments. The middleware used is based on VDT 1.1.11 which includes: EDG-VOMS for authentication, GLUE Schema for MDS Information Providers, MonaLisa for monitoring and MOP for production control.

The MOP System itself uses Dagman and Condor-G for job specification and submission and a Condor-based match-making process to select resources.

The system produced¹⁾ 7.7 millions pythia events, 2.3 CMSIM events and about 10 millions OSCAR events corresponding to about 80000 jobs, 650 *KSI*2000-month and 13 TB of data.

5.2.2 CMS-LCG Regional Center

The CMS-LCG Regional Center used the CMS/LCG-0 system for about 3 months and then part of the official LCG-1 testbed for about another month.

The CMS/LCG-0 was a CMS-wide testbed owned by CMS. It has been a joint CMS, DataTAG-WP4 and LCG-EIS effort started in June 2003. It is based on the LCG pilot distribution (VDT 1.1.6 and EDG 1.4.X) but includes also components from DataTAG (VOMS for VO management, GLUE schemas and info providers, GridICE for monitoring). It consisted of about 170 CPU's and 4 TB disk distributed over several sites²⁾. It allowed not only to test the system in a production environment but also to develop tools for DC04 itself (e.g. develop interfaces between COBRA and the LCG replication tools).

The system produced 0.5 million *heavy*³⁾ Pythia events, 1.5 million CMSIM events and 0.6 million OSCAR events (on LCG-1) corresponding to about 10000 jobs, 100 *KSI*2000-month and 3 TB of data.

5.3 DC04 Reconstruction Challenge

A *fake-DAQ* process will stage data from Castor MSS to a disk buffer at a rate of 40 MB/s which corresponds to the requested 25 Hz rate. The data files will be registered to a POOL RLS catalogue at CERN.

A computing farm of about 400 CPU's located at CERN will provide the 180 *KSI*2000-month CPU needed to run the ORCA reconstruction jobs for approximately 20 hours per day during the month of the challenge. Both the input raw data files and the output DST files will be saved to Castor MSS (~50 MB/s) and copied to a distribution buffer. All data will be registered to the RLS POOL catalogue and to a distribution database.

5.4 DC04 Data Distribution Challenge

All of the DST files and a fraction of the raw data files will be distributed to one or more Tier-1 centers. The overall transfer rate exiting from CERN will be about 1-2 TB/day.

An ad-hoc developed *Transfer Management database* will be used to steer data distribution. The RLS catalogue at CERN used as a POOL catalogue during reconstruction is also used to locate the physical instances of the data files CMS-wide.

Several data transfer tools will be used: the LCG Replica Manager tools, native SRM (Storage Resource Manager) and SRB (Storage Resource Broker). Each tool will have its own dedicated export buffer. All transfer tools will be able to store information about physical instances of the files into the RLS catalogue.

¹⁾ Situation on February 9th, 2004

²⁾ Bari, Bologna, Bristol, Brunel, CERN, CNAF, Ecole Polytechnique, Imperial College, ISLAMABAD-NCP, Legnaro, Milano, NCU-Taiwan, Padova, U.Iowa.

³⁾ The duration of these jobs is about 8 hours, compared to the 1.5 minutes of the usual Pythia jobs.

The central RLS catalogue at CERN will be replicated to some other Tier-1 using the *Oracle multi-master mirroring* procedure that should grant a reasonable synchronization of the catalogues. Other catalogues (like MySQL POOL catalogues or the SRB GMCAT catalogue) will need to be synchronized at each site using ad-hoc procedures.

5.5 DC04 Calibration and Physics Groups Analysis Challenge

Selected Tier-1 sites will run calibration procedures. This implies a rapid distribution of the calibration samples (within hours at most) to the Tier-1 site and automatically scheduled jobs to process the data as they arrive. The results will be published through the RLS catalogue and made available to the Tier-0 for incorporation in the calibration database. This challenge includes the ability to switch calibration database at the Tier-0 on the fly and to be able to track from the meta-data of each data file which calibration table has been used.

Selected Tier-1 sites will run Physics Group Analysis. The processed data should be made available to the Tier-0 and to the Tier-2 centers connected to the same *region* (e.g. CNAF has to make files available to at least the italian Tier-2's). The same RLS catalogue will be used to store information about the data produced during this analysis step.

5.6 DC04 User Analysis Challenge

About 8-10 Tier-2 centers will also participate to the Data Challenge. The main task for the Tier-2 centers is to provide to the end-users the ability to access the computing resources and the data (either local or at the Tier-1). Most of the Tier-2 centers will exploit the LCG tools to get access to remote resources: submission of jobs through the LCG Resource Broker, replication and download of data to local storage via the Replica Manager, sharing of private data with remote collaborators.

Private analysis on distributed data is outside DC04 scope but will be kept as a low-priority milestone. A new tool based on BOSS and optimized for use on the grid, GROSS, is being developed for this purpose.

6 Conclusions

Computing is a CMS-wide activity that doesn't include only the development of the computing system in time for the LCH startup but also the support to other CMS activities i.e. detector, DAQ, Trigger and Physics studies. The nature and primary goals of the CMS distributed data challenges reflect this mandate.

The CMS computing system is increasing in size and complexity. Data challenges started in 1999 with 1 TB produced in 1 month in 1 site and today it is producing about 170 TB in 6 months in 50 sites. The grid paradigm has been adopted and there is a strong collaboration with the LCG project at CERN and with all the other EU and US grid projects. Several prototypes have been deployed and used demonstrating an increasing level of usability and efficiency.