Experience with the CMS Computing Model from commissioning to collisions

To cite this article: Daniele Bonacorsi and the CMS Computing project 2011 J. Phys.: Conf. Ser. 331 072005

View the article online for updates and enhancements.

You may also like

- Large scale commissioning and operational experience with tier-2 to tier-2 data transfer links in CMS
  J Letts and N Magini

- Challenges for the CMS computing model in the first year
  I Fisk and (on behalf of the Cms Offline and Computing Project)

- Exploiting CMS data popularity to model the evolution of data management for Run-2 and beyond
  D Bonacorsi, T Boccali, D Giordano et al.
Experience with the CMS Computing Model from commissioning to collisions

Daniele Bonacorsi (for the CMS Computing project)
University of Bologna, Department of Physics, Viale Berti Pichat 6/2, Bologna, Italy
E-mail: daniele.bonacorsi@bo.infn.it

Abstract. In this presentation we will discuss the early experience with the CMS computing model from the last large scale challenge activities through the first six months of data taking. Between the initial definition of the CMS Computing Model in 2004 and the start of high energy collisions in 2010, CMS exercised the infrastructure with numerous scaling tests and service challenges. We will discuss how those tests have helped prepare the experiment for operations and how representative the challenges were to the early experience with data taking. We will outline how the experiment operations has evolved during the first few months of operations. The current state of the Computing system will be presented and we will describe the initial experience with active users and real data. We will address the issues that worked well in addition to identifying areas where future development and refinement is needed.

1. The CMS Computing Model

The CMS experiment at the LHC accelerator designed and built a Computing Model [1] as a distributed system of computing resources and services relying on Grid technologies. A general software layer able to access and use the global computing infrastructure of Grid resources and services is offered to the LHC experiments by the Grid projects. For CMS, they mainly are Enabling Grids for E-sciencE (EGEE) and European Grid Infrastructure (EGI) in Europe and Asia [2], and Open Science Grid (OSG) in the US [3]. CMS has been building its own specific application layer on top of this basic layer, using the Grid components and services as building blocks of its computing infrastructure. The Worldwide LHC Computing Grid (WLCG) project [4] provides the overall coordination among LHC experiments and Grid projects. Since the MONARC project [5], a hierarchical set of computing resources was proposed, with a Tier-0 (T0) center plus a CERN Analysis Facility (CMS-CAF) located at CERN, 7 Tier-1 (T1) centers at large regional computing sites, about 50 Tier-2 (T2) centers distributed worldwide and many Institutional Tier-3 (T3) centers. These resources together comprise the computing, storage and connectivity systems that CMS relies on in performing all its workflows: data recording and archiving, data transfer and processing, Monte Carlo event simulations, physics end-users analysis. The group of computing Tiers supporting the CMS organization form an inter-connected and inter-operating ensemble and work as a coherent system, although each Tier level inherently provides different resources and services.

The T0, a unique centre located at CERN, accepts RAW data from the CMS Online Data Acquisition and Trigger system, archives the RAW data to tape for custody, groups RAW data into data streams, feeds the prompt first-pass reconstruction step, classifies the reconstructed...
Figure 1. Readiness of CMS Tier-1 (a) and Tier-2 (b) monitored as a function of time. Green, red, purple dots refer to ‘ready’, ‘not ready’, ‘in scheduled downtime’ sites, respectively.

(RECO) data into up to 50 ‘primary’ datasets according to their physics content (i.e. trigger path), makes them available for transfers to the next Tier stage resources (i.e. T1s), hosts and maintains the CMS-CAF resources devoted to latency-critical analysis tasks (trigger performance services, detector diagnostics, derivation of calibration and alignment constants). The next Tier level consists of a set of 7 large T1 computing centers in collaborating countries: ASGC in Taiwan, GridKA/KIT in Germany, INFN-CNAF in Italy, IN2P3 in France, PIC in Spain, RAL in UK, FNAL in the US. Each T1 receives some subset of the primary datasets from the T0, provides tape archiving capability and substantial CPU power for several scheduled data-intensive tasks (e.g. re-reconstruction, skimming), distributes RECOs, skims and AODs to other Tier levels, thus feeding and populating the distributed storage at next Tier levels. A numerous set of T2 centers (currently about 50) - with consistent CPU resources but limited disk space (and no tape archives) - provide the computing capacity for physics users analysis and Monte Carlo simulations. The T2s import the datasets from T1s and other T2s over a fully deployed transfer matrix, and rely upon T1 tape systems to safely store the results of the simulations.

2. Commissioning the CMS computing infrastructure

The coherent operation of the worldwide computing resources to serve the CMS experiment requires a stable and reliable behaviour of each component of the underlying infrastructure. Over the years CMS has established procedures to extensively test all the most relevant aspects of Grid sites, such the efficient use of the network to transfer data inbound or outbound, the good functionality of all the on-site services used by CMS, and the capability to sustain all the various CMS computing workflows at the required scale. A logical combination of selected tests is used as a single estimator to rank the ‘readiness’ of a Tier. This is important for the sites - as a tool to find out and debug site problems - and for the CMS Computing Operations teams - to monitor how the situation evolves over time [6]. The outcome of such tests are overviewed daily, and the problems spotted by the tests are regularly followed up. The CMS Computing shift team is responsible for identification and notification of a problem via tickets to problematic sites, which are hence encouraged to provide a better service if they want to be used to execute CMS workflows. The “Site Readiness” for all CMS T1 and T2 sites since late 2008 is shown in Fig 1. The results obtained up to the start of LHC data taking at 7 Tev show a clear improvement in the number of ‘ready’ sites and an approximately constant level of reliability of sites which is not expected to change much over time.

Along with regular activities to commission and monitor site services, ad-hoc efforts were devoted to improve the overall system readiness in view of the start of LHC data taking. Over the recent years a set of very wide and complete computing scale tests (“challenges”) were organized
and run - both by WLCG and by each LHC experiment. They were focussed exercises, with a limited duration in time and well specified goals, with the purpose of verifying the status of the overall Grid infrastructure, identify fragile areas and focus the planning towards the most critical areas. CMS was very active and exercised its computing systems with yearly-based challenges of increasing complexity, i.e. a Data Challenge in 2004 [7], a set of repeated Computing Software and Analysis challenges in 2006, 2007 and 2008 [8, 9, 10, 11]. Additionally, since early in the 2000-2010 decade, WLCG organized and coordinated several Service Challenges (SC1, SC2, SC3, SC4 [12]), which culminated in two major worldwide computing exercises, namely the WLCG Common Computing Readiness Challenge (CCRC) in 2008 [13, 14] and the Scale Test for the Experiment Program (STEP) in 2009 [15]. All these efforts involved dozens of computing centers worldwide and few hundreds of physicists and computer scientists. In particular, the last two challenges (CCRC’08 and STEP’09) emphasized the simultaneous test of the computing systems of all 4 LHC experiments, in the most realistic manner. The STEP09 challenge was crucial as a final exercise before the start of 7 TeV data taking, since it emphasized the simultaneity of many tests and the overlap among experiments.

3. From commissioning to collisions
The CMS computing workflows ran very smoothly at the start of the LHC data taking at 7 TeV in 2010. During collision running CMS performed workflows and activities as predicted in the Computing Model, i.e. the WLCG Tiers performed: i) prompt processing at the T0; ii) output to the CAF for prompt feedback and alignment/calibration activities; iii) transfer to T1s for storage and distribution to T2s; iv) prompt skimming and reprocessing at T1s; v) Monte Carlo production and analysis activities at T2s. In 2010 the LHC live time was lower than we expected from the original planning, hence the total data volume was smaller than foreseen and CMS experienced less load on computing systems than could have been accommodated. The predicted activities were performed more frequently, e.g. reprocessing and analysis were regularly exercised and the data were subscribed to all T1s and even oversubscribed to T2s. It’s remarkable that all components of the model were exercised and largely demonstrated to be functional at the needed scale in 2010 with no major problems or bottlenecks. The luminosity ramp and the increase in total data will be keeping Computing more and more busy in next months. Room for changes in order to better use available resources was also identified - especially at T1s - and it will be important when moving to a resource-constrained operational environment.
The T0 system performed efficiently in 2010: the different kind of jobs (related to the express processing and prompt reconstruction tasks) recorded an averaged job efficiency >99.95%. The CMS software and the system reliability is good, and the level of automation is high. The CAF has also been heavily used in burst: jobs started quickly in low-latency queues, and >250 active users were observed over last 6 months. The processing latency for time-critical applications (e.g. express) was well within the design goals (see Fig. 2a), and the overall system demonstrated the ability to react promptly to LHC fills with remarkably good performances (see Fig. 2b).

The Tier-1 sites functioned well for prompt skimming and reprocessing, with the high level activity driven by the 7 TeV re-reconstruction passes. All 7 CMS T1s participated to the reprocessing passes, receiving and running a number of jobs that depends on the fraction of custodial data hosted on-site. The T1 usage pattern will change already in 2011, when CMS Computing will need to operate in a resource-constrained environment. CMS demonstrated the ability to quickly turn over the data passes needed for Summer conferences (e.g. ICHEP’10, HCP’10): over more than 10 re-reconstruction passes on 2010 pp data at 7 TeV, the time needed to complete the reprocessing of the full sample was never higher that 7-10 days.

The aggregate data transfer volume by CMS among Grid sites since 2004 is shown in Fig. 3. The improvements over years (note the logarithmic scale) are clearly visible, thanks to ad-hoc computing challenges of increasing complexity and to the regular computing commissioning activities. The transfer of the 7 TeV data sample recorded so far by CMS smoothly flew on various routes, mainly from T0 to T1 sites (highly correlated with LHC fills), from T1 to T2 sites (serving data to the CMS analysis layer), and among Tiers at the same level (T1-T1 and T2-

![Figure 3. CMS PhEDEx aggregate data volume since 2004, on a log scale. Different colors show the volume transferred in specific computing challenges or in constant commissioning activities. The normal, production traffic of the CMS experiment is labelled as “Production”.

---

4
A total data volume of about 5 PB was transferred from CERN to T1s in 2010, with remarkable peaks at >200 TB/day sustained for more than one month in October-November 2010. All T1s imported their needed fraction of the data effectively with a remarkably good transfer quality. The T1-T1 traffic was relatively low in size but very useful, since triggered by PhEDEx from various source T1 locations to allow a quicker sample completion at destination T1 sites. A serious T2-T2 traffic was operated in production for the first time: the majority of links among T2s were also used, since the full T2 transfer mesh was deployed in early 2010 to allow flexibility and to increase the sample replication speed. During this commissioning effort, up to 30 new T2-T2 links were commissioned per day, averaging at about 7 links/day over last 6 months [18]. As a result of this commissioning effort, smooth transfers of physics results among all T2s associated to same analysis groups were observed during 2010 data taking, as well as a full support to newly deployed fall-back stage-out tactics within CRAB (the CMS Remote Analysis Builder used to handle the analysis tasks [19]). Fig. 4 shows the traffic among Tiers month by month in 2010, with the breakdown in the different routes. The T2-T2 component is highly visible, and gave an important contribution to the overall PhEDEx traffic in 2010. The number of T2-T2 links used each month peaks to more than 600 connections used for production data transfers during May-June 2010 (not always the same ones were used). Nearly 7 PB of official samples are hosted at T2s for analysis access, and the community of CMS T2s

Figure 4. Total transfer volume in 2010 with monthly breakdown into transfer routes. The line at the top indicates the number of Tier-2 to Tier-2 links active every month (see text).

Figure 5. (a) Number of CMS analysis jobs completed in the period March-April 2010; the ramp up after the start of LHC data taking at 7 TeV is visible. (b) Number of CMS analysis users (per week) as a function of time.
provided an efficient access to large datasets with skimmed collision data and the corresponding simulation samples. The number of analysis users is now steadily increasing: we now record >800 individual analysis users per month, on average >400 individuals submitting analysis jobs every week since the start of the 7 TeV data taking in March 2010, with daily peaks at >100k jobs/day and weekly peaks at >140k jobs/day (see Fig. 5).

The Tiers usage by Monte Carlo production is driven by requests, and is optimized on all the computing resources available to CMS. The Monte Carlo production, originally designed to run on T2 (and opportunistic T3) resources, were in fact opened at T1 sites also, as from August 2010. In September-October 2010, up to about 18k simultaneously running jobs were observed at CMS Tiers.

4. Summary
The CMS Computing system has been deployed through many years of daily commissioning work, as well as through ad-hoc computing challenges, both CMS-specific and in a WLCG context. The deployed tools and services are working smoothly for LHC physics at 7 TeV so far, and are able to cope with the load in all sectors. CMS is not running in a resource constrained environment yet, but the planning calls for interesting times soon, as the integrated volume of data and the live time of accelerator will increase.

Acknowledgements
The author, on behalf of the CMS Computing project, would like to thank to all the CMS contacts and the site admins at the distributed WLCG Tiers for the hard, competent and constant work, as well as the overall WLCG collaboration and management.

References
[1] C. Grandi, D. Stickland, L. Taylor et al., “The CMS Computing Model”, CERN-LHCC-2004-035/G-083 (2004)
[2] EGEE: www.eu-egee.org/; EG: www.egi.eu/
[3] OSG: www.opensciencegrid.org/
[4] WLCG project: http://lcg.web.cern.ch/lcg/
[5] MONARC project: http://monarc.web.cern.ch/MONARC/
[6] S. Belforte et al., “The commissioning of CMS Computing Centres in the WLCG Grid”, ACAT’08, Erice, Nov. 2008 - PoS (ACAT08) 043 (2008)
[7] D. Bonacorsi et al., “Towards the operation of the INFN Tier-1 for CMS: lessons learned from CMS DC04”, ACAT’05, DESY, Zeuthen, 2005
[8] I. Fisk, “CMS Experiences with Computing Software and Analysis Challenges”, CHEP’07, Victoria BC, Canada, Sept. 2007
[9] O. Gutsche et al., ‘WLCG scale testing during CMS data challenges, J. Phys. Conf. Ser. 119 062033 (2008)
doi:10.1088/1742-6596/119/6/062033
[10] CMS CSA’07 twiki portal: https://twiki.cern.ch/twiki/bin/view/CMS/CSA07
[11] D. Futyan et al., “The CMS Computing, Software, Analysis Challenge 2008”, CHEP’09, Prague, Mar. 2009
[12] D. Bonacorsi, “WLCG Service Challenges and Tiered architecture in the LHC era”, IFAE, Pavia, Apr. 2006
[13] J.D. Shiers et al., The (WLCG) Common Computing Readiness Challenge(s) - CCRC’08, contribution N29-2, Grid Computing session - Nuclear Science Symposium, IEEE (Dresden), October 2008
[14] L. Bauerdick and D. Bonacorsi, “CMS results in the Combined Computing Readiness Challenge (CCRC’08)”, IPRD’08, Siena, Jun. 2010
[15] D. Bonacorsi and O. Gutsche, “CMS from STEP’09 to Data Taking: CMS Computing experiences from the WLCG STEP’09 challenge to the First Data Taking of the LHC era”, ISGC’10, Taipei, Taiwan, Mar. 2010
[16] L. Tuura et al., “Scaling CMS data transfer system for LHC start-up”, CHEP’07, Victoria BC, Canada, Sep. 2007 - J. Phys.: Conf. Ser. 119 072030 (2008)
[17] N. Magini et al., “The CMS Data Transfer Test Environment in Preparation for LHC Data Taking”, IEEE Nuclear Science Symposium, Dresden, 2008
[18] J. Letts, “Large Scale Commissioning and Operational Experience with T2-T2 Data Transfer Links in CMS”, this Conference
[19] D. Spiga, “Running the CMS Computing for the analysis of the rst data, IPRD’08, Siena, Italy, Jun. 2010