Experience Building and Operating the CMS Tier-1 Computing Centres

M.Albert[1], J.Bakken[1], D.Bonacorsi[2], C.Brew[4], C.Charlot[5], Chih-Hao Huang[1], D.Colling[6], C.Dumitrescu[1], D.Fagan[1], F.Fassi[6], I.Fisk[1], J.Flix[7], L.Giacchetti[1], G.Gomez-Ceballos[8], S.Gowdy[12], C.Grandi[2] ,O.Gutsche[1], K.Hahn[8], B.Holzman[1], J.Jackson[4,13], P.Kreuzer[9], C.M.Kuo[14], D.Mason[1], N.Pukhaeva[6], G.Qin[10], G.Quast[11], P.Rossman[1], A.Sartirana[2,3], A.Scheurer[11], G.Schott[11], J.Shih[10], P.Tader[1], R.Thompson[1], A.Tiradani[1], A.Trunov[11]

[1]FNAL, [2]INFN-Bologna and Univ. of Bologna, [3]LLR., [4]STFC – RAL, [5]E.Polyt., [6]Imperial College, [6]CC-IN2P3, [7]PIC and CIEMAT, [8]MIT, [9]RWTH Aachen IIIa, [10]ASGC, [11]GridKa/FZK and Univ. of Karlsruhe, [12]CERN, [13]Univ. of Bristol, [14]NCU

Claudio.Grandi@cern.ch

Abstract. The CMS Collaboration relies on 7 globally distributed Tier-1 computing centres located at large universities and national laboratories for a second custodial copy of the CMS RAW data and primary copy of the simulated data, data serving capacity to Tier-2 centres for analysis, and the bulk of the reprocessing and event selection capacity in the experiment. The Tier-1 sites have a challenging role in CMS because they are expected to ingest and archive data from both CERN and regional Tier-2 centres, while they export data to a global mesh of Tier-2s at rates comparable to the raw export data rate from CERN. The combined capacity of the Tier-1 centres is more than twice the resources located at CERN and efficiently utilizing this large distributed resources represents a challenge. In this article we will discuss the experience building, operating, and utilizing the CMS Tier-1 computing centres. We will summarize the facility challenges at the Tier-1s including the stable operations of CMS services, the ability to scale to large numbers of processing requests and large volumes of data, and the ability to provide custodial storage and high performance data serving. We will also present the operations experience utilizing the distributed Tier-1 centres from a distance: transferring data, submitting data serving requests, and submitting batch processing requests.

1. Introduction

The CMS Computing Model [1] makes use of the hierarchy of computing Tiers as has been proposed in the MONARC Project [2]. The resources are made available to CMS through the Worldwide LHC Computing Grid Project (WLCG) [3].

The Tier-0 is located at CERN where the experiment data are recorded. The Tier-0 has the responsibility to archive on tape a first copy of the detector (RAW) data, provide the prompt first pass of reconstruction, and dispatch the data to the Tier-1 centres.

The Tier-1 centres assure the distributed custodial storage of a fraction of the RAW data and for the simulated data produced at the connected Tier-2 centres; provide computing resources for their further re-processing (re-reconstruction, skimming, etc…) and for high priority analysis; control the data transfer to the Tier-2 centres for analysis.

© 2010 IOP Publishing Ltd

The Tier-2 centres provide CPU and disk resources for group and user analysis and for data simulation.

2. Role of Tier-1 centres in CMS

According to the CMS Computing Model [1] and the CMS Computing Technical Design Report [4] Tier-1 regional centres have aspects of custodial data storage, re-reconstruction, data analysis and are also responsible for serving data to Tier-2s for analysis, MC storage and user support. Each Tier-1 centre has the following roles in CMS:

- Securing, and making available to the collaboration, a second copy of a share of the RAW data and reconstructed RECO data
- Receiving and making available a copy of the full CMS Analysis Object Data (AOD) data-set.
- Participating, with the Tier-0, to the timely calibration and feedback to the running experiment.
- Running large scale Processed Datasets skims and selected reprocessing for analysis groups and individuals of CMS.
- Serving data-sets to the Tier-2 and other regional or institute computing facilities
- Securing and distributing Monte Carlo simulated samples produced in the Tier-2 and other centres.
- Running production reprocessing passes of Primary Datasets and Monte Carlo Samples.

3. The CMS Tier-1 centres

There are seven Tier-1 centres supporting CMS. In addition at CERN there are disk and tape resources to support the activities of the Russian Tier-2 centres. Their main characteristics are shown in Table 1.

Site	Location	Storage type	Tape (TB)	Disk (TB)	Batch system	Job slots
CERN	Switzerland	Castor	400	30	-	-
FZK	Germany	dCache/TSM	900	600	PBSPro	850
PIC	Spain	dCache/Enstore	1614	584	Torque/Maui	390
CCIN2P3	France	dCache/HPSS	1650	850	BQS	1100
CNAF	Italy	Castor + Storm	681	525	LSF	650
ASGC	Taiwan	Castor	585	675	Torque/Maui	765
RAL	UK	Castor	600	650	Torque/Maui	545
FNAL	USA	dCache/Enstore	4700	2000	Condor	5000

Table 1: Characteristics of CMS Tier-1 centres in Q2/2009

4. CMS Tier-1 centres operations

4.1. CCRC08

During the first months of 2008 CMS took part to the WLCG Combined Computing Resource Challenge 2008 (CCRC08). During that activity several aspects of the CMS computing system have been tested. In particular the AOD distribution among Tier-1 centres, the reprocessing of the data at full capacity and the data recall from tape to the disk buffer were tested.

For what concerns the Tier-1 activities several goals have been achieved. Transfer from the Tier-0 to the Tier-1 centres reached a sustained rate of 600 MB/s as daily average during a period of more than seven days in a row, with enough headroom and hourly peaks up to 1.7 GB/s. The latency for AOD replication among Tier-1 centres has been verified and all centres reached the prefixed target. About 90% of the transfer links between Tier-1 and Tier-2 centres were tested (178 out of 193 links corresponding to about 2/3 of the full mesh). Peak rates up to 38 times the target were achieved. About 200 thousand reprocessing jobs and 15 thousand skimming jobs have been executed at the Tier-1 centres. This helped finding the limitations and the bottlenecks of the different centres and allowed

addressing them in the following months. Finally the data recall from tape to the disk buffer has been tested. This last test is particularly important for CMS as the model foresees only a "T1D0" system at Tier-1s. This means that data are automatically migrated to tape when they arrive on the storage and are automatically recalled from tape to the disk buffer when requested by a job running on the local farm or when requested for transfer to another site.
At each site one or more 10 TB datasets were selected among those present on the local storage. All the files were removed from the disk buffer then a stage-in request was manually submitted to the system by he local site administrators. The time to submit the request and the time to retrieve the data on tape were measured at all sites. The results are in Table 2. Besides the specific results, the test was important to identify potential problems and understand the strategies for optimization.

Site	Data (TB)	# Files	# Tapes	Stage req. time	Staging time	Throughput
FZK	10.0	4000	50	180'	27h	90 MB/s
PIC	11.6	4744	38	300'	33h	100 MB/s
CCIN2P3	10.0	11061	68	2'	120h	23 MB/s
CNAF	10.8	7235	426	45'	79h	40 MB/s
ASGC	13.2	5632	360	18'	22h	150 MB/s
RAL	10.5	5376	19	10'	10h	290 MB/s
FNAL	10.0	5736	270	12'	25h	110 MB/s

Table 2: Results of the tape recall test during CCEC08 (Feb. 2008)

CMS will repeat the tape recall test during the Scale Testing for the Experiment Program in 2009 (STEP09). A tool is being developed to trigger tape recalls at sites via a central request of the data operation team. This will become part of the standard procedures for data processing at the Tier-1 centres.

4.2. Cosmic data processing in 2008

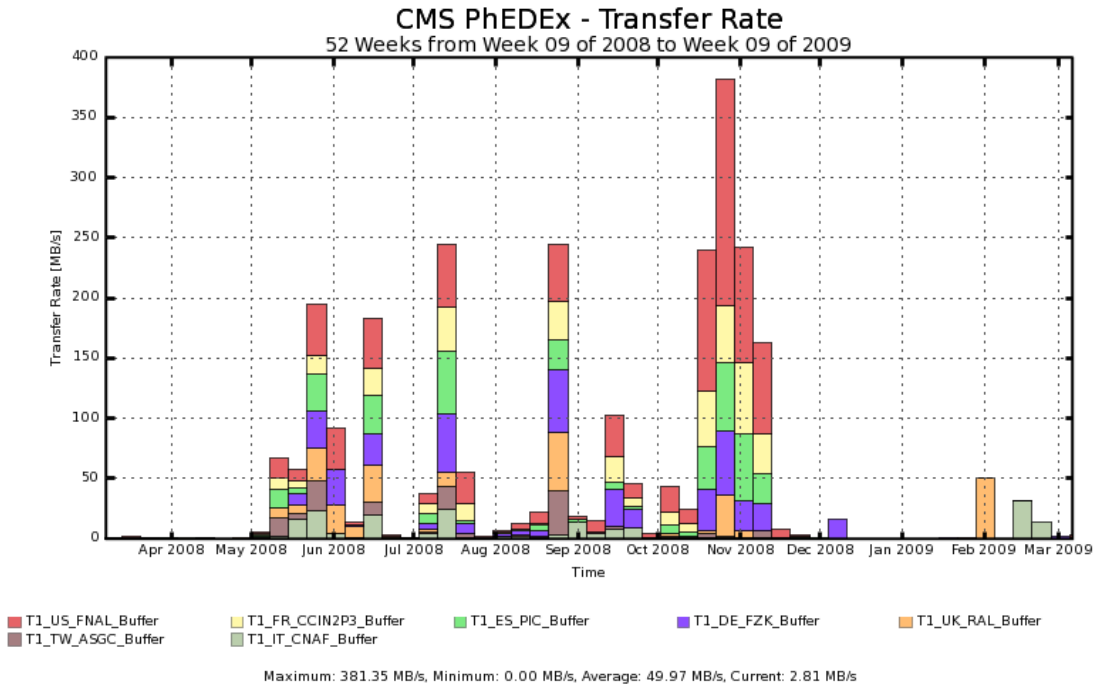

Figure 1: Transfer rates between CERN and the CMS Tier-1 centres

During the 2009 preparation for LHC beams and during the cosmic runs taken after the LHC incident in September, CMS collected real data that are of paramount importance for understanding the detector performance and for its fine tuning. The data went through the full processing chain up to the Tier-2 centres for their analysis.

Figure 1 shows the transfer rates from CERN to each Tier-1 as function of time. Over 1.3 PB were transferred in the period from March 2008 to February 2009. The peaks correspond to the CMS Global Run periods. Daily averages of over 500 MB/s were reached. This is compatible with the expected rates at the beginning of CMS data taking. In particular between 16 October and 11 November 2008 a total of 370 million cosmic triggers were collected with the CMS detector with full magnetic field (CRAFT), fully reconstructed at the Tier-0 and transferred to the Tier-1s.

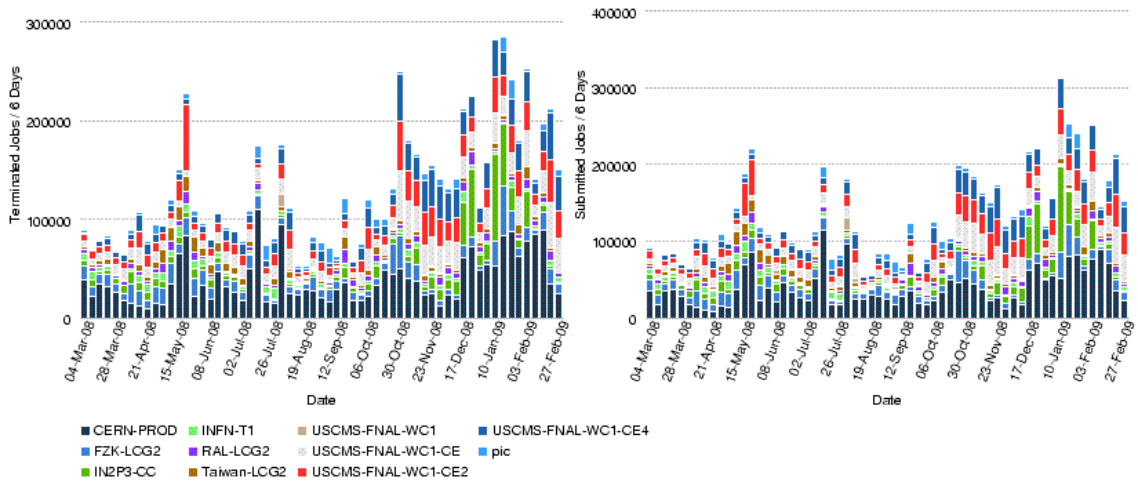

Figure 2: Terminated and submitted jobs in the CMS Tier-0 and Tier-1 centres

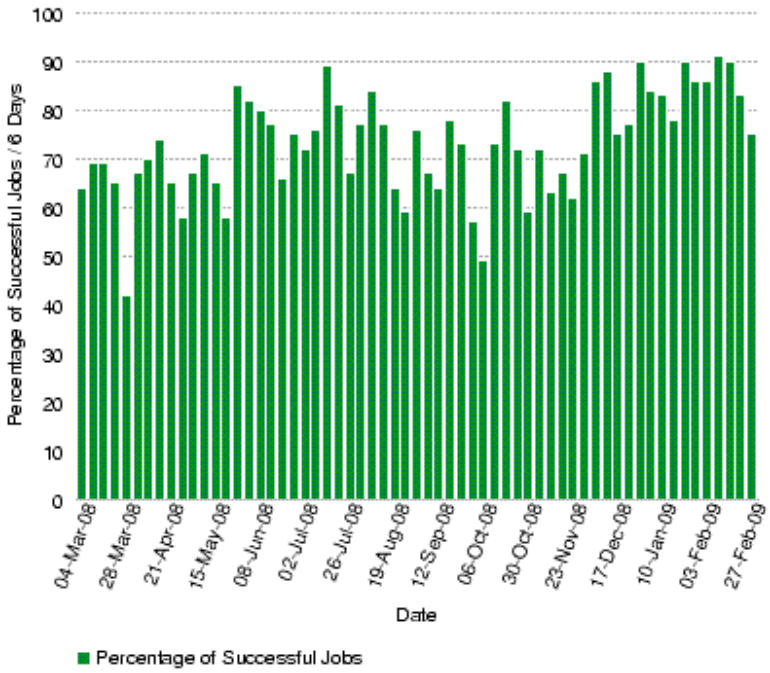

Figure 3: Fraction of successful jobs at the Tier-0 and Tier-1 centres as function of time

Organized data processing activities are carried on at the Tier-1 centres by the Data Operations team. These include re-processing of raw data, production of AOD and large scale physics skims and processing. Figure 2 shows the number of jobs submitted and terminated at the Tier-0 and Tier-1 centres as function of time. During the last months about 4000 jobs per day were executed in average, demonstrating the capability of the framework to exploit all the available resources. The percentage of successful jobs has been increasing during the last year and reached about 90% in certain periods (unsuccessful jobs include both application and infrastructure failures) as it can be seen in Figure 3.

Apart from the job failure rate, the main reason of inefficiency in the utilization of the Tier-1 centres and of instability of their infrastructure is related to the access to the mass storage and in particular to the recall of data from tape that is currently triggered by the process needing the data, i.e. the processing job or the transfer request. The impact of this non-optimized procedure will be evaluated during the abovementioned STEP09 period and eventually procedures to automatically trigger a controlled pre-stage of tapes will be introduced in the normal operations.

4.3. Support to Simulation Operations

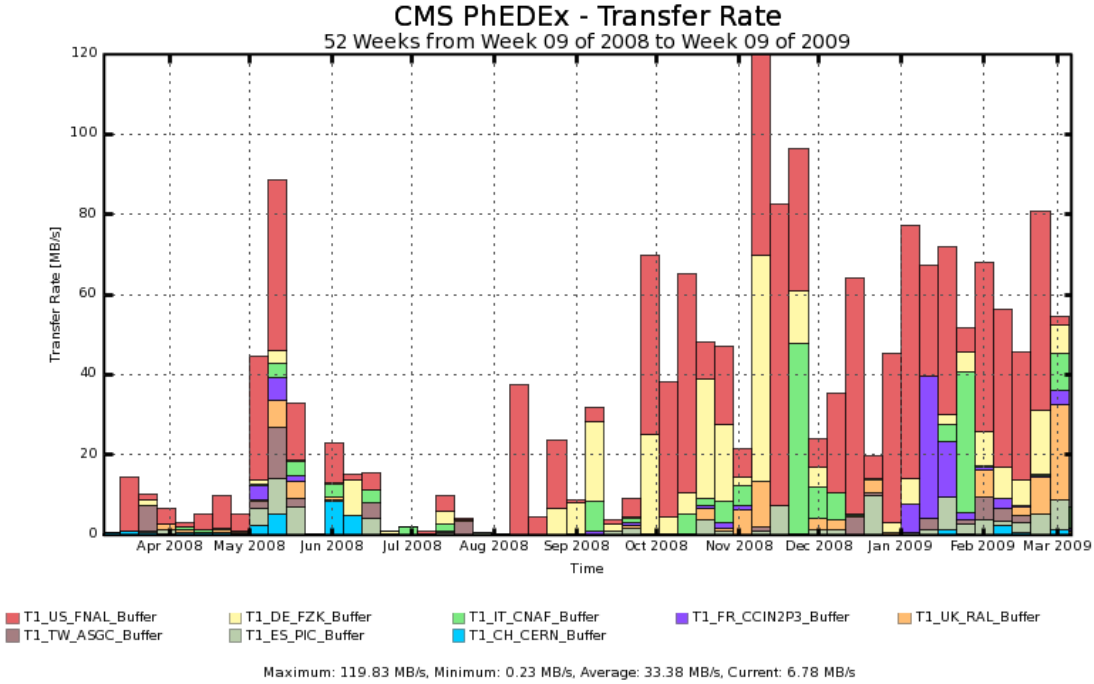

Figure 4: Transfer rates between Tier-2 and the CMS Tier-1 centres

Simulated data are produced by the Tier-2 centres. After the simulation step the data are transferred to a Tier-1 for custodial and reprocessing. Figure 4 shows the transfer rates from Tier-2s to each Tier-1. The requirements for Tier2-Tier1 transfers are less demanding, nevertheless about 1.1 PB were transferred from the Tier-2 centres to the seven Tier-1 centres in the period from March 2008 to February 2009 and daily averages of over 250 MB/s (integrated over the Tier-1 centres) were reached. This is compatible with the expected performances.

4.4. Support to analysis

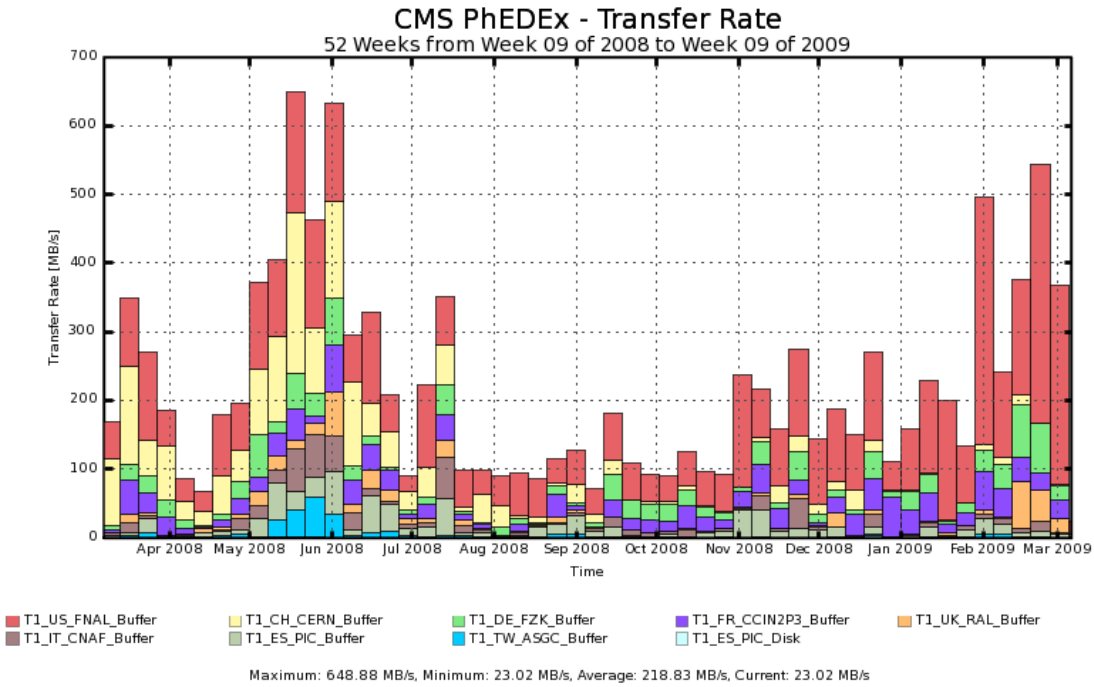

Figure 5: Transfer rates between Tier-1 and other CMS centres

Reprocessed real and simulated data are made available to all CMS sites for analysis. Figure 5 shows the transfer rates from each Tier-1 to all other CMS centres. Over 6 PB were served by the seven Tier-1 centres the period from March 2008 to February 2009 and daily averages of over 1.3 GB/s were reached (integrated over the Tier-1 centres). At the current scale we didn't observe saturation effects but this rate will need to be substantially increased during normal CMS operations.

Tier-1 centres provide a very controlled running environment where the production team carries out its operation. The only other users allowed to process data at the Tier-1s are selected physics group members that perform high priority analysis (e.g. Data Quality Monitoring and Calibrations). Access policies are enforced via VOMS [5] credentials.

5. Tier-1 monitoring

Sites are constantly monitored and their commissioning status is used by the Data Operation team to best schedule transfers and processing activities. In particular the CMS Dashboard [6] provides a source of information on the status of all the components of the CMS computing infrastructure. In figure 6 is shown a "map" of all CMS sites from which it is possible to have an overall feeling of the status. Detailed information on each site can be obtained.

In order to improve the site reliability for production activities, CMS defined metrics to determine whether a site is ready for production. The metrics are based on the result of common WLCG tests and CMS specific tests. Failing the metrics for more than 2 days over the last five causes the site to be considered not ready for production. A consecutive period of 2 days with all the metrics satisfied must be achieved before being considered ready again [7].

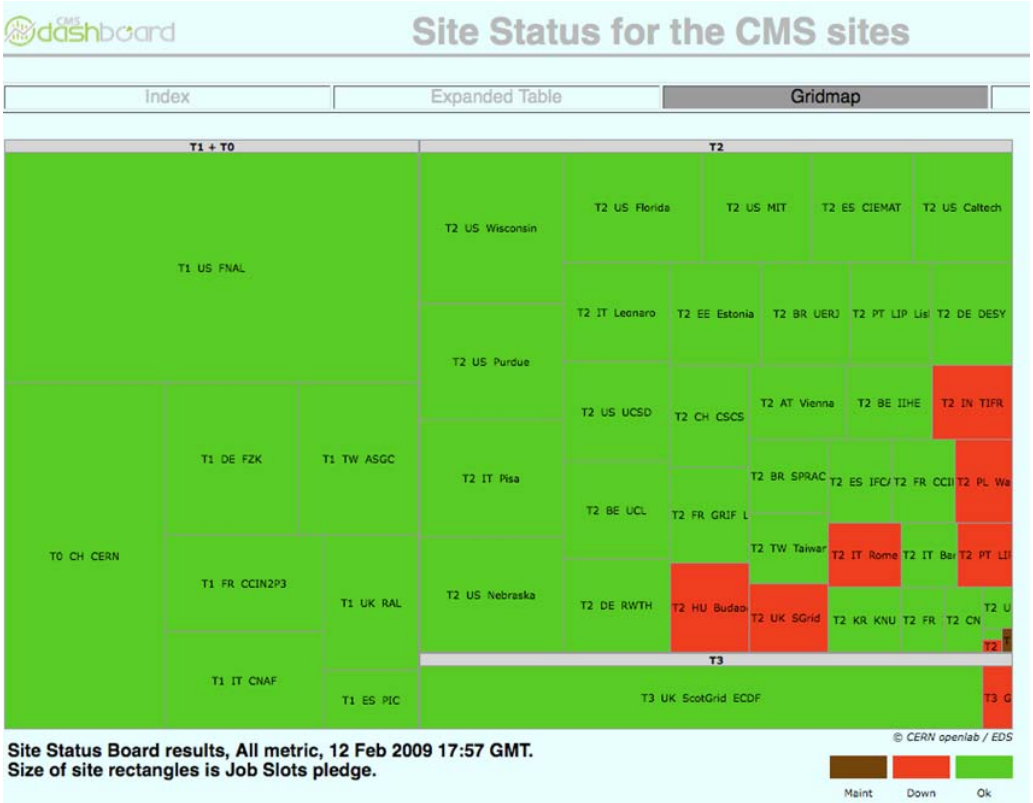

Figure 6: Status of CMS Centres from the CMS Dashboard

6. Summary

During the last year the CMS Tier-1 centres progressively increased their scale and improved their reliability. This has been verified through dedicated facility challenges and most importantly during the day by day activity to support the experiment data taking activities with cosmic events. Data operations and facilities operations procedures are now consolidated. CMS is confident that the infrastructure can run at full capacity for the forthcoming LHC data taking.

References

[1] C.Grandi, D.Stickland, L.Taylor et al. “The CMS Computing Model” CERN-LHCC-2004-035/G-083 (2004)

[2] M. Aderholz et al., “Models of Networked Analysis at Regional Centres for LHC xperiments (MONARC) - Phase 2 Report,” CERN/LCB 2000-001 (2000).

[3] I. Bird et al. “LHC computing Grid. Technical design report” CERN-LHCC-2005-024 (2005)

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

[5] A.Ceccanti et al., “Virtual Organization Management Across Middleware Boundaries”, Proceedings of the International Conference on e-Science and Grid Computing (e-Science 2007), Bangalore, India (2007)

[6] J. Andreeva et al. "Dashboard for the LHC experiments", Proceedings of the Conference on Computing in High Energy and Nuclear Physics (CHEP07), Victoria, Canada (2007)

[7] J. Flix et al., “The Commissioning of CMS Sites: Improving the Site Reliability,” To appear in the Proceedings of the Conference on Computing in High Energy and Nuclear Physics (CHEP09), Prague, Czech Republic (2009)