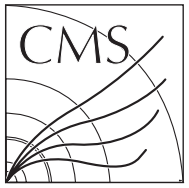
Available on CMS information server

CMS CR -2010/273



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

26 November 2010 (v2, 13 January 2011)

CMS Distributed Computing Workflow Experience

Jeffrey David Haas for the CMS Collaboration

Abstract

The vast majority of the CMS Computing capacity, which is organized in a tiered hierarchy, is located away from CERN. The 7 Tier-1 sites archive the LHC proton-proton collision data that is initially processed at CERN. These sites provide access to all recorded and simulated data for the Tier-2 sites, via wide-area network (WAN) transfers. All central data processing workflows are executed at the Tier-1 level, which contain re-reconstruction and skimming workflows of collision data as well as reprocessing of simulated data to adapt to changing detector conditions. This paper describes the operation of the CMS processing infrastructure at the Tier-1 level. The Tier-1 workflows are described in detail. The operational optimization of resource usage is described. In particular, the variation of different workflows during the data taking period of 2010, their efficiencies and latencies as well as their impact on the delivery of physics results is discussed and lessons are drawn from this experience. The simulation of proton-proton collisions for the CMS experiment is primarily carried out at the second tier of the CMS computing infrastructure. Half of the Tier-2 sites of CMS are reserved for central Monte Carlo (MC) production while the other half is available for user analysis. This paper summarizes the large throughput of the MC production operation during the data taking period of 2010 and discusses the latencies and efficiencies of the various types of MC production workflows. We present the operational procedures to optimize the usage of available resources and we the operational model of CMS for including opportunistic resources, such as the larger Tier-3 sites, into the central production operation.

Presented at *CHEP2010: International Conference on Computing in High Energy and Nuclear Physics 2010*

CMS distributed computing workflow experience

Jennifer Adelman-McCarthy[1], Oliver Gutsche[1], Jeffrey D. Haas[2], Harrison B. Prosper[2],Valentina Dutta[3], Guillelmo Gomez-Ceballos[3], Kristian Hahn[3], Markus Klute[3], Ajit Mohapatra[4], Vincenzo Spinoso[5], Dorian Kcira[6], Julien Caudron[7], Junhui Liao[7], Arnaud Pin[7], Nicolas Schul[7], Gilles De Lentdecker[8], Joseph McCartin[9], Lukas Vanelderen[9], Xavier Janssen[10],Andrey Tsyganov[11], Derek Barge[12] and Andrew Lahiff[13]

[1] Fermi National Accelerator Laboratory, USA
[2] Florida State University, USA
[3] Massachusetts Institute of Technology, USA
[4] University of Wisconsin-Madison, USA
[5] INFN Bari, Italy
[6] California Institute of Technology, USA
[7] Universite Catholique de Louvain, Belgium
[8] Universite Libre de Bruxelles, Belgium
[9] Universiteit Gent, Belgium
[10] Universiteit Antwerpen, Belgium
[11] Joint Inst. for Nuclear Research, Russian Federation
[12] University of California Santa Barbara, USA
[13] Rutherford Appleton Laboratory, United Kingdom

Abstract. The vast majority of the CMS Computing capacity, which is organized in a tiered hierarchy, is located away from CERN. The 7 Tier-1 sites archive the LHC proton-proton collision data that is initially processed at CERN. These sites provide access to all recorded and simulated data for the Tier-2 sites, via wide-area network (WAN) transfers. All central data processing workflows are executed at the Tier-1 level, which contain re-reconstruction and skimming workflows of collision data as well as reprocessing of simulated data to adapt to changing detector conditions. This paper describes the operation of the CMS processing infrastructure at the Tier-1 level. The Tier-1 workflows are described in detail. The operational optimization of resource usage is described. In particular, the variation of different workflows during the data taking period of 2010, their efficiencies and latencies as well as their impact on the delivery of physics results is discussed and lessons are drawn from this experience. The simulation of proton-proton collisions for the CMS experiment is primarily carried out at the second tier of the CMS computing infrastructure. Half of the Tier-2 sites of CMS are reserved for central Monte Carlo (MC) production while the other half is available for user analysis. This paper summarizes the large throughput of the MC production operation during the data taking period of 2010 and discusses the latencies and efficiencies of the various types of MC production workflows. We present the operational procedures to optimize the usage of available resources and we the operational model of CMS for including opportunistic resources, such as the larger Tier-3 sites, into the central production operation.

1. Introduction

In 2010, the LHC [1] at CERN started its physics program with the first long run collecting proton-proton collisions at a center-of-mass energy of 7 TeV. 2010 also marked the final transition of the CMS [2] computing systems from preparation to operation. This paper will describe the experience of the CMS collaboration with the data processing and Monte-Carlo (MC) production in 2010.

Several ingredients were necessary to complete these tasks, which include: software [5], workload and data management tools [6, 7], grid infrastructure [3, 4], CMS Tier-1 and Tier-2 sites and the operation teams to keep everything alive and working [8]. We would like to thank the developers of our tools, our integration teams, the CMS facility operations group, those who care for the functionality of sites, and all the rest of CMS who contribute to this team effort.

The Compact Muon Solenoid (CMS) Computing Model [9] is designed as a hierarchical structure of computing centers with well defined roles, located throughout the world. The CMS Computing resources follow a tree model of tier levels (computing centers) ranging from Tier 3 to Tier 0. These resources are part of the World-wide Large Hadron Collider Computing Grid (WLCG [10]).

A large majority of the CMS computing capacity is not located at the LHC host laboratory CERN, but is distributed around the world. CERN is at the top of the hierarchical structure as the only Tier-0 center. The Tier-0 is responsible for the safe keeping of the first copy of experimental RAW data (archived on tape, considered a cold backup copy not intended to be accessed frequently), prompt data processing, prompt calibration, and the distribution of data to all Tier-1 centers.

There are a total of 7 Tier-1 centers, located at large universities and national laboratories in France, Germany, Italy, Spain, Taiwan, the United Kingdom and the United States. Tier-1s are at the center of the data flow. The Tier-0 sends the raw and reconstructed data for custodial care (archived on tape) to the Tier-1s. Monte-Carlo simulations produced at the Tier-2s are also sent to the Tier-1s for custodial care (archived on tape). The Tier-1s perform event re-reconstruction and skimming workflows on the data, where the outputs are distributed to the Tier-2s. Since August 2010, the Tier-1s also process Monte Carlo simulations of data if resources are available.

The Tier-2 centers are located at about 50 sites around the world. The Tier-2s do not have tape systems available, all data are cached on disk for analysis. The Tier-2 level represents the primary analysis facilities for CMS. Monte Carlo simulations are mainly carried out at the Tier-2 level as well. The Tier-2s rely on Tier-1s as their link to CMS data and MC simulations for analysis access.

The Tier-3 level is special in the sense that it is not a pledged resource of the experiments, but rather voluntarily provided to CMS. A Tier-2 must have sufficient CPU and disk space to produce Monte Carlo simulations and to support CMS analysis activities, while a Tier-3 does not have these requirements. Therefore, while CMS can use Tier-3 resources for opportunistic purposes, it cannot rely on their availability.

2. Processing at Tier-1 level

The Tier-1 level takes care of all processing that needs input from samples custodially archived on tape. In the following, CMS' concepts of data partitioning and the characteristics of Tier-1 workflows are introduced followed by the summary of processing during 2010.

2.1. Data Partitioning

CMS stores events recorded by the detector system and its derived products in files of different contents. The following main data tiers characterize the content of these files:

- RAW: RAW event data contains detector data and trigger information. The largest contributor to the RAW event size of the order of 500 kB is the silicon strip detector.

- RECO: The Reconstructed data (RECO) contain reconstructed physics quantities derived from RAW data. Detector calibration constants are applied and physics objects are identified. A RECO event is about 400 kB in size.
- AOD: The Analysis Object Data (AOD) contains a small subset of the RECO data format, sufficient for 90% of all physics analyses. An AOD event is about 120 kB in size.
- GEN-SIM-RAW: The RAW data tier originating from Monte Carlo (MC) simulations with the Generator (GEN) information and the Simulation (SIM) information added. A GEN-SIM-RAW event is about 1000 kB in size.
- GEN-SIM-RECO: Corresponds to RECO using GEN-SIM-RAW as input and contains small amounts of generator and simulation information. A GEN-SIM-RECO event is about 500 kB in size.
- AODSIM: The AODSIM format contains a small subset of the GEN-SIM-RECO data format sufficient for 90% of all analyses. An AODSIM event is about 140 kB in size.

CMS determines the luminosity corresponding to the recorded data in granularity of a Luminosity Section (LS) which constitutes 23 seconds of data taking. In case of MC simulation, a LS holds the events of a single MC production job. A LS is always kept intact in a single file and not split between several files to guarantee bookkeeping of the luminosity during processing and analysis. The size of individual files is 2-10 GB, optimized for tape storage.

Files are grouped together into file-blocks of 500 to 1000 files. Blocks contain no more than one run. Site location is tracked on a block level and only complete blocks are available for processing, partially transferred blocks have to wait for the completion of the transfer to be processed.

Data from the detector is split into Primary Datasets (PD) by trigger selections by physics interest. Examples are the Electron, Photon and Jet PDs. MC samples are split by their generator configuration, like QCD or TTBar.

2.2. Processing workflows

The Production Agent (PA) [11] is the main component of the CMS Workload Management System, which enables large processing of data using CMS software. It is a message based modular python workload management system. There are 16 autonomous components, python daemons, within the PA. These components take care of job creation, submission, tracking, error handling, job cleanup, data merging and data publication.

CMS distinguishes two main processing workflow types on the Tier-1 level: data re-reconstruction and MC re-digitization/re-reconstruction, (a) and b) in Fig. 1). During data re-reconstruction, the Tier-1 sites process RAW data with newer software releases and/or updated alignment and detector condition constants, producing data in both the RECO and AOD formats. The processing jobs can produce output file sizes that are too small and non-optimal for tape storage. Therefore, a dedicated merge step combines the unmerged outputs of several processing jobs with the same data format. The Tier-1s then skim the reconstructed data, extracting events of interest, in a separate step. These events are written out into files of RECO or a combination of RECO and RAW formats and follow the same processing/merge cycle as the re-reconstruction (see a in Fig. 1).

The Tier-1 sites also reprocess Monte Carlo generated events with newer software versions and/or newer alignment and calibration constants. The GEN-SIM-RAW input is re-digitized producing an updated version of the simulated RAW data, which are then re-reconstructed. In order to eliminate multi-step processing (processing of a dependent workflow after waiting for completion of the merge step of the previous workflow), maximize computer resources and improve the production efficiency significantly, Chained Processing (CI) was established (see b in Fig. 1). In Chained Processing, all workflow steps are processed one after the other on

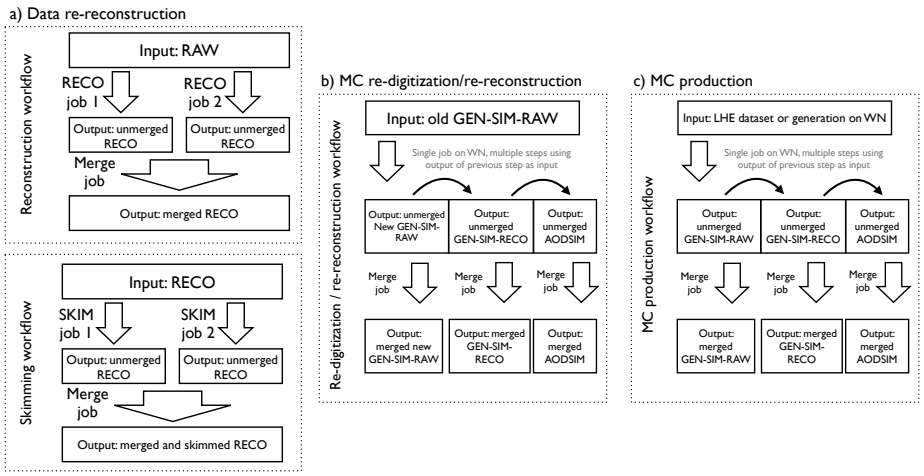


Figure 1. Schematic overview of CMS' processing and production workflows: a) data re-reconstruction, b) MC re-digitization/re-reconstruction, c) MC production.

the same workernode using the output of the previous step as input for the following step. In Chained Processing, the outputs are merged individually eliminating the need to wait for the completion of the merge step of the previous workflow.

Processing of input samples is not split between different sites but rather processed completely at a single site. The processing of a complete sample is optimized by splitting it into smaller jobs. Each job should run about 8-12 hours to optimize resource usage. Job splitting is done by file to keep luminosity sections intact. We also follow a run-based merging policy to avoid having files contain more than one run. During the processing, the intermediate output is stored on disk-only areas.

2.3. Processing experience in 2010

The Tier-1 sites have been stable during the 2010 collisions data taking period [8]. Apart from their custodial allocation, all RAW collision data samples have been distributed to all Tier-1s to increase processing flexibility. This was possible because of the small total data size. CMS collected cosmic data early in 2010. In March, the Large Hadron Collider (LHC) provided proton-proton collisions, but the bulk of the integrated luminosity has been collected since September, when beam luminosity increased due to the use of closely packed bunches in proton bunch trains. CMS recorded collisions at a data taking rate of 300Hz, with spikes reaching 700Hz. The primary datasets per data acquisition era are shown in Table 1.

CMS performed 3 MC re-digitization/re-reconstruction campaigns in 2010 (see Tab. 2) that produced significantly more output than the data taking including the over 16 re-reconstruction passes (see Tab. 3). The CPU needs for the re-reconstruction passes were small compared to the needs for the MC campaigns but increased after September 2010 with the increasing collected luminosity (see (*left*) in Fig. 2).

The Tier-1 production has been very successful during 2010 and the tools and operation teams significantly contributed to the timely publication of the first physics results of CMS, but not without challenges. Production efficiency suffered from lengthy debugging efforts before production quality of the workflows could be reached. The large number of requests extended the time spent on bookkeeping and completion of the workflows. This caused additional false starts due to pilot error. All processing of data requires a detailed post-mortem for each failed job; this was labor intensive and time consuming with the tools at hand. The production infrastructure

PD	Com10	2010A	2010B
Cosmics	593.1	264.2	68.2
MinimumBias	1339.9	119.2	19.0
ZeroBias	438.7	34.9	14.9
JetMETTau		168.0	
JetMET		31.6	
BTau		27.8	12.5
EG		61.8	
Mu		56.0	10.6
MuOnia		37.4	11.8
Commissioning		181.9	7.2
Jet			13.1
MultiJet			1.1
METFwd			8.2
Total	2371.7	982.7	166.9

Table 1. Number of Million events in Primary Datasets per Data Acquisition Era in 2010.

	Spring10	Summer10	Fall10
GEN-SIM-RAW			
Events (M)	658.6	592.5	469.0
Size (TB)	481.3	412.4	322.5
GEN-SIM-RECO			
Events (M)	744.6	592.0	469.0
Size (TB)	267.7	234.5	165.7
AODSIM			
Events (M)	658.0	588.0	469.0
Size (TB)	78.6	57.3	39.3

Table 2. Number of Million events (M) per MC re-digitization/re-reconstruction campaign in 2010.

	CMS Internal	Events (M)	Luminosity	Start date	Completion (days)
1	Jan23ReReco	40.7	N/A	01/23/10	2
2	Jan29ReReco	44.6	N/A	01/29/10	3
3	Feb9ReReco	44.6	N/A	02/09/10	2
4	Mar1rstReReco	6.2	N/A	03/01/10	5
5	Mar3rdReReco	223.0	N/A	03/03/10	5
6	Apr1ReReco	10.5	0.032	04/01/10	3
7	Apr20ReReco	168.8	0.516	04/20/10	1
8	May6thReReco	338.7	1.663	05/06/10	4
9	May27thReReco	824.4	18.195	05/27/10	3
10	Jun9thReReco (ICHEP)	1003.3	19.593	06/09/10	7
11	Jun14thReReco	1012.0	50.343	06/14/10	6
12	Jul6thReReco	26016800	83.291	07/06/10	1
13	Jul15thReReco	40.4	193.420	07/15/10	1
14	Jul16thReReco (ICHEP)	16.6	132.605	07/16/10	1
15	Jul26thReReco	11.5	115.010	07/26/10	1
16	Sep17ReReco	1295.8	3493.308	09/17/10	10

Table 3. Re-Reconstruction Passes During 2010; 7 TeV re-reconstruction passes start April 1st, 2010

imposed its own restrictions due to performance reasons; a single instance was limited to 3000 jobs running in parallel. Due to the messaging based design, jobs got lost during processing whose recovery was lengthy and difficult if not impossible in some cases.

3. MC production at the Tier-2 level

Tier-2s represent the primary CMS MC production and analysis facilities, where 50% of the resources are committed to producing MC simulations and 50% are committed for use in CMS analysis. Output from the MC production is archived on tape at the Tier-1 centers. The Tier-2s

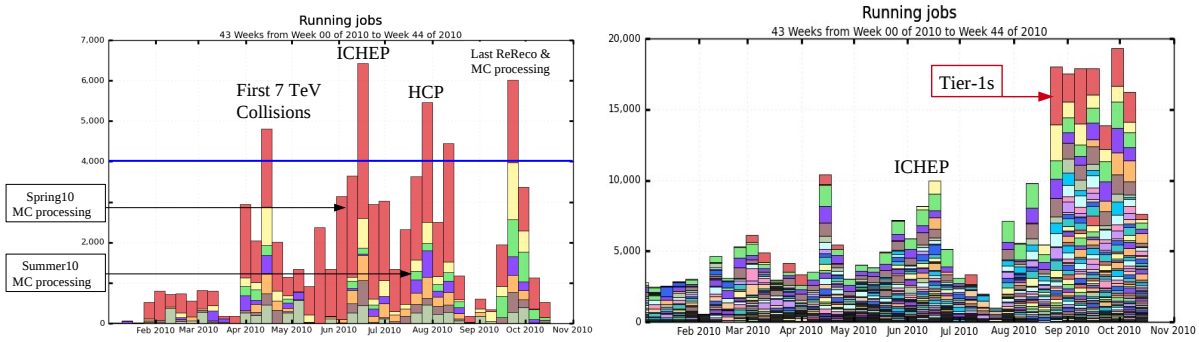


Figure 2. CMS processing (*left*) and MC production (*right*) jobs during 2010.

are divided up into geographic regions, grouped around Tier-1s. The NorduGrid region stands on its own because of their different middleware technology. These regions are managed by 5 operator teams.

3.1. MC Production

CMS requires a large number of MC events to supplement the data physics studies. The task of generating billions of MC events in a timely manner is accomplished using PAs like the Tier-1 processing. MC events are produced at all Tier-2 sites, a few opportunistic Tier-3 sites and, as of August 2010, Tier-1 sites in order to make better use of free resources. The MC production workflow is executed using a Chained Processing workflow type (see c in Fig. 1) where 3 outputs for GEN-SIM-RAW, GEN-SIM-RECO and AODSIM are stored. During 2009 and 2010, CMS produced over 3.5 billion events. Normal MC production capacity is about 300 Million events per month, however during September 2010 500 Million events were produced due to low-occupancy event compositions. Figure 2 shows the number of MC production jobs running in parallel in 2010. The increase in number of jobs in August 2010 is due to the significantly increased demand for MC events and the possibility to use free resources at the Tier-1 sites for production.

The MC production in 2009/2010 was very successful and could meet all demands. Also here, some issues were noticed. Apart from the same production infrastructure issues like the Tier-1 processing, the large number of different sites created a multitude of individual problems. Although the GRID infrastructure increased in stability over time and was very good in 2010 [8], occurring problems with basic services like compute and storage elements or individual workernodes increased the time effort for debugging and decreased the production efficiency.

4. Conclusions & Outlook

This has been a successful year for CMS' distributed workflow management in delivering input for successful first physics analysis with LHC proton-proton data: data were re-reconstructed 22 times; 3 Monte Carlo re-digitization/re-reconstruction campaigns were completed since the start of the 2010 run. Over the last 2 years, 3 billion Monte Carlo events were produced.

Looking into the future, developments to overcome shortcomings of the current workload management system (PA) are undergoing integration tests. The architecture of the new CMS Workload Management system (WMAgent) is based on a state machine rather than a messaging system to keep track of each and every processing job reliably and with 100% accountability. The new system will be the bases of all processing tasks at Tier-0, Tier-1, and MC production and analysis. The expected increase in production efficiency will make sure that CMS will meet its demands in producing input for physics analysis in the years to come.

5. Acknowledgements

We would like to thank the production tools and software development teams, the site administrators and facility teams, operation teams and all other CMS collaborators who contributed to a successful operation of all production and processing workflows on CMS' distributed computing infrastructure. We thank the international funding agencies amongst those the Department of Energy and the National Science Foundation for their support of this work.

[1] O. Bruning, P. Collier, P. Lebrun, S. Myers, R. Ostojic, J. Poole and P. Proudlock, "LHC design report. Vol. I: The LHC main ring", *CERN-2004-003*

[2] CMS Collaboration, "CMS, the Compact Muon Solenoid: Technical proposal",*CERN-LHCC-94-38*

[3] Enabling Grids for E-sciencE `http://www.eu-egee.org/`

[4] Open Science Grid `http://www.opensciencegrid.org/`

[5] C. Jones et al, "The new CMS event data model and framework", Proceedings for Computing in High-Energy Physics (CHEP '06), Mumbai, India, 13 Feb - 17 Feb 2006

[6] Egeland, R. and others, "Data transfer infrastructure for CMS data taking", Proceedings for XII Advanced Computing and Analysis Techniques in Physics Research, Erice, Italy (Nov. 2008)

[7] A.Afaq,et.al.,The CMS Dataset Bookkeeping Service,J.Phys.Conf.Ser,119,072001(2008).

[8] D. Bonacorsi, Experience with the CMS Computing Model from commissioning to collisions, CHEP 2010

[9] The CMS Collaboration CMS Computing Technical Design Report, CERN-LHCC-2005-023, (2005)

[10] Worldwide LHC Computing Grid (WLCG) urlhttp://lcg.web.cern.ch/LCG/public/default.htm

[11] Wakefield, S. and others, "Large Scale Job Management and Experience in Recent Data Challenges within the LHC CMS experiment, Proceedings for XII Advanced Computing and Analysis Techniques in Physics Research, Erice, Italy, Nov. 2008