# Monitoring techniques and alarm procedures for CMS Services and Sites in WLCG

**J Molina-Perez[1,2], D Bonacorsi[3], O Gutsche[4], A Sciabà[1], J Flix[5,6], P Kreuzer[7,1], E Fajardo[8], T Boccali[9], M Klute[10], D Gomes[11], R Kaselis[12], R Du[13], N Magini[1], I Butenas[12] and W Wang[13]**

[1] European Organization for Nuclear Research, Geneva, Switzerland
[2] University of California San Diego, USA
[3] Università di Bologna, Italy
[4] Fermi National Accelerator Laboratory, USA
[5] Port d'Informació Científica, Spain
[6] Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas, Spain
[7] RTWH Aachen University, Germany
[8] Universidad de los Andes, Colombia
[9] INFN Sezione di Pisa, Italy
[10] Massachusetts Institute of Technology, USA
[11] Universidade do Estado do Rio de Janeiro, Brazil
[12] Vilniaus Universitetas, Lithuania
[13] Institute of High Energy Physics, Beijing, China

E-mail: Jorge.Amando.Molina-Perez@cern.ch

**Abstract**. The CMS offline computing system is composed of roughly 80 sites (including most experienced T3s) and a number of central services to distribute, process and analyze data worldwide. A high level of stability and reliability is required from the underlying infrastructure and services, partially covered by local or automated monitoring and alarming systems such as Lemon and SLS; the former collects metrics from sensors installed on computing nodes and triggers alarms when values are out of range, the latter measures the quality of service and warns managers when service is affected. CMS has established computing shift procedures with personnel operating worldwide from remote Computing Centers, under the supervision of the Computing Run Coordinator at CERN. This dedicated 24/7 computing shift personnel is contributing to detect and react timely on any unexpected error and hence ensure that CMS workflows are carried out efficiently and in a sustained manner. Synergy among all the involved actors is exploited to ensure the 24/7 monitoring, alarming and troubleshooting of the CMS computing sites and services. We review the deployment of the monitoring and alarming procedures, and report on the experience gained throughout the first two years of LHC operation. We describe the efficiency of the communication tools employed, the coherent monitoring framework, the proactive alarming systems and the proficient troubleshooting procedures that helped the CMS Computing facilities and infrastructure to operate at high reliability levels.

## 1. Introduction

The Compact Muon Solenoid (CMS) experiment on the Large Hadron Collider at CERN records data on collisions at the energy of 7 TeV since Spring 2010, having increased up to 8 TeV in 2012. There

are 200 institutes involved in the Collaboration and nearly 3500 users spread over five continents. The CMS offline computing system [1] is supporting that large community in order to distribute, process and analyze the detector data.

This requires a high level of reliability together with efficient recovery procedures for problem solving if needed, which will be described in this paper: first by showing the current monitoring and infrastructure framework, then by describing the actors involved and finally by exposing the communication tools used. We show how all these entities interact to ensure the 24/7 monitoring, alarming and troubleshooting of the CMS computing sites and services.

## 2. Monitoring Framework

### *2.1. CMS Central Services*

Monitoring is carried out through the CERN IT LHC Era Monitoring (Lemon) [2] and Service Level Status (SLS) [3] system tools. Lemon focuses mainly on the computing nodes themselves, while SLS checks the reliability of the services running on them.

Lemon is a server/client based monitoring system: an agent is launched on every VOBox (dedicated server for developing or running services that are provided and managed by Virtual Organizations such as ATLAS and CMS) and collects metrics from defined sensors, triggering alarms when values are out of range. These alarms are first handled by CERN IT operators who operate as first line support trying to apply basic recovery routines. If problem persists, alarms are then escalated to system administrators and/or service managers.

SLS provides a web-based display and collects service information in XML format. It dynamically shows availability, basic information and statistics concerning CMS/IT Computing Services, as well as the dependencies among them. These SLS instances are integrated in the CMS Critical Services Gridmap [4] (figure 1). When the service availability is affected, shifters contact the Computing Run Coordinator and the latter applies tested and documented recovery procedures and/or asks for further support to experts.

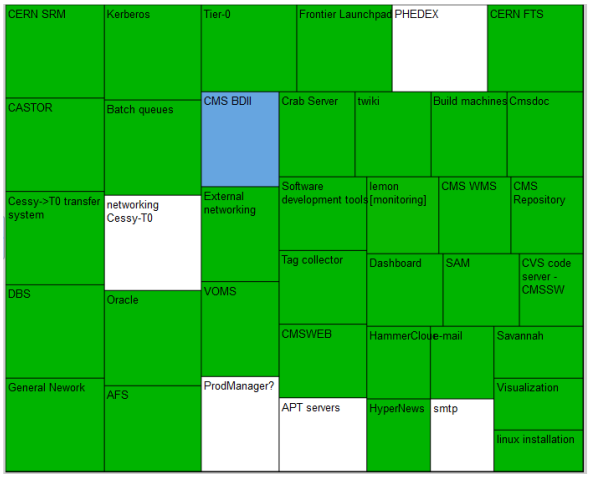


**Figure 1.** CMS Critical Services Gridmap. Green boxes correspond to operational CMS/IT services and change colour when service is affected. Blue/white boxes appear for services not yet tested.

### *2.2. CMS Sites*

The CMS Grid Infrastructure is composed of the Tier-0 at CERN, seven Tier-1, and ~100 Tier-2 and Tier-3 sites (n.b. Tier-3 site support provided on best effort basis) spread throughout the world. Sites

monitoring is performed though the site functional testing tools HammerCloud [5] and Site Availability Monitoring (SAM) [6].

HammerCloud replaces JobRobot [7] and sends automated "analysis-like" jobs to all sites to test their functionality. This tool can also send stress tests on demand using real analysis jobs, which are very useful for site stress testing, CMSSW evaluation and site performance comparison.

Nagios [8], a popular infrastructure monitoring system, has been adopted by CERN IT-GT for WLCG to replace the old SAM client and CMS uses it to run the SAM functional tests.

The Debugging Data Transfers task force [9] defined metrics to test transfers between sites. About 800 production links are continuously exercised.

The prior monitoring sources, as well as the scheduled downtimes and transfer qualities on sites, are used by the CMS Site Readiness [10] to assess the readiness of a site.

Readiness information for all sites is then collected/displayed (as seen on figure 2) through the Site Status Board [11], developed by the CERN IT Dashboard team. It provides automatic notification on site status and allows cross-reference to open tickets.

**Figure 2.** CMS Site Status Board screen capture.

## 3. CMS VOBox Infrastructure

The CMS VOBox Infrastructure is exploited by the CMS Offline-Computing community to develop and run services which cover the whole CMS data handling chain, starting with data recording from detector and ending with the analysis performed on the WLCG. It comprises the VOCMS cluster and the Critical Services Documentation project, and requires a high level of stability and reliability. Synergy between the hardware improvements of nodes used for jobs submission and the effort of Computing Operations make possible to send up to 50,000 parallel production running jobs into the Grid [12].

### *3.1. VOCMS cluster*

This cluster hosts the CMS Central Services, and is composed of about 150 VOBoxes corresponding to 8.3 kHS06. This amount of CPU is expected to decrease due to the progressive virtualization of the cluster (currently, 41% of the nodes are virtual machines). VOBoxes are managed through Quattor (~60 customization profiles) and monitored by Lemon.

Quattor [13] is a system administration software which provides a powerful, portable, and modular set of tools for the automated installation, configuration, and management of medium and large clusters. It allows automated server re-installation from scratch to move a machine into production within a couple of hours in order to replace a faulty computing node.

### *3.2. Critical Services Documentation*

This project is being developed in collaboration with CMS service managers, and intends to list the recovery procedures to be applied by CERN IT operators and/or CMS computing experts. Services are ranked according to their criticality, which is defined depending on the expected response time (see figure 3).

| Criticality Level | Meaning | Response Time (h) |
|---|---|---|
| 10 | CMS stops operating | 0.5 |
| 9 | CMS stops transferring from Cessy | 0.5 |
| 8 | T0 Production stops | 0.5 |
| 7 | T1/T2 Production/analysis stops | 2 |
| 6 | Services Critical when needed but not needed all the time (currently includes documentation) | 4 |
| 5 | A service monitoring or documenting a critical service | 8 |
| 4 | CMS development stops if service unavailable | 12 |
| 3 | Services not critical for CMS | 24 |
| 2 | Services not required for CMS | 72 |
| 1 | Used by an insignificant fraction of CMS | 72 |
| 0 | Not used or dicouraged by CMS | Forever |

**Figure 3.** Definition of criticality levels according to the response time.

The recovery procedures are readable only by the concerned operators and application managers, selected computing experts and the CMS Computing Management, because they contain expert-on-call telephone numbers. After the recovery procedure is validated for a service (figure 4), a new SLS entry is then included in the CMS Critical Services Gridmap.

| Service | Impact | Service Description | Complete |
|---|---|---|---|
| **Data Management** | | | |
| DBS | 8 | DBS used for simulation and reconstruction | Complete: |
| Frontier Launchpad | 8 | Used for serving Oracle information from Conditions into the distributed SQUID infrastructure | Complete: |
| PhEDEx Agent | 8 | Used to support local PhEDEx agents. | Complete: |
| **Workflow Management** | | | |
| T0 Bookkeeping | 9 | Central T0 Submission Service | Complete: |
| Glideinwms | 7 | Used for submitting simulation jobs | Complete: |
| CRAB | 7 | CMS Analysis Submission tool | Complete: |
| **Web Tools Services** | | | |
| Frontend | 8 | Web Server used for WebTools Backend Applications | Complete: |
| Web Tools Base | 3 | Masthead component used by WebTools Backends | Complete: |
| SiteDB | 7 | List of sites, contacts, service names, and resources used by a variety of CMS Services | Complete: |
| PhEDEx Data Service | 8 | Programmatic interface to query PhEDEx Information | Complete: |
| PhEDEx Web | 7 | Web page to identify transfer activity and space used by groups | Complete: |
| PhEDEx Graph | 5 | Web page with phedex graph | Complete: |
| T0 Monitoring | 6 | Tier-0 monitoring system that indicates progress on repacker, reconstruction, and alca-reco | Complete: |

**Figure 4.** Critical Services Documentation: status overview of the main instances.

## 4. Roles for 24/7 Service Operations

### *4.1. Computing Shift Person (CSP)*

Dedicated 24/7 CSP is contributing to detect and react timely on any unexpected error and hence ensure that CMS computing operations are carried out in an efficient and sustained manner. There is a team of ~140 shifters distributed worldwide at ~30 institutions. 3 CSP/day cover the main time zones (Asia, Europe and America), thus there is no need for night shifts. While the amount of CSP manpower (see table 1) is close to optimal in the American Zone, there is a large pool of shifters in

Europe and a lack of personnel in the Asian zone. The latter is very important for 24/7 coverage because there are neither core operators nor developers in that time zone.

CSP are fully integrated in the monitoring/alarming/recovery procedures for the CMS Critical Services and Infrastructure, and have proven to be necessary in many occasions since the LHC startup.

**Table 1.** CSP statistics.

| Time Zone | **Asia** | **Europe** | **America** |
|---|---|---|---|
| Shifters | 27 | 74 | 35 |
| Institutes | 3 | 18 | 9 |
| Operations [a] | 12am-8am | 8am-4pm | 4pm-12am |

[a] These hours of operation correspond to CERN local time.

### *4.2. Computing Run Coordinator (CRC)*

The CRC is a central and essential role for the CMS Computing Operations, having the best overview of the Offline-Computing plan and status, participating in daily run meetings and keeping track of still open computing issues/problems. The CRC supports the CSP and can safely apply recovery procedures on already tested services, escalating alarms to on-call experts when problems persist. There is a pool of ~15 people available for CRC shifts, who are alternatively on duty during one week and play a 24/7 on-call role. Even though physical presence at CERN is preferred, remote shifts are allowed.

### *4.3. Computing Operations Operator*

There is a pool of ~15 operators at CERN who respond as on-call expert during day operations, perform routine service operations, and monitor the CMS Central Services. The most critical services services (Tier-0, Frontier, DBS, Phedex, WMAgent, CMSWEB, CRAB) have dedicated operators, being the Virtual Organization Contact (VOC) the main responsible of the VOCMS cluster management and administration. Among others, activities comprise new service re-installations, Quattor configurations, ports opening in local/central firewall, privileges/access authorization, and handling of security issues.

### *4.4. CMS Local Site Admin/Contact*

There are typically 1-5 people who play the admin role on CMS sites. The bigger the site, the more the admins involved (1-2 enough for a Tier-2, while typically 3-5 for a Tier-1 site). CMS Site Contacts are particularly critical for Tier-1 sites; they conduct overall site troubleshooting issues and address CMS specific operations to Site Admins.

## 5. Communication Tools

### *5.1. Service Now (SNOW)*

Provided by CERN GS, it is a single service desk system [14] with standard processes for all service providers. It is an implementation of best practice (ITIL V3) applied to CERN needs, and is available for incident management and request fulfilment.

SNOW defines a set of persons per experiment (power users) who are able to open tickets to any service at any support level and with any priority. They are able to phone the service desk to request escalation of any existing ticket to a higher level of support. There are ~50 CMS power users composed of the Computing-Offline managers and a number of central service operators and application managers.

### *5.2. CMS Computing Logbook*

This web application is provided by CERN IT and based on GNU open-source ELOG [15]. It is subdivided in several categories according to main CMS Computing workflows, and logbooks are shared among service experts and CSPs. The possibility to use threads and automatic email notifications make this tool very useful for discussion forums.

### *5.3. Global Grid User Support (GGUS)*

GGUS is the main ticketing system for WLCG operations [16]. There are three types of tickets: *User*, *Team* and *Alarm*. A User ticket can be submitted by anyone, whereas Team and Alarm tickets submission requires the appropriate permissions within CMS.

Team tickets are targeted to problems at sites and are co-owned by the CRC, shifters, and a number of CMS experts. Alarm tickets apply uniquely to CERN and Tier-1 sites, and can only be submitted by the CRC and a reduced group of CMS experts. These tickets notify admins about serious problems on their sites at any time, independent from usual office hours, and may trigger 24/7 phone calls. Team tickets might be escalated to Alarm.

### *5.4. Savannah*

This ticketing state based system tool [17] provided by WLCG is subdivided in *trackers* and *squads*. The former refer to the three different issue trackers offered by Savannah (support, bug, task), while the latter are the group of experts to be targeted for solving issues. CMS mapped all Sites' admins/contacts and the tickets are assigned to them efficiently. In addition, the squads are not only related to Sites but also to CMS computing services.

In contrast with the other LHC experiments, the CMS Computing Savannah tracker is equipped with a GGUS bridge, to automatically convert a Savannah item into a GGUS ticket. The bridge forwards the Savannah ticket to GGUS and as long as the case inside GGUS is open, the Savannah ticket stays on-hold and inactive (in order to avoid tracking overlaps).

The Savannah instance at CERN is no longer actively maintained, and might be replaced by open-source Trac or commercial software JIRA.

## 6. Service and Site troubleshooting procedure

Figure 5 shows the global support workflow: the existing correlation among the monitoring framework, all involved actors, and communication tools which preserves a high level of stability and reliability to the CMS Computing Sites and Services.

On the one hand there is the core system alarming (Lemon) at CERN where IT operators act as first line support when an alarm is triggered. When basic routines do not solve the existing issue, alarm is then escalated to CRC and service managers who regularly operate those services.

On the other hand the CSP checks the monitoring status of services and sites. When services/sites suffer degradation, CSP updates the corresponding category in the Computing Logbook so that concerned experts and the CRC are informed.

Depending on the severity of the problem, a different problem tracking tool can be used. Non urgent issues can be handled through Savannah, SNOW and GGUS Team tickets, whereas Tier-0/Tier-1 critical alarms require the escalation of GGUS Team tickets or the creation of GGUS Alarm tickets.

If there isn't any answer from experts within an hour and alarm becomes critical, the CRC might be called in order to either apply the corresponding service recovery procedure or create a GGUS Alarm ticket to the concerned site. If problem persists, then the CRC might phone on-call Experts/Site Contacts to accelerate the troubleshooting process.

The entire process followed is systematically reported in the Computing Logbook for transparency purposes and to learn from experience when facing similar problems, as well as tickets get continuously updated until the solution is found.

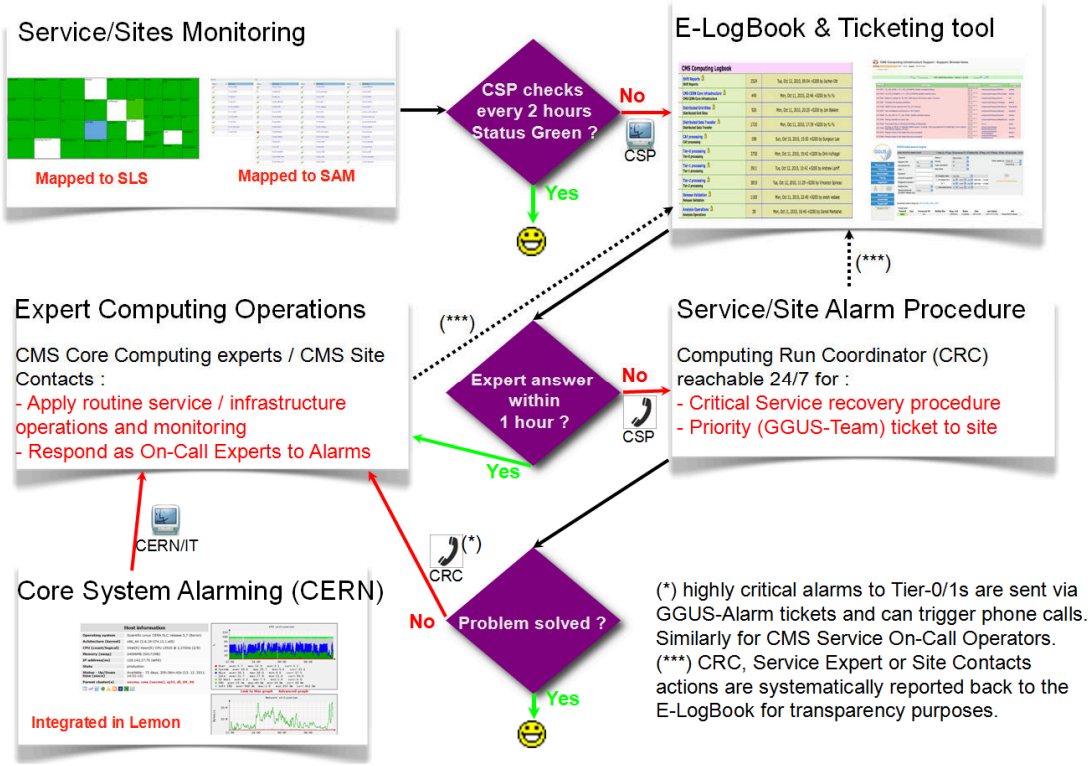


**Figure 5.** Service and Site troubleshooting procedure: the global support workflow.

## 7. Conclusion

Service availability level has been satisfactory since the beginning of LHC operation. Critical service operations are now well integrated into the 24/7 computing shift monitoring/alarming/recovering procedures, which are constantly reviewed and updated.

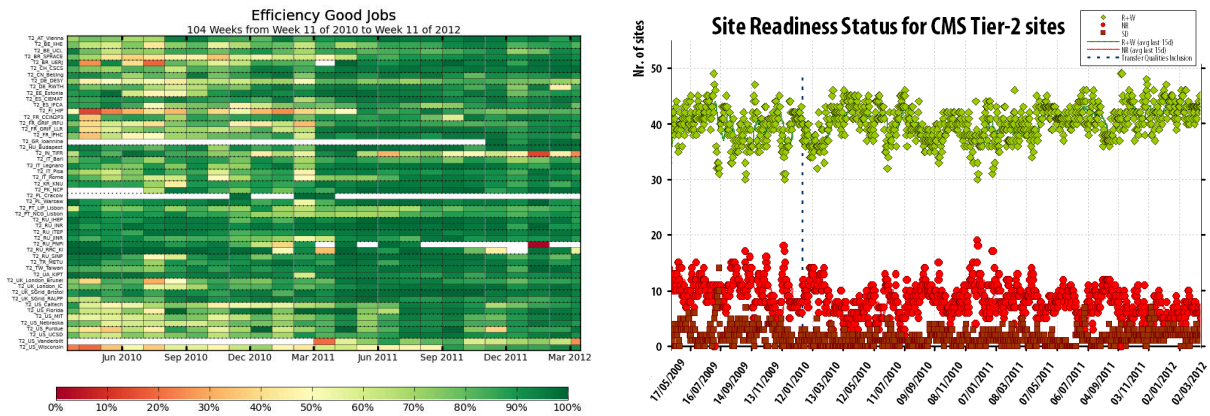


**Figure 6.** Site readiness and jobs efficiency for Tier-2 sites during the last two years.

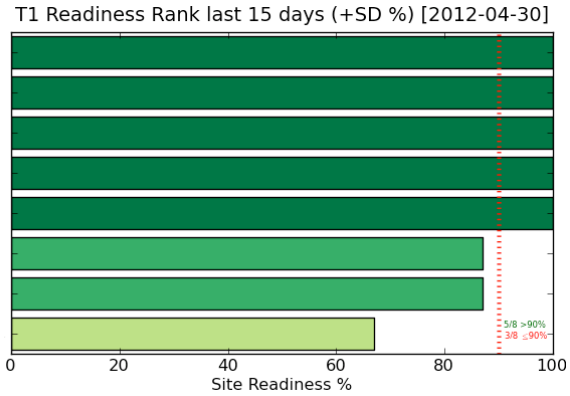


**Figure 7.** Tier-1 sites readiness rank screen capture.

The average of CMS Site Readiness has been around 80% during the last 2 years. The Site Readiness integrated over a 2 weeks period is regularly reviewed and typically reaches the 90 % (80%) level for a large fraction of Tier-1 (2) sites, being used by production teams and users to select reliable sites and thus helping to increase the jobs efficiency (figures 6 and 7).

In addition, further effort is foreseen in the near future to fit CMS computing services into the ongoing major transformation of CERN computing facilities, tools and processes [18].

## References

[1] CMS Collaboration 2005 CMS Computing Technical Design Report *CERN-LHCC-2005-023*
[2] Babik M, Fedorko I, Hook N, Lansdale T H, Lenkes D, Siket M and Waldron D 2011 LEMON - LHC Era Monitoring for Large-Scale Infrastructures *J. Phys.: Conf. Ser.* 331 052025
[3] Lopienski S 2008 Service level status - a new real-time status display for IT services *J. Phys.: Conf. Ser.* 119 052025
[4] http://cms-critical-services.cern.ch/
[5] Van Der Ster D C, Elmsheuser J, Úbeda García M and Paladin M 2011 HammerCloud: A Stress Testing System for Distributed Analysis *J. Phys.: Conf. Ser.* 331 072036
[6] https://twiki.cern.ch/twiki/bin/view/LCG/SAMOverview
[7] https://twiki.cern.ch/twiki/bin/view/CMSPublic/JobRobot
[8] http://wlcg.web.cern.ch/samnagios
[9] Bagliesi G *et al* 2010 Debugging Data Transfers in CMS *J. Phys.: Conf. Ser.* 219 062055
[10] Bagliesi G, Bloom K, Brew C, Flix J, Kreuzer P and Sciabà A 2012 Towards higher reliability of CMS Computing Facilities *J. Phys.: Conf. Ser.*
[11] Saiz P *et al* 2012 Collaborative development. Case study of the development of flexible monitoring applications *J. Phys.: Conf. Ser.*
[12] Fajardo E, Gutsche O, Foulkes S, Linacre J, Spinoso V, Lahiff A, Gomez-Ceballos G and Klute M 2012 A new era for central processing and production in CMS *J. Phys.: Conf. Ser.*
[13] Jouvin M 2008 Quattor: managing (complex) grid sites *J. Phys.: Conf. Ser.* 119 052021
[14] https://cern.service-now.com/service-portal/
[15] http://midas.psi.ch/elog/
[16] https://ggus.eu/pages/home.php
[17] https://savannah.cern.ch/
[18] Andrade P, Bell T, Van Eldik J, McCance G, Panzer-Steindel B, Coelho Dos Santos M, Traylen S and Schwickerath U 2012 Review of CERN Computer Centre Infrastructure *J. Phys.: Conf. Ser.*