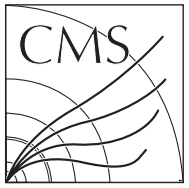
Available on CMS information server

CMS CR -2009/094



The Compact Muon Solenoid Experiment

Conference Report

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

14 May 2009

Operational Experience with CMS Tier-2 Sites

I. González Caballero for the CMS Collaboration

Abstract

In the CMS computing model, more than one third of the computing resources are located at Tier-2 sites, which are distributed across the countries in the collaboration. These sites are the primary platform for user analyses; they host datasets that are created at Tier-1 sites, and users from all CMS institutes submit analysis jobs that run on those data through grid interfaces. They are also the primary resource for the production of large simulation samples for general use in the experiment. As a result, Tier-2 sites have an interesting mix of organized experiment-controlled activities and chaotic user-controlled activities. CMS currently operates about 40 Tier-2 sites in 22 countries, making the sites a far-flung computational and social network. We describe our operational experience with the sites, touching on our achievements, the lessons learned, and the challenges for the future.

Presented at *Computing in High Energy and Nuclear Physics, CHEP2009,21-27/03/2009,Prague,Czech Republic,15/05/2009*

Operational Experience with CMS Tier-2 Sites

I. González Caballero ***for the CMS Collaboration***

Universidad de Oviedo, Dpto. Física
Calvo Sotelo s/n, 33007 – Oviedo, Spain

E-mail: Isidro.Gonzalez.Caballero@cern.ch

Abstract. In the CMS computing model, more than one third of the computing resources are located at Tier-2 sites, which are distributed across the countries in the collaboration. These sites are the primary platform for user analyses; they host datasets that are created at Tier-1 sites, and users from all CMS institutes submit analysis jobs that run on those data through grid interfaces. They are also the primary resource for the production of large simulation samples for general use in the experiment. As a result, Tier-2 sites have an interesting mix of organized experiment-controlled activities and chaotic user-controlled activities. CMS currently operates about 40 Tier-2 sites in 22 countries, making the sites a far-flung computational and social network. We describe our operational experience with the sites, touching on our achievements, the lessons learned, and the challenges for the future.

1. Introduction

We are at the door of an exciting new era in High Energy Physics to be driven by the largest and most ambitious particle accelerator installation ever built: the Large Hadron Collider (LHC). Located on the border between Switzerland and France, at the European Laboratory for Particle Physics (CERN, Geneva), the LHC is expected to resume regular operations before the end of 2009.

The LHC is built in the 27 km long circular tunnel left by the LEP accelerator around 100 meters underground. Two proton beams will circulate in opposite directions guided by 1232 superconducting dipoles. They will collide at a 40 MHz rate and at a center-of-mass energy of 14 TeV when nominal operation is reached. Four main particle detectors (ALICE, ATLAS, CMS and LHC-b) will collect the results of part of those collisions to try to understand the fundamental nature of matter including the search for evidences of new physics.

The CMS (Compact Muon Solenoid) [1] collaboration has built one of two general purpose particle detectors at the LHC. CMS is a large collaborative effort of around 3500 scientists and engineers from more than 180 institutes worldwide.

The different hardware and software triggers will filter the events produced at each collision so that only those whose content is more promising are stored lowering the initial rate to values of the order of a 100 Hz. Even with that tight selection, the amount of data produced at nominal luminosity and energy during a year is expected to get to values above several petabytes.

2. The CMS Computing Model

The unprecedented level of data that needs to be stored, distributed and analyzed in CMS poses a big challenge in the design of the CMS computing model. In order to cope with all the needs of the CMS

detector and scientists, CMS has developed a computing model which is, among other characteristics, distributed, hierarchical and data driven.

The CMS model is built on top of the biggest worldwide computer resource: the World-wide LHC Computing Grid (WLCG) [2], supported by the major grid infrastructures around the world (EGEE [3], NorduGrid [4] and Open Science Grid [5]). These infrastructures provide the collaboration with more than 50 computation and storage sites scattered on the five continents connected through dedicated network links of 1-10 Gbps.

A hierarchical structure with several levels or tiers is used. A unique Tier-0 located at CERN is responsible for the storage of the data directly coming from the detector as well distributing it to the Tier-1 centers. Prompt reconstruction happens also at CERN. Among the duties for the 7 Tier-1's located in America, Europe and Asia are the custodial archiving of reconstructed data, data reprocessing and the distribution of data to the Tier-2 sites.

The CMS Tier-2 centers are expected to provide all the MC simulation that the collaboration may need, as well as the resources required for the CMS users' physics analysis. They should also be able to transfer data to their associated Tier-3's if any. It is worth noting that, by design, the users' physics analysis are driven by the necessities of the physicists and physics groups and, therefore, the resources are used in bursts of activity that are very difficult to plan. On the other hand the MC simulation workflow is centrally coordinated by the experiment, and thus the occupancy of the resources can be systematically filled and the activity scheduled according to their availability.

Given the huge amount of data that is produced and the fact that the CMS computing centers (and thus the storage elements they host) are scattered around the globe, the collaboration computing model is designed to minimize the amount of data that is moved among the sites. Data is transferred and stored in a well structured way and jobs are expected to run on the nodes holding that data. In this context, tools to handle that data and to find where it lays become very important and a special effort has been put by the collaboration in the development of such services.

3. Tier-2 sites in CMS

CMS operates more than 40 Tier-2 sites typically located at universities and research institutes in 22 countries. Tier-2 centers are a very important resource for CMS computing since the sum of all of them accounts for more than 50% of the total CMS computing power and around 40% of the global collaboration disk storage capacity. A more precise accounting of the resources located at the Tier-2's during 2008 and expected to be deployed during 2009 is detailed in Table 1.

Table 1. Total storage and computing resources in CMS compared to those located in the collaboration Tier-2 sites, both during the year 2008 and foreseen for 2009.

	CPU (MSI2k)			Disk storage (PB)		
Year	CMS Total	Tier-2 (absolute)	Tier-2 (percentage)	CMS Total	Tier-2 (absolute)	Tier-2 (percentage)
2008	39.7	22.3	56.1 %	12.7	4.8	37.8 %
2009	44.0	28.0	63.6 %	18.0	7.7	42.3 %

As mentioned above, CMS is running two main workflows on the collaboration Tier-2 sites: MC simulation and user analysis. Both activities require a minimum grid infrastructure to be set at the Tier-2 sites:

- A grid computing cluster with support for the CMS Virtual Organisation. EGEE, OSG or ARC based middleware may be deployed at the sites.
- A storage cluster with any of the technologies available (CASTOR, dCache, DPM, Lustre, GPFS, etc) provided an SRM version 2 frontend is integrated with it. CASTOR, dCache and DPM provide such tool, while other systems are currently using StoRM [6] or BeStMan [7] as the SRMv2 frontend. The space provided should be big enough to handle

the datasets needed for the user analysis together with a smaller amount for the simulated data.

The CMS collaboration has developed tools to improve the efficiency at which data can be transferred and accessed in any site. The Tier-2's contributing to CMS computing are expected to install at least two of this services: PhEDEx [8] and FroNTier [9].

PhEDEx provides the data placement and the file transfer system for the CMS experiment. PhEDEx manages and optimises data transfers connecting sites through SRM and using the FTS [10] service to schedule them. A medium-sized machine needs to be set up as a normal grid User Interface and the PhEDEx software installed following a well defined mechanism (using the APT tools). Updates are released periodically to fix bugs and to provide new functionalities. A set of independent agents take care of data transfers, data consistency checks, data removal and transfer monitoring. Given the broad variety of systems and cases that need to be supported, the number of things to configure for the whole PhEDEx service is large and, therefore, the configuration and optimisation of the service is a complex task. However the great amount of documentation and examples available, together with an important group of motivated developers and users helps setting it initially at the sites.

The FroNTier system provides a Squid based cache system serving conditions data (such as alignment and calibration constants) to the local cluster at the CMS centers. Data is distributed once from central databases to each of the tiers when required so to avoid overloading the central CMS servers. Within each site the FroNTier system distributes the data to all the worker nodes. The current recommendation is to deploy a FroNTier server for every 800 computing slots or so. Most of the Tier-2's are well below that number so only one such machine needs to be set up at them

On top of the previous mentioned services every Tier-2 site implements more or less sophisticated tools to monitor the status of the batch queues, storage disks, network bandwidth, etc.

4. Data handling at a CMS Tier-2

For all the reasons mentioned in section 2, but mainly due to the role that data plays in the CMS computing model, the whole design is very dependent on the way data is transferred. An efficient and flexible data transfer system has been built by the computing project in which a complex topology arises (see Figure 1 for and schematic view).

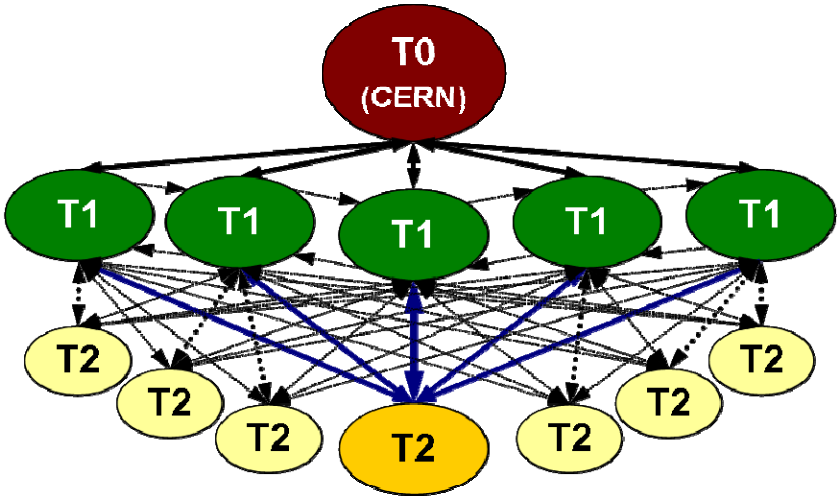


Figure 1. Schematic view of the CMS transfers topology. For simplicity connections to Tier-3's have been omitted.

For every Tier-2 in CMS the concept of an associated Tier-1 refers mainly to the place where first aid with computing problems may be sought, but in the area of data transfers they need to be able to efficiently export and import data to and from any of the Tier-1's in the collaboration. It may be worth noting that, although transfers among Tier-2's are not encouraged, they are allowed. These links are currently being developed. Despite the addition a new level of complexity, they have proved to be very useful when they are set between two Tier-2's associated to the same physics groups by adding a new path by which any of them can get the interesting datasets.

The CMS topology increases the normal complexity in the operation of the Tier-2 network since multiple SRM connections must be managed by the sites. Since the CMS centers are spread all over the world, very different network latencies need to be managed making the tuning of the network parameters a delicate task. Moreover, the different time zones in which they are located affects the speed at which problems are communicated and solved.

To make sure sites behaving badly or with wrong configurations do not affect functioning sites, the CMS Facility Operations area has launched a program to mark every link between two CMS computing centers as either commissioned or not-commissioned. Only commissioned links may be used to transfer real data. Fake data is used in an independent PhEDEx instance for the tests that decide if a link passes the metric so as to be commissioned. In order to be commissioned a Tier-1 to Tier-2 links (downlink) need to show the ability to sustain a transfer rate above 20 MB/s for a day. The minimum requirement for a link in the opposite direction (uplink) is set to 5 MB/s. Links are periodically exercised to check they keep their ability to maintain sustained rates and qualities.

Figure 2 shows the number of downlinks commissioned from any CMS Tier-1 to any CMS Tier-2. As can be seen there, the mesh is almost green since about 85% of those links have been commissioned and kept in that state. This means that data for user analysis can in most cases be transferred very quickly to the Tier-2's. The amount of links commissioned in the opposite direction, needed by the MC simulation workflow, is already above 50%. A special effort, DDT (Debugging Data Transfers) [11], has been put in place by CMS in order to help the CMS sites meet the link commissioning metrics and to improve any aspect of the data transfer model that may arise (for example, reducing the data latency). The status of the link mesh and the improvement in transfer rate and system stability thanks to the DDT team is progressing at a very good pace.

The disk space in the CMS Tier-2 sites is distributed according to the structure in Figure 3. Further details may be found in [12]. The space controlled centrally has little impact on the Tier-2 local operations since it is transparent to the site. Each Tier-2 in CMS is associated with 1 to 3 physics analysis and/or detector groups. A restricted number of persons in each of these groups are responsible to decide which datasets can be stored in the 30 TB assigned to the associated group. PhEDEx keeps track of the ownership of the data in this area, making it easy to follow the correct use of the data at the sites. The Local Space is devoted to the geographically close physics community and each Tier-2 has its own rules. In order to have better control on the way the space at the Tier-2's is managed, CMS created the role of the Data Manager at every site. The Data Manager has to review every transfer or deletion request and, according the site commitments and the situation of the local Storage Element, approve or deny it. This quite consuming activity assures that the space is used efficiently and following CMS rules. Finally, every user in CMS is associated to a Tier-2, usually based on his location and geographical proximity. Around 0.5 to 1 TB of space is reserved for his use. CMS currently provides no mechanism to do the accounting of this area and freedom is given to the sites to manage it as they use, including the use or not of strict quotas.

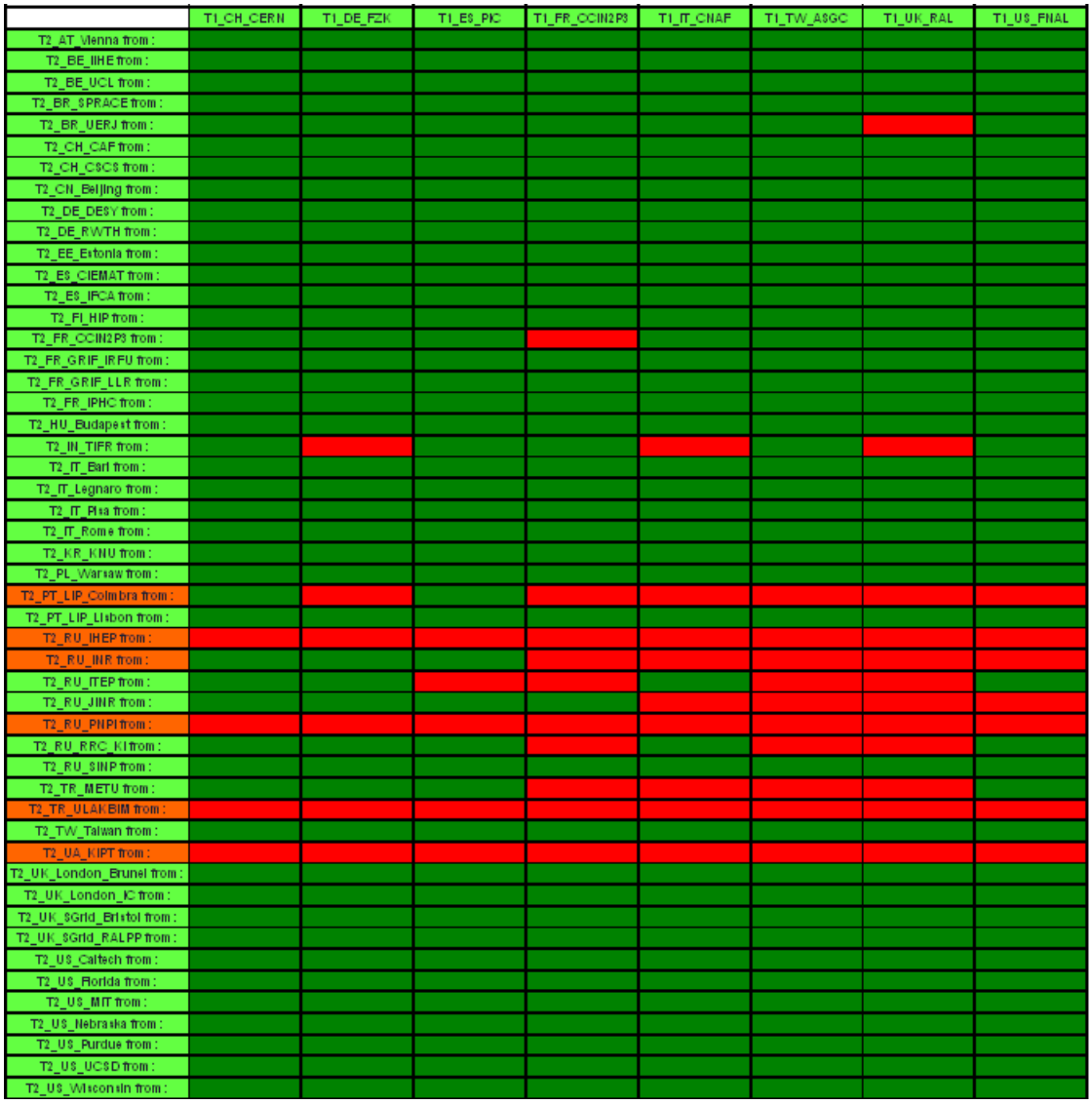


Figure 2. Commissioned (green) and not-commissioned (red) downlinks from the CMS Tier-1's (top row) and Tier-2's (left column).

5. Computing at CMS Tier-2

Though CMS uses standard grid infrastructures to run the physics jobs, it requires special tunings so as to have everything working properly.

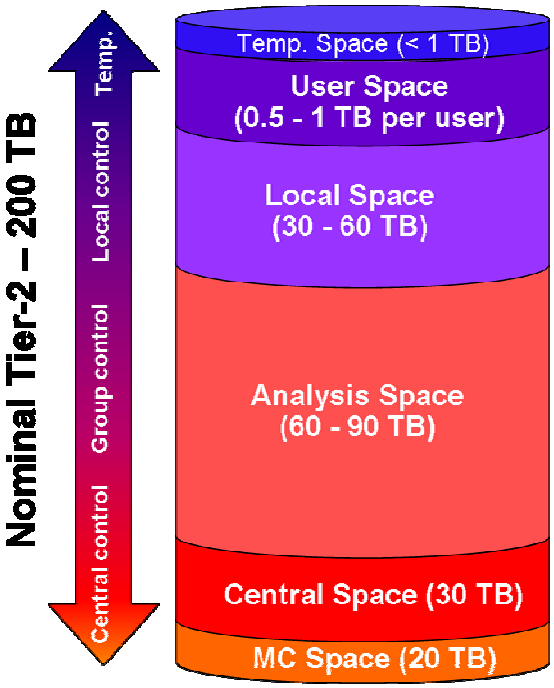


Figure 3. Distribution of the storage space at a nominal CMS Tier-2 (200 TB of total disk space). The bottom data stored in the blocks at the bottom (MC Space and Central Space) is centrally controlled by the experiment. The Analysis Space is managed by the associated physics and analysis groups. The Local and User areas are however handled by the local physicists.

The jobs run by CMS at the sites need the CMS simulation, reconstruction and analysis framework software, CMSSW, to be installed at the computing cluster. The installation of these packages is centralised and done through grid jobs executed under the software manager role. Sites are expected to provide mechanisms to ensure that the software is available for all the worker nodes in the farm. Given that a single installation job per release will run in just one node, the site has to make sure the installation is propagated to the rest of the farm. The usual configuration just sets a software installation area, writable by the software manager role, which is then shared among all the computing nodes in the Tier-2 (via NFS for example). Since the user holding the software manager role might not always be the same further special configurations might be needed. In the past the heavy requirements in terms of memory of the CMSSW installer software forced the site to ensure the installation jobs would arrive at powerful computers. The newer releases have been highly improved and this is no more a concern.

While data is moved among the CMS sites using SRM, more efficient protocols are favored for the jobs accessing local data: POSIX, RFIO, dCache… This is configured at every site through the Trivial File Catalog (TFC), an XML file mapping the logical file names to physical file names. CMSSW jobs look into the TFC to know where local data resides and how to get it.

The CPU share that each of the two workflows running at the Tier-2's take needs to be also configured locally. CMS expects that half of the CPU power at a given site is reserved for MC production, so the site batch queues have to be set accordingly.

Though not a particular requisite of a Tier-2, they often offer one or more User Interfaces to their local community of users. CMS has developed a special tool, CRAB, which can be installed at the User Interfaces to enhance the way users interact with the grid, and to facilitate the job partitioning, submission and retrieval.

6. Central operation of CMS Tier-2's and monitoring

From the CMS central point of view, operating the CMS Tier-2 sites is a complex task, not only due to the big number (more than 40 sites), but also because of the heterogeneity of technologies used and their geographical distribution. Mechanisms to communicate important news, configuration changes, requirements and problems are crucial, and so they have been put in place. One dedicated Tier-2

Hypernews forum exists and several others more tool or service specific are available. Operators and managers at the Tier-2 need to subscribe to those forums where aspects affecting their site tools are discussed. News is continuously communicated using this tool. Problem and bug tracking is mostly achieved through the LCG Savannah portal [13].

At the same time CMS has developed a set of tools and metrics to monitor the sites and to establish their ability to contribute to CMS computing activities. The most relevant of such metrics is the Site Readiness [14] which, based on the number of commissioned links, the results of fake analysis jobs (JobRobot) and the Site Availability Monitoring [15] tests output classifies the sites as ready, not-ready or in warning state (i.e. in danger of becoming not-ready). Through this single value, site operators have a very easy way to evaluate how well the site is behaving in what contributing to CMS is concerned.

Many other tools to monitor almost any aspect of the Tier-2's activities are available for both local and central operators (see for example [16]). The CMS dashboard provides tools to check the status of the analysis and production jobs, and the level of activity happening at any CMS site. PhEDEx implements a complete monitoring system including ways to plot and find, among others, the rate, volume and quality of the transfers, the errors detected and the reasons of those errors, latencies, routing details, etc. Historic charts can be used for accounting purposes and to study the behaviour of a Tier-2 over time to find inefficiency patterns. The plots showing the current state of the services are also very useful to identify and correct problems as soon as possible.

7. Results and Conclusions

CMS has designed a computing model in which the Tier-2's play a key role providing more than one third of the resources. They handle a mixture of centrally controlled and bursty activities supporting two crucial workflows for the collaboration: MC production and user analysis. The user physics analysis requires data to be efficiently and quickly transferred from the Tier-1's. The MC data produced is continuously moved in the opposite direction where it may be redistributed to other CMS centres.

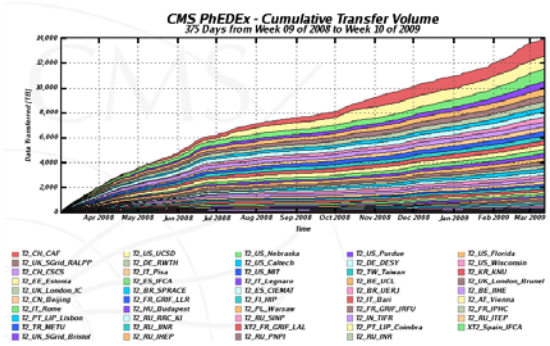


Figure 4. Total data (in TB) transferred to any Tier-2 site in CMS over 12 months from March 2008 to March 2009. Almost 14 PB of data where downloaded during that period. The cumulative graph colours are associated to each of the sites as explained in the legend below it.

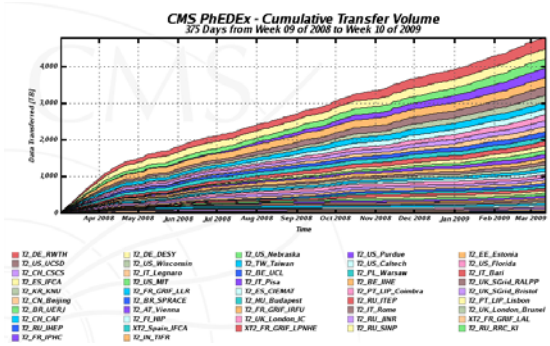


Figure 5. Total data (in TB) transferred from any Tier-2 site in CMS over 12 months from March 2008 to March 2009. More than 4.5 PB of data where uploaded during that period. The cumulative graph colours are associated to each of the sites as explained in the legend below it.

CMS has developed PhEDEx to manage both data flows in the complex topology developed by the collaboration. As can be seen in Figure 4 an aggregated volume of almost 14 PB of data was transferred to the CMS Tier-2 sites using PhEDEx while more than 4.5 PB of data was exported from the Tier-2's as shown in Figure 5.

More than 2 billion events have been processed by the simulation and reconstruction CMS software in the Tier-2's as can be seen in Figure 6.

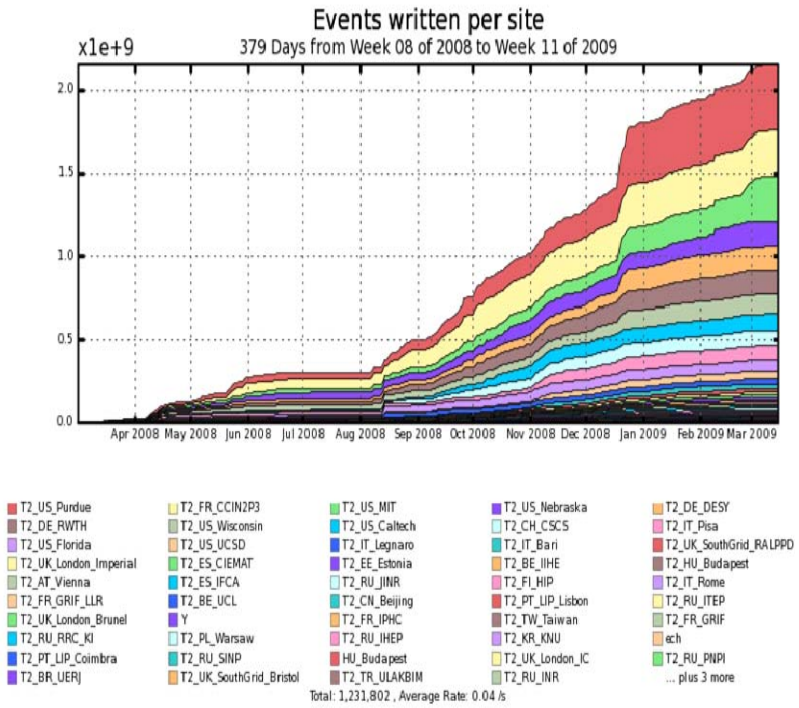


Figure 6. Number of events produced (simulated and reconstructed) in the CMS Tier-2 sites during the last 12 months.

CMS users are intensively using the resources at the Tier-2 sites [17]. Almost 9 million user analysis jobs have been run in the last year as shown in Figure 7. It is worth noting that these jobs have not only on MC produced samples, but also on real data taken during several cosmic runs exercised with and without magnetic field. The overall application efficiency is above 60% and goes down to 55% if grid failures are taken into account.

Given that the nature of the CMS computing model is very much dependent on the correct handling and placement of the data, CMS has built tools to efficiently move and locate physics samples. These tools have been deployed to the Tier-2 sites and have proved to be able to cope with the requisites of a demanding environment. A Data Manager appointed at every site links CMS central data operations with the local management.

The CMS collaboration has established metrics to validate the availability and readiness of the Tier-2's to contribute efficiently to the collaboration computing needs by verifying the ability to transfer and analyze data. At the same time, CMS has set up specialized teams to help sites finding the solution to the problems that may appear and meeting the level CMS requires in its services. A big number of tools have been developed by CMS and CERN IT division to monitor every aspect of a Tier-2 in order to better identify and correct the problems in the day by day operations.

CMS Tier-2 sites have proved to be already well prepared for massive data MC production, dynamic data transfer and efficient data serving to local clusters. Moreover they provide CMS physicists with the infrastructure and the computing power to perform their studies and analysis fast and reliably. Several physics papers have been published based on the analysis done on MC data stored and generated at Tier-2 sites using these resources.

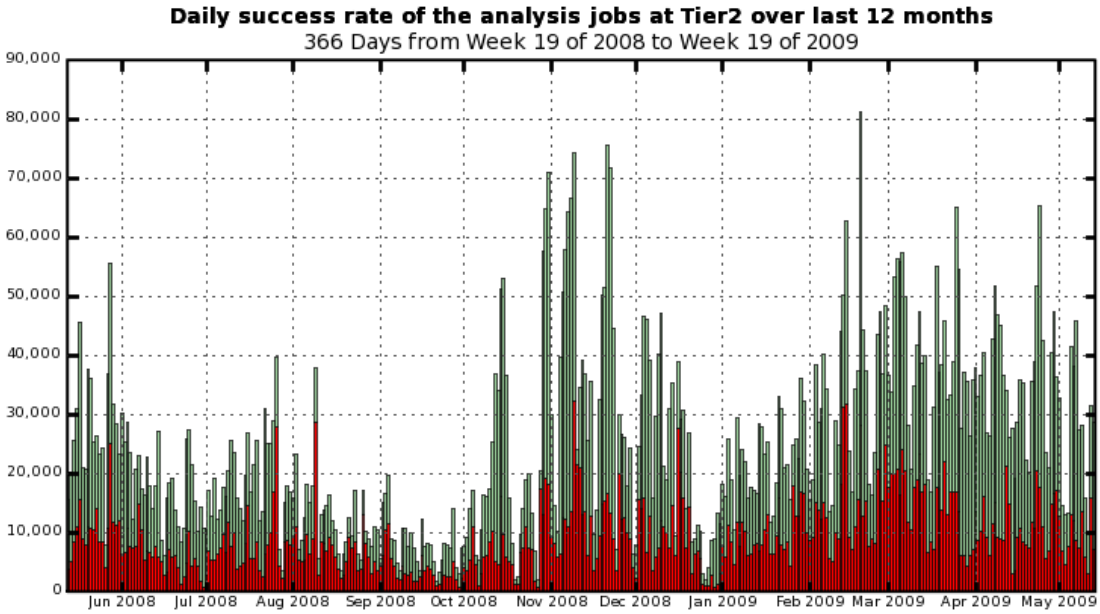


Figure 7. User analysis jobs executed at all the CMS Tier-2 sites over the last 12 months. Successful jobs are shown in green while application failed jobs are rendered in red. Almost 9 million jobs were run by CMS users.

Acknowledgments

We thank the technical and administrative staff at CERN and other CMS Institutes, and acknowledge support from: FMSR (Austria); FNRS and FWO (Belgium); CNPq, CAPES, FAPERJ and FAPESP (Brazil); MES (Bulgaria); CERN; CAS, MST and NSFC (China); MST (Croatia); RPF (Cyprus); Academy of Sciences and NICPB (Estonia); Academy of Finland, ME and HIP (Finland); CEA and CNRS/IN2P3 (France); BMBF, DFG and HGF (Germany); GSRT and Leventis Foundation (Greece); OTKA and NKTH (Hungary); DAE and DST (India); IPM (Iran); SFI (Ireland); INFN (Italy); KICOS (Korea); CINVESTAV, CONACYT, SEP and UASLP-FAI (Mexico); PAEC (Pakistan); SCSR (Poland); FCT (Portugal); JINR (Armenia, Belarus, Georgia, Ukraine, Uzbekistan); MST and MAE (Russia); MSD (Serbia); MCINN and CPAN (Spain); Swiss Funding Agencies (Switzerland); NSC (Taipei); TUBITAK and TAEK (Turkey); STFC (United Kingdom); DOE and NSF (USA).

References

[1] CMS Collaboration 1994 *The Compact Muon Solenoid Technical Proposal* (CERN/LHCC 94-38)

[2] Worldwide LHC Computing Grid, http://lcg.web.cern.ch/LCG/

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

[4] Nordic Grid facility, http://www.nordugrid.org/

[5] Open Science Grid, http://www.opensciencegrid.org/

[6] Storage Resource Manager, http://storm.forge.cnaf.infn.it/

[7] Berkeley Storage Manager, http://datagrid.lbl.gov/bestman/

[8] Rehn J et al 2006 PhEDEx high-throughput data transfer management system *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics, CHEP06* (Mumbai, India)

Egeland R 2009 PhEDEx Data Service *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics, CHEP09* (Prague, Czech Republic)

[9] Blumenfeld B, Dykstra D, Lueking L and Wicklund E 2007 *CMS Conditions Data Access using FroNTier* FERMILAB-CONF-07-526-CD

See also http://frontier.cern.ch/

[10] File Transfer Service, http://egee-jra1-dm.web.cern.ch/egee-jra1-dm/FTS/
[11] Letts J 2009 Debugging Data Transfers in CMS *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics, CHEP09* (Prague, Czech Republic)
[12] Kress T 2009 CMS Tier-2 resource management *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics, CHEP09* (Prague, Czech Republic)
[13] See https://savannah.cern.ch/userguide/
[14] Flix J 2009 The commissioning of CMS sites: improving the site reliability *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics, CHEP09* (Prague, Czech Republic)
[15] Duarte A, Nyczyk P, Retico A and Vicinanza D 2008 Testing and integrating the WLCG/EGEE middleware in the LHC computing *J. Phys.: Conf. Ser. 119 (2008) 062020*
[16] Saiz P 2009 Generic monitoring solutions for LHC site commissioning activity and LHC computing shifts *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics, CHEP09* (Prague, Czech Republic)
[17] Letts J 2009 CMS analysis operations *Proc. Int. Conf. on Computing in High Energy and Nuclear Physics, CHEP09* (Prague, Czech Republic)