Journal of Physics: Conference Series **331** (2011) 072020 doi:10.1088/1742-6596/331/7/072020

Monitoring the Readiness and Utilization of the Distributed CMS Computing Facilities

J Flix[1,2], J M Hernández[2], A Sciabà[3]

[1] Port d'Informació Científica (PIC), Bellaterra, Spain
[2] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas (CIEMAT), Madrid, Spain
[3] CERN, Geneva, Switzerland

E-mail: jose.hernandez@ciemat.es

Abstract. The CMS experiment has adopted a computing system where resources are distributed worldwide in more than 100 sites. The operation of the system requires a stable and reliable behavior of the underlying infrastructure. CMS has established procedures to extensively test all relevant aspects of a site and their capability to sustain the various CMS computing workflows at the required scale. The Site Readiness monitoring infrastructure has been instrumental in understanding how the system as a whole was improving towards LHC operations, measuring the reliability of sites when running CMS activities, and providing sites with the information they need to solve eventual problems. This paper reviews the complete automation of the Site Readiness program, with the description of monitoring tools, the impact in improving the overall reliability of the Grid from the point of view of the CMS computing system, as well as the resource utilization and performance seen at the sites during the first year of LHC running.

1. Introduction

The Large Hadron Collider (LHC) at CERN started colliding protons in November 2009. The Compact Muon Solenoid (CMS) [1] is one of the four detectors that record the collisions. It is foreseen to collect few PB of data every year. The large scale data processing and analysis requires exploiting computing and storage resources from outside CERN. The CMS distributed computing system integrates via the Worldwide LHC Computing Grid (WLCG [2]) distributed resources in a tiered structure, consisting of a Tier-0/CAF at CERN, 7 remote Tier-1 sites, 50 Tier-2 sites placed in 23 countries and a growing number of about 50 Tier-3 sites. Tier-1s execute organized data processing workflows while user analysis is run at the Tier-2 sites. The operation of the system requires a stable and reliable behavior of the underlying infrastructure to sustain the various workflows (Monte Carlo simulation, event reconstruction, reprocessing and skimming, data analysis). CMS has established procedures to extensively and continuously test all relevant aspects of a Grid site, such as the ability to efficiently use their network to transfer data, the correct behaviour of all the site services relevant for CMS and the capability to sustain the various CMS computing workflows at the required scale. Monitoring plays an essential role. CMS extensively uses the LHC Experiment Dashboard [3].

2. Site Readiness Monitoring

The Site Readiness Monitoring framework [4] has been developed to assess site readiness. It consists of three main ingredients: i) Site Avalability Monitoring (SAM [5]) tests to check site specific services; ii) Monitoring jobs to test the data processing workflows; iii) Monitoring data transfers to evaluate the capability of replicating data between sites. Site readiness is assessed by means of a set of defined metrics using the results of these tests.

2.1. Site readiness tests

In the WLCG all Grid services are periodically tested using the SAM framework by means of high priority monitoring jobs submitted to the sites every hour. These tests provide a way to measure site availability and reliability. CMS has adopted SAM to run custom tests on the Computing and Storage Elements (CE and SE) instances at the sites. A failure of any of these tests determines the unavailability of the service instance where the test ran. If all instances of a given service type in a site are unavailable, the service itself is considered unavailable. Finally, if either the CE or the SRM service is unavailable, the site itself is considered unavailable.

The capability to run data processing workflows is tested submitting to the sites monitoring jobs similar to the real data processing jobs. A tool called Job Robot was developed for automatic job preparation, submission, tracking, collection and evaluation. Few hundred jobs are submitted daily at regular intervals to the 50+ Tier-1 and Tier-2 sites. In aggregate 25k jobs per day are generated by this activity. The Job Robot daily statistics are used to measure the success rate for each site. The Job Robot tool will be replaced by the HammerCloud system[6].

For sites to be usable, data transfer links need to be operational. CMS created in 2008 the Debugging Data Transfers (DDT) task force [7] with the mandate of defining appropriate metrics, procedures and tools to certify data transfer links and to assist sites in solving eventual problems. The link certification procedure consists of demonstrating the capability of the site to sustain a certain average throughput during 24 hours (20 MB/s for Tier-0 to Tier-1, Tier-1 to Tier-1 and Tier-1 to Tier-2 links, and 5 MB/s for Tier-2 to Tier-1 and Tier-2 to Tier-2 links). More than 2000 links of the CMS full-mesh network transfer topology has been certified.

Data transfer quality is continuously probed at low rate (few kB/s) in all certified links. These monitoring transfers generate about 1 GB/s network traffic CMS-wide. They allow to quickly detect data transfer problems, not only at the network level, but also in the data transfer services and the storage infrastructure.

Figure 1. Snapshot of the CMS Site Status Board monitor.

Table 1. Site Readiness metrics.

Tier-1 sites	Tier-2 sites
daily SAM availability $\geq 90\%$	daily SAM availability$\geq 80\%$
daily Job Robot efficiency $\geq 90\%$	daily Job Robot efficiency $\geq 80\%$
commissioned link from Tier-0	commissioned links to Tier-1s ≥ 2
commissioned links to Tier-2s ≥ 20	commissioned links from Tier-1s ≥ 4
commissioned links from/to other Tier-1s ≥ 4	

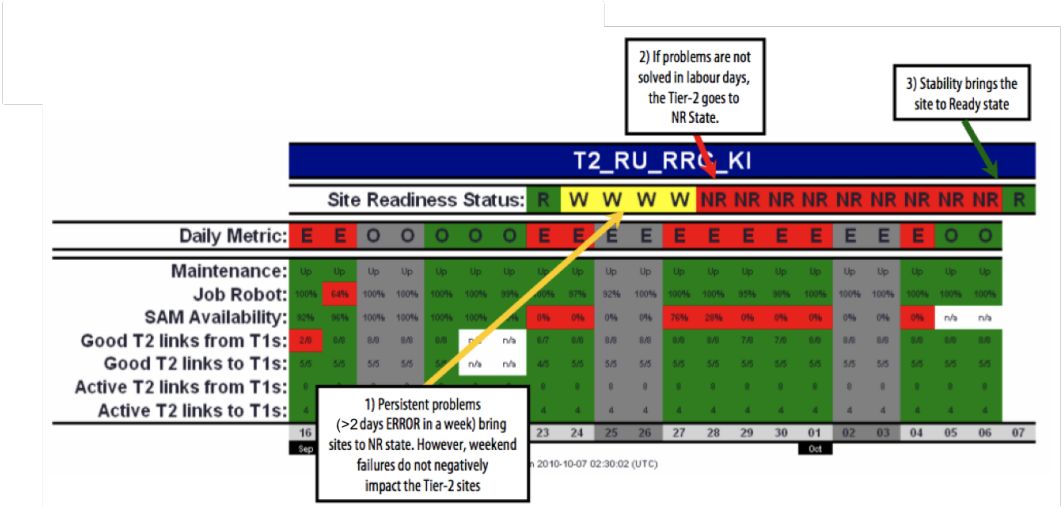

Figure 2. Example of site readiness status tracking.

2.2. Site readiness evaluation

The information of the site readiness tests is collected in the Site Status Board (SSB [8]). It provides a synoptic view of the status of all CMS computing sites and a central point of information for sites and computing shifters to monitor site status. Figure 1 shows a snapshot of the SSB. It provides clickable links to drill down into detailed information.

The site readiness metrics are combined into a single daily *site readiness status* site readiness status. Sites must satisfy all the metrics listed in table 1 for at least 5 days in the previous 7 days to be in the *Ready* status. A failure in any of the metrics transitions the site into the *Warning status*. If the number of days with failing metrics exceeds 2 in the last 7, the site goes into the *Not-Ready* status. To return to the *Ready* status the site must fulfill all the metrics during 2 days. Scheduled downtimes (and weekends for the Tier-2s) are not taken into account in the evaluation of the site readiness status.

Each site is provided with an overall picture of its last 15 days site readiness status and daily metric status, as well as the independent daily measurements. Figure 2 shows an example of such a view for a Tier-2.

Figure 3 shows a historical view of the number of Tier-1 and Tier-2 sites in each readiness status since the start of the site readiness program. A steady increase of the number of good Tier-2 sites is visible, from 15 sites at the beginning of October 2008 to about 40 by April 2009. Since then this number has been kept stable. Tier-1 sites have also improved its readiness with the course of the time although there is still some work needed to reach the level of all Tier-1

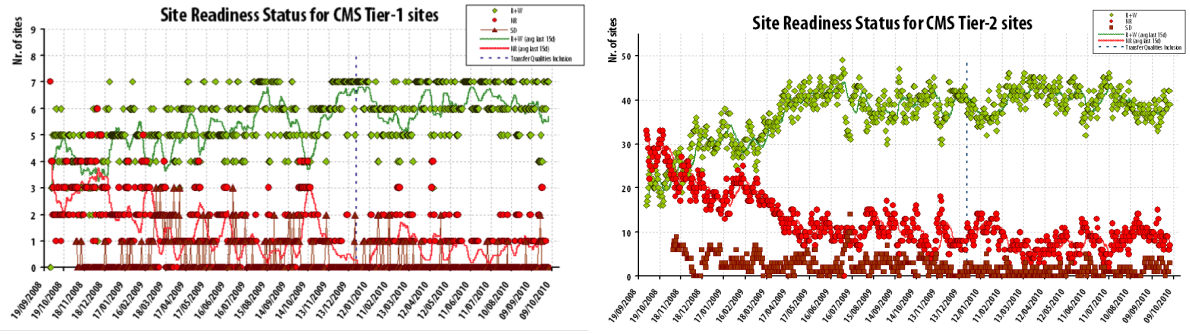

Figure 3. History of number of sites passing/failing the site readiness metrics since October 2008 for Tier-1 sites (left) and for Tier-2s (right).

sites in Ready status for prolongued periods of time. It is apparent that the site readiness program has had very positive effects in the improvement of the reliability of the sites.

The site readiness status history is used to flag good/bad sites. To be usable sites are required to be in Ready status more than 80% of the time in the last 15 days for Tier-2s and 90% for the Tier-1s. The list of bad sites is available to the workload management tools to exclude them from running workflows.

3. Monitoring resource utilization and performance

CMS closely monitors how efficiently computing resources are used in terms of utilization (slot usage, processing share among sites, utilization level relative to the pledged resources) and performance (job success rates and job CPU efficiency). Optimizing efficiency requires identifying and investigating performance issues related to sites, to the Grid or CMS workload management tools, operation problems, imbalance of the data pre-located at the sites with respect to its available compute resources, etc. At the moment CMS is not resource-constrained to process the available amount of data but trying to balance the resource utilization to make the most efficient use of the available resources will pay off in future where a resource-constrained scenario is forseen.

3.1. Tier-1 utilization and performance

At the Tier-1 sites, where organized processing workflows are executed, the slot utilization, processing share and utilization level are closely monitored. Figure 4 shows the slot usage, i.e., the average number of slots occupied running processing, for each of the 7 CMS Tier-1s for the last six months. The site usage is spiky due to the fact that at the moment LHC is not producing collisions continuously. It has been decided in the last months to complement the Tier-1 workflows with Monte Carlo production to make use of the available resources. The processing share, i.e., the fraction of the processing done at each Tier-1, is plotted in figure 5 for the last six months. This plot is quite useful to balance resource utilization. Replicas of the data are available at several Tier-1s so the processing done at each site can be adjusted. In addition, the fraction of the data distributed to each site can be changed to match the available compute resources.

Figure 6 shows the utilization level, i.e., the fraction of the pledged number of slots actually used. As noted before, sites are not yet fully utilized in a continuous way due to the overprovission of resources relative to the available data. Tier-1s are becoming busier in the last months due to Monte Carlo production activities.

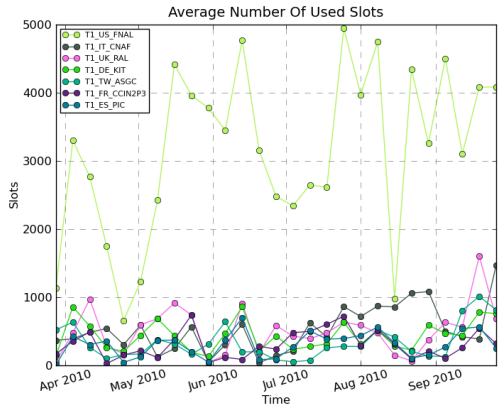

Figure 4. Average number of slots occupied running jobs at the Tier-1 centres.

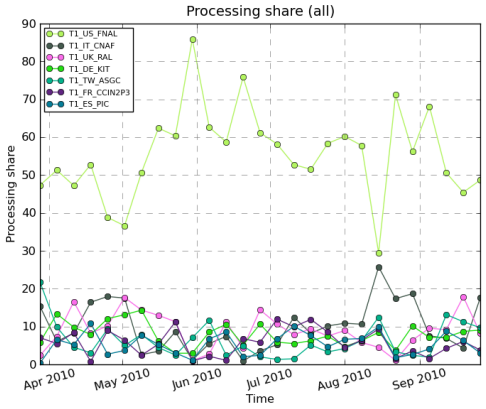

Figure 5. Fraction of the processing done at each of the Tier-1 centres.

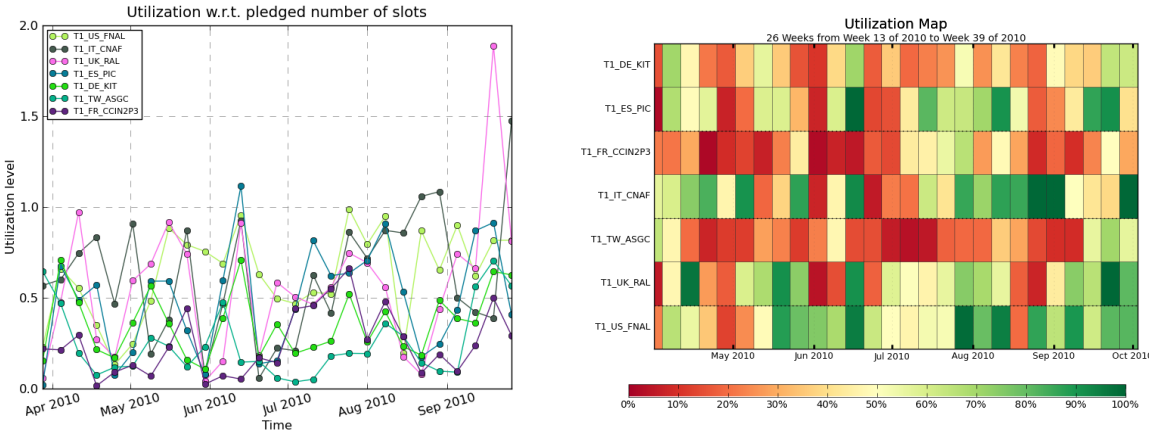

Figure 6. Utilization level at each of the Tier-1 centres.

Job performance is continuously monitored in terms of job success rates and job CPU efficiency (CPU time over wallclock time). In Figure 7 the history of the job success rates for each Tier-1 is shown. The average success rate is about 90%. After a few automatic resubmissions almost all jobs are done. Some of the processing workflows at the Tier-1s, e.g. data skimming or data merging, are I/O bound which reflects in a lower CPU efficiency as can be seen in Figure 8. It is visible in these plots that some sites systematically have a lower CPU efficiency than the others which points to a problem in their data serving system. There is a lot of work ongoing in CMS to improve the CPU efficiency of the processing application (e.g. by optimizing the data layout of files and the reads) as well as identifying and fixing bottlenecks at the sites in the data serving infrastructure.

3.2. Tier-2 utilization and performance

Tier-2 sites are continuously used for analysis activities and the production of simulated data. Figure 9 shows the slot usage for analysis and simulation. In average 6000 slots are continuously occupied with analysis jobs. When simulation production campaigns take place the slots available at the Tier-2 sites are exhausted. All Tier-2s are regularly used for analysis (see

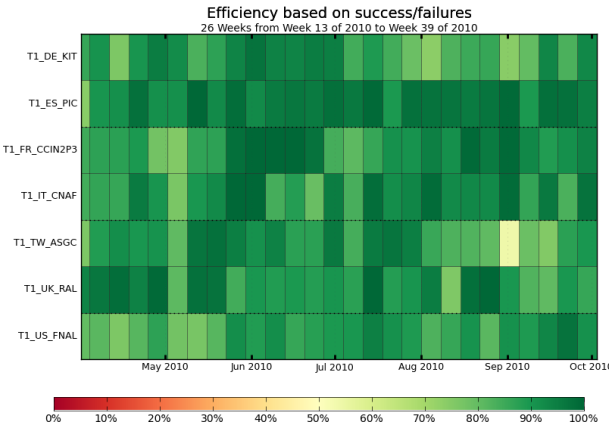

Figure 7. History of job success rates at each of the Tier-1 centres.

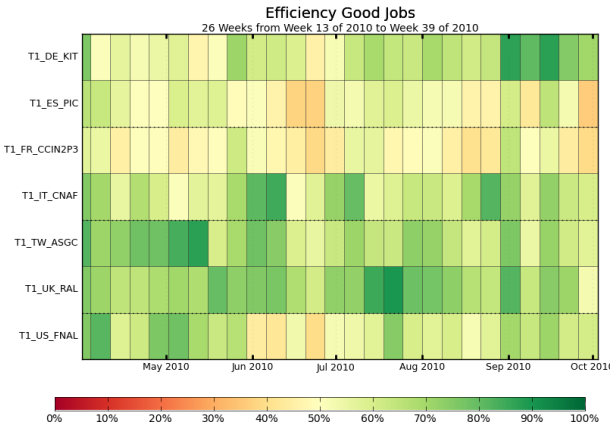

Figure 8. History of the job CPU efficiency at each of the Tier-1 centres.

Figure 10). About half of them contribute significantly (more than 1% of the jobs) while very few sites occasionally run more than 10% of the analysis jobs.

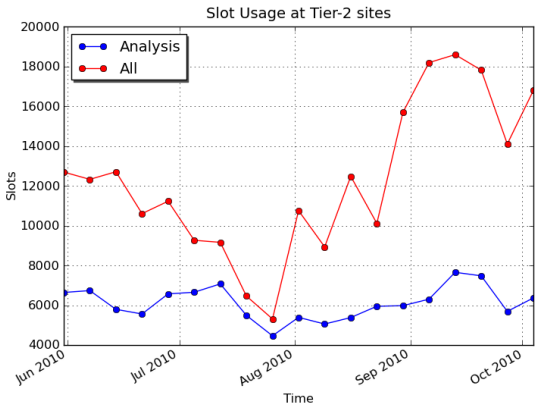

Figure 9. Average number of slots occupied running jobs at the Tier-2 centres.

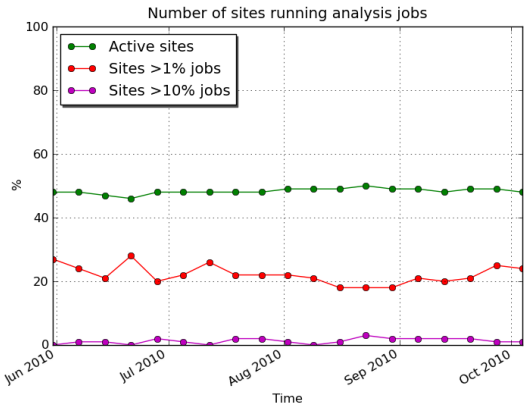

Figure 10. Number of Tier-2 sites running analysis jobs.

The number of distinct users is also significant, as can be seen in figure 11. More than 400 distinct CMS users per week utilize the analysis resources.

Analysis job CPU efficiency is somewhat low, around 55% in average (see figure 12). There is here a potentially large gain with the ongoing improvements in the I/O system of the CMS application framework.

Analysis job success rate is about 70% in average (see figure 13). Only a small fraction of the failures can be attributed to the sites (about 5%). The rest is caused by failures in the application (crash of user code, wrong configuration, failures in the remote stageout of the output data, etc).

It is worth noting that the fraction of jobs aborted by the Grid infrastructure is significant, about 10%, as can be seen in figure 14. This number has not changed much in the last years, so it looks CMS has to live with this level of intrinsic unreliability of the WLCG infrastructure.

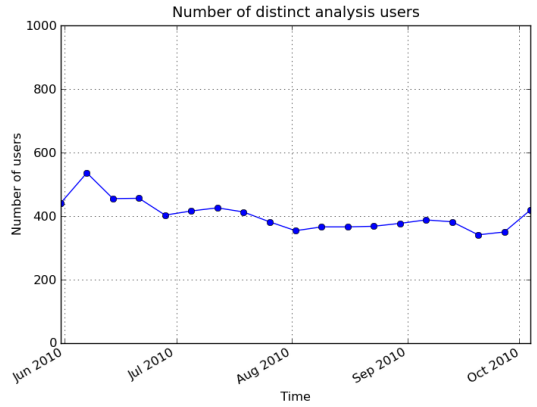

Figure 11. Number of distinct users per week running analysis jobs at the Tier-2 centres.

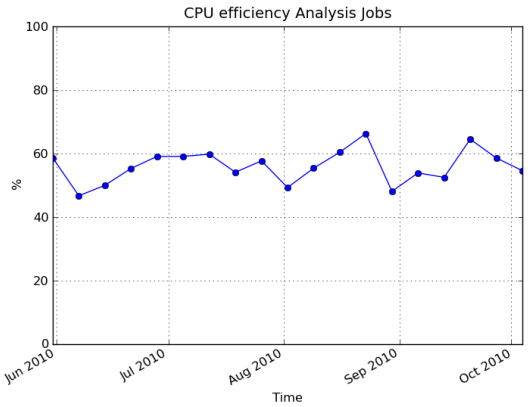

Figure 12. CPU efficiency of analysis jobs at Tier-2 sites.

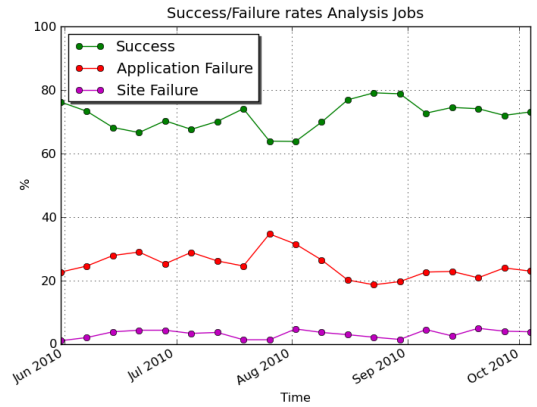

Figure 13. Sucess rate of analysis jobs at the Tier-2 centres.

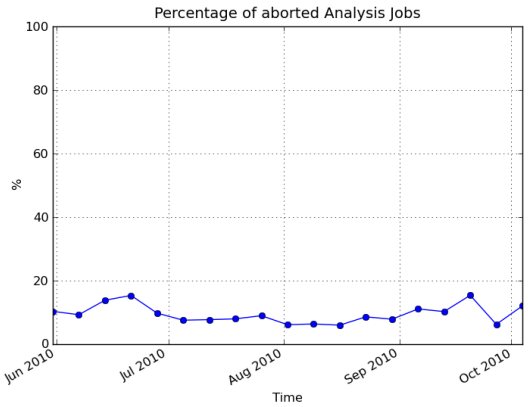

Figure 14. Fraction of analysis jobs aborted by the Grid at Tier-2 sites.

4. Conclusions

Site readiness monitoring has been instrumental in bringing sites into stable and reliable operations. A continuous monitoring of the site readiness to sustain the various CMS computing workflows at the required scale is essential in keeping sites performing efficiently and reliabily.

Monitoring the resource utilization and the execution performance of the CMS workflows has been beneficial in improving its execution efficiency. It has allowed to identify inneficiencies at various levels and to optimize the balance of the utilization of the available computing resources.

References

[1] S Chatrchyan et al, The CMS collaboration, The CMS experiment at the CERN LHC. JINST 3 S08004, 2008.
[2] Worldwide LHC Computing Grid. Technical Design Report. E-ref: http://lcg.web.cern.ch/LCG/TDR/LCG TDR v1 04.pdf
[3] J Andreeva et al., Experiment Dashboard for monitoring of the LHC distributed computing systems, these proceedings.
[4] S Belforte et al, The commissioning of CMS sites: improving the site reliability, J. Phys.: Conf. Ser. 219 062047, 2010

[5] A Duarte, P Nyczyk, A Retico, D Vicinanza, Monitoring the EGEE/WLCG Grid Services, J. Phys.: Conf. Ser. 119 052014, 2008.
[6] D. Van Der Ster et al, HammerCloud: A Stress Testing System for Distributed Analysis, these proceedings.
[7] G Bagliesi et al, Debugging data transfers in CMS, J. Phys.: Conf. Ser. 219 062055, 2010
[8] Dashboard applications to monitor experiment activities at sites, J. Phys.: Conf. Ser. 219 062003, 2010