Available on CMS information server

CMS CR -2009/114

The Compact Muon Solenoid Experiment

Conference Report

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

18 May 2009

The CMS Online Cluster: IT for a Large Data Acquisition and Control Cluster

Gerry Bauer[7], Barbara Beccati[3], Ulf Behrens[2], Kurt Biery[6], Angela Brett[6], James Branson[5], Eric Cano[3], Harry Cheung[6], Marek Ciganek[3], Sergio Cittolin[3], Jose Antonio Coarasa[5,3], Christian Deldicque[3], Elizabeth Dusinberre[5], Samim Erhan[4], Fabiana Fortes Rodrigues[1], Dominique Gigi[3], Frank Glege[3], Robert Gomez-Reino[3], Johannes Gutleber[3], Derek Hatton[2], Jean-Francois Laurens[3], Constantin Loizides[7], Juan Antonio Lopez Perez[3]fnal, Frans Meijers[3], Emilio Meschi[3], Andreas Meyer[2,3], Remi Mommsen [6], Roland Moser[3]a, Vivian O'Dell[6], Alexander Oh[3]b, Luciano Orsini[3], Vaios Patras[3], Christoph Paus[7], Andrea Petrucci[5], Marco Pieri[5], Attila Racz[3], Hannes Sakulin[3], Matteo Sani[5], Philipp Schieferdecker[3]c, Christoph Schwick[3], Josep Francesc Serrano Margaleff[7], Dennis Shpakov[6], Sean Simon[5], Konstanty Sumorok[7], Marco Zanetti[3]

Abstract

The CMS online cluster consists of more than 2000 computers running about 10000 application instances. These applications implement the control of the experiment, the event building, the high level trigger, the online database and the control of the buffering and transferring of data to the Central Data Recording at CERN. In this paper the IT solutions employed to fulfil the requirements of such a large cluster are revised. Details are given on the chosen network structure, configuration management system, monitoring infrastructure and on the implementation of the high availability for the services and infrastructure.

Presented at *Computing in High Energy and Nuclear Physics (CHEP*

[1] Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, Rio de Janeiro , Brazil

[2] DESY, Hamburg, Germany

[3] CERN, Geneva, Switzerland

[4] University of California, Los Angeles, Los Angeles, California, USA

[5] University of California, San Diego, San Diego, California, USA

[6] FNAL, Chicago, Illinois, USA

[7] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[a] Also at University of Technical University of Vienna, Vienna, Austria

[b] Now at University of Manchester

[c] Now at Universitaet Karlsruhe

Presented at *Computing in High Energy and Nuclear Physics (CHEP'09),21-27 March 2009,Prague,Czeck Republic,15/05/2009*

The CMS Online Cluster: IT for a Large Data Acquisition and Control Cluster

G Bauer[1], B Beccati[5], U Behrens[2], K Biery[3], A Brett[3], J Branson[4], E Cano[5], H Cheung[3], M Ciganek[5], S Cittolin[5], J A Coarasa[4,5], C Deldicque[5], E Dusinberre[4], S Erhan[5,6], F Fortes Rodrigues[7], D Gigi[5], F Glege[5], R Gomez-Reino[5], J Gutleber[5], D Hatton[2], J F Laurens[5], C Loizides[1], J A Lopez Perez[3,5], F Meijers[5], E Meschi[5], A Meyer[2,5], R Mommsen[3], R Moser[5,8], V O'Dell[3], A Oh[5,9], L B Orsini[5], V Patras[5], C Paus[1], A Petrucci[4], M Pieri[4], A Racz[5], H Sakulin[5], M Sani[4], P Schieferdecker[5,10], C Schwick[5], J F Serrano Margaleff[1], D Shpakov[3], S Simon[4], K Sumorok[1] and M Zanetti[5]

[1]MIT, Cambridge, USA; [2]DESY, Hamburg, Germany; [3]FNAL, Chicago, USA; [4]UCSD, San Diego, USA; [5]CERN, Geneva, Switzerland; [6]UCLA, Los Angeles, USA; [7]CEFET/RJ, Brazil; [8]also at Technical University, Vienna, Austria; [9]now at University of Manchester, Manchester, UK; [10]now at Universität Karlsruhe, Karlsruhe, Germany.

E-mail: coarasa@cern.ch

Abstract. The CMS online cluster consists of more than 2000 computers running about 10000 application instances. These applications implement the control of the experiment, the event building, the high level trigger, the online database and the control of the buffering and transferring of data to the Central Data Recording at CERN. In this paper the IT solutions employed to fulfil the requirements of such a large cluster are revised. Details are given on the chosen network structure, configuration management system, monitoring infrastructure and on the implementation of the high availability for the services and infrastructure.

1. Introduction

The Compact Muon Solenoid (CMS) experiment [1] is one of the four experiments at the Large Hadron Collider [2] (LHC) at the *European Organization for Nuclear Research* (CERN). It is a general-purpose experiment covering a wide range of physics at the TeV scale [3].

Its Data Acquisition (DAQ) system [4] reads data from more than 50 million channels and handles unprecedented data volumes: with an event size of approximately 1 Mbyte and a maximum first level trigger rate of up to 100 kHz, the DAQ system has to handle a throughput of up to 100 Gbytes per second.

Data produced by the CMS detector are fed into the central DAQ system via approximately 650 custom links. At the receiving end, data are fed into commercial networking hardware. Subsequent stages of the data flow are implemented using commercial components.

Event selection in CMS is based on a two level trigger system. The first level, implemented in custom hardware, reduces the event rate to a maximum of 100 kHz. The second level, also called high level trigger (HLT), is implemented in software running on a dedicated computer farm receiving events from the event-building network. The selected events are subsequently transferred to the central data recording center at CERN (Tier 0).

In CMS event building (EVB) is done in 2 stages [5]. The first stage is based on a Myrinet [6] switched network whereas the second is based on multigigabit Ethernet networks.

In order to implement the aforementioned functionalities and the control of the experiment a computing cluster with currently 2000 computers and 120 networking devices has been built up. This cluster gives service to a community of approximately 800 users. This paper discusses details of its information technologies (IT) implementation.

In section 2 the requirements and constraints that have conditioned the design of the cluster are presented. In the third section its implementation is discussed.

2. Requirements, constraints and implications for the CMS Online Cluster IT

While the experimental area, the custom electronics and control computers, are located in a cavern (USC) approximately 80 m below the ground level, most of the online cluster and the control room are located in a building in the surface (SCX5).

Given the environment for the CMS Online Cluster, its IT requirements can be classified into two categories: the more general ones, commonly found in many highly available IT sites, and the ones coming from the specific purpose of the cluster, which has been described in the previous section.

The general requirements are:

- The online cluster must be autonomous, i.e. independent of any other networks, including the CERN campus network, so that data taking can continue even without a link to the CERN network. The implication is that all IT services and infrastructure must be local;
- The IT infrastructure and services must allow uninterrupted operation 24 hours a day, 7 days a week, i.e. the IT infrastructure has to have high availability;
- Strict security is directly implied due to the experiment control being hosted in the cluster;
- The IT infrastructure and services have to be continuously monitored and allow remote control, so that the diagnostic and service of malfunction can be quickly carried out;
- IT infrastructures and services must be scalable to accommodate expansions;
- The commissioning phase of the experiment, with evolving applications being deployed in the cluster, has needed and still needs a fast configuration turnaround of the cluster and services.

The more technical requirements coming from the fact that the cluster is used for the CMS Data Acquisition system are:

- The computers connected to the experiment frontend must have enough networking bandwidth to read from the frontend and feed the subsequent event builder stages;
- The data network must allow traffic of data throughput up to 100 Gbytes/s from the frontend to the high level trigger computers;
- The cluster must contain a farm allowing the operation of the high level trigger [7]. The execution time of the high level trigger software for one event has been measured to be 50 ms on average on a 2.5 GHz core. This implies, at 100 kHz level 1 rate, that at least 5000 cores running at 2.5 GHz are needed;
- There must be enough storage to operate the experiment for up to 2 days without network connection to the Central Data Recording Center at CERN (Tier 0), so there must be at least 300 Tbytes of local storage;
- It must be possible to send up to 1 Gbytes/s to the CMS Tier 0.

All these requirements have to be dealt with under certain constraints, the most important being:

- The computers connected to the frontend have to be quickly replaced in case of failure so as not to interrupt the data acquisition. This implies that there is need for spare computers and fast turnaround in the reinstallation and/or reconfiguration.
- The manpower dedicated to network and computing service is 5 FTE. This implies the use of automatic procedures where possible, which in turn enhances reliability;

- The cluster is running all the time during the commissioning phase not only of the CMS experiment but also of the infrastructure of the LHC. This implies that it has to cope with unexpected cooling failures and power cuts, which requires, for example, implementing an automatic shutdown at the computer level triggered by a temperature over the set limit.

3. The CMS Online Cluster implementation

In the following sections, the implementation of the CMS Online Cluster is presented. The technological choices and implemented services chosen to fulfill the requirements listed in the previous section are discussed.

3.1. Overview of the CMS IT Online Cluster resources

The CMS Online cluster is composed of more than 2000 computers, more than 120 networking switches (Myrinet and Gigabit Ethernet) and a network-attached storage (NAS) system. The cluster has been growing to its actual size in stages.

A more detailed breakdown of the cluster components, according to the different functionality of the computers, follows:

- 640 computers equipped with two Myrinet and three independent 1 Gbit Ethernet links for data networking. They run applications for both stages of the CMS event building;
- 720 computers (named BUFUs) with 8 CPU cores equipped with two 1 Gbit Ethernet lines for data networking. They run applications for the second stage of the event building and the high level trigger;
- 16 computers (storage manager) with access to a *Fiber Channel* storage area network of 300 Tbytes and equipped with four 1 Gbit Ethernet lines for data networking and 2 trunked additional 1 Gbit Ethernet lines for networking to the Tier 0. They run applications to locally buffer the event data and handle the transfer to Tier 0;
- More than 350 computers to control and monitor the subdetector electronics. This includes 90 Windows computers;
- 12 computers for the experiment Online Database implemented via an ORACLE RAC [8];
- 15 computers running the CMS run control system called RCMS [9];
- 50 computers as desktop control terminals for the different subdetectors;
- 10 computers hosting the cluster IT services;
- 5 computers to allow the access to the online cluster from the CERN campus network;
- 200 computers for commissioning and testing;
- 120 active spare computers.

3.2. Networking for the CMS Online Cluster

The Ethernet networking structure in the cluster is depicted in figure 1. The CMS Online Cluster makes use of 5 networks in total:

- The CERN campus public network. This network connects the CMS online cluster to the outside world through the CERN routers and the CMS access servers, acting as computer gateways;
- The LHC technical private network. This network connects all the accelerator related computers (only a few computers are connected to this network. It is not depicted in figure 1);
- Two CMS private networks:
  - The CMS service network. All computers in the CMS cluster are connected to this network. The IT general cluster services are provided via this network, i.e: DNS, Kerberos, LDAP, installation, configuration, etc. This network spans the two physical locations of the CMS computing infrastructure: the underground area and the surface building. The redundant interconnection of the two sites is implemented with 4 routers as shown in figure 2. Four 10 Gbit lines are used to interconnect the routers. The router pairs in each location are configured in a failover mode. The interconnectivity

within each location is also redundant: in each rack, a switch connects to all computers in the rack and to both routers. This cabling tolerates single failures up to the rack switch.

  - The CMS data networks. These are eight networks used for the event building, connectivity to the high level trigger filter farm and to the storage manager. To implement these high bandwidth networks, eight *Force 10* [10] 1 Gbit Ethernet switches with 6 high density line cards each are used. One of these networks connects 80 readout units, 90 BUFUs and all 16 storage manager computers through all of their data network interfaces [4], totalling 3576 ports for the eight switches. In order to make use of the aggregate bandwidth of all interfaces in each computer the data networks are divided in to Virtual LANs (VLANs) and the computers are configured to use source routing;

- The central data recording private network. This network allows transferring data from the storage manager computers to the Tier 0 at CERN for later analysis and archiving. It is serviced by two 10 Gbit lines in redundant failover configuration.

In building up these networks more than one hundred gigabit Ethernet switches are used.

An additional Myrinet network connects the computers involved in the first stage event building through twelve Myrinet switches. More details about the custom protocols used there can be found in [4].

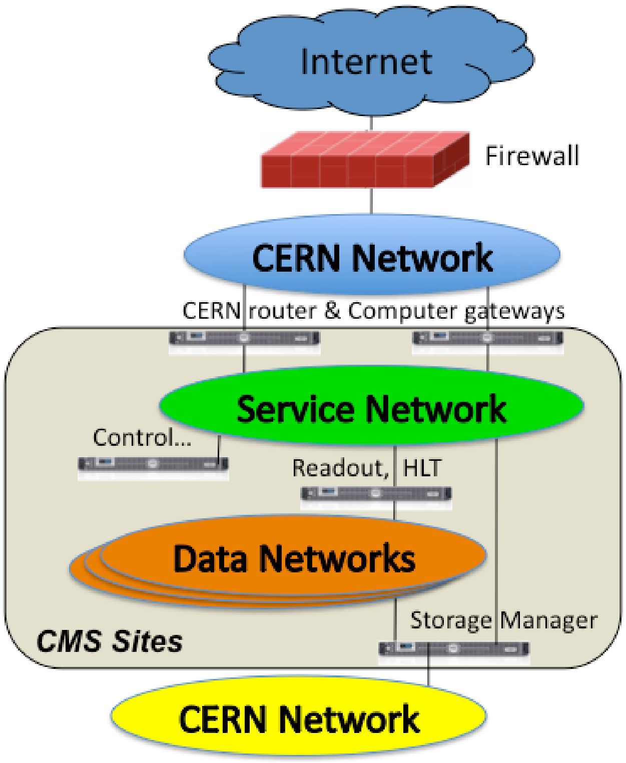

Figure 1. Ethernet networking structure of the CMS online cluster. The LHC Technical network is not depicted here.

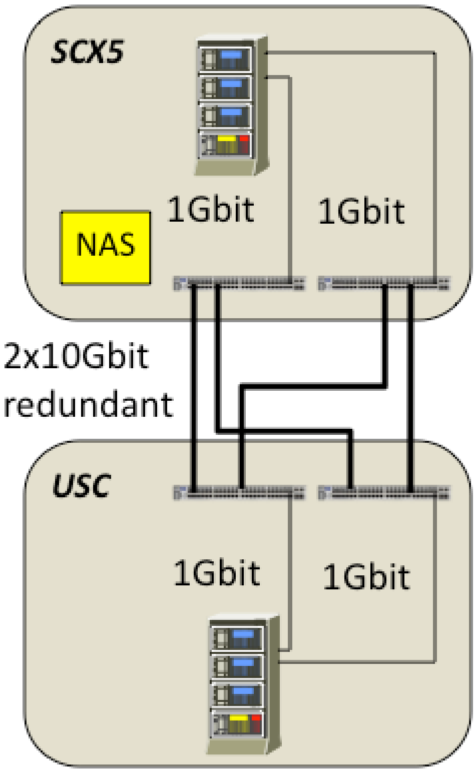

Figure 2. Redundancy of CMS private service network connecting the two locations of the CMS experiment.

3.3. The Network-attached storage

Important data that do not need to be read by many computers at the same time and need to be kept safe are stored in a *NetApp* network-attached storage filer (NAS) [11]. These data account at the present for the user home directories, CMS analysis data such as calibration and data quality monitoring, system administration data, software repositories and configuration management data.

The actual configuration of the *NetApp* filer (figure 3) is consisting of two sets of one head (model FAS3040-R5) and 3 storage shelves (model DS14MkII). Each set is located in a different rack and powered with a UPS and non-backed up power for redundancy. The two heads are in failover configuration. The 3 shelves in one rack are mirrored to the ones in the other rack, and each shelf is configured with internal Dual Parity RAID 6. The internal snapshot feature is active so that the necessity to restore a backup is minimized.

The *NetApp* filer heads are connected to the CMS network routers with redundant 10 Gbit links. Before commissioning the 10 Gbit redundant links, the system was operated with four 1 Gbit trunked lines connecting one of the heads to one of the routers. Measurements carried out with this configuration were limited by the bandwidth of the trunked link. The bandwidth was measured to be 380 Mbytes/s.

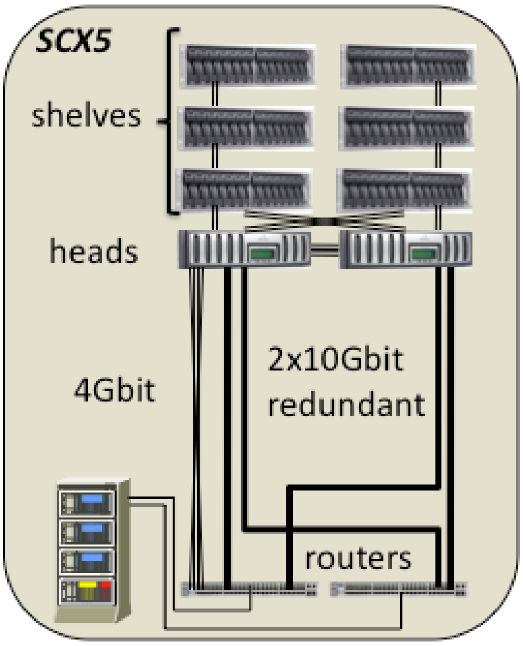

Figure 3. Redundancy in the CMS *NetApp* network-attached storage configuration and its networking.

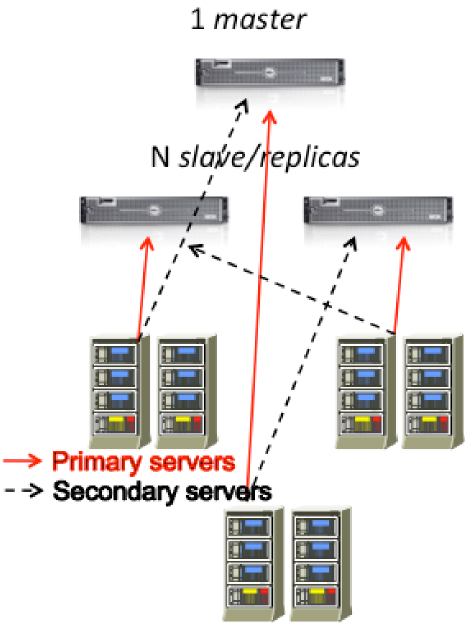

Figure 4. Model for the redundancy and shared load of some of the CMS structural services.

3.4. The CMS IT structural services

The scale of the cluster and the required high availability call for redundancy in the IT structural services. The cluster services must be scalable to accommodate future expansion.

For the CMS IT structural services (see table 1) a model with one master server and N (where N is now 2) replica servers has been chosen. The administration is done in the master server. It is the single point of administration for the service. This requires that all services' data are replicated to the replica servers. Some of the services allow for standard replication (i.e. slurpd for LDAP or the master/slave mechanism for the DNS data) but some other services do not provide a mechanism to do it (i.e. the Kerberos or the DHCP services). For such services, in-house developments were created to replicate the data from the master servers to the replica servers. For more details refer to table 1.

The services are provided, where possible, by a DNS aliased name. In case of need (administration purposes, recovering from a crashed computer, etc.) the DNS alias can be changed to point to a different computer minimizing the interruption of the service.

Following standard practices in IT, these services are hosted in computers with mirrored system disks and powered with a UPS and non-backed up power. For extended redundancy, the different computers providing the same service are located in different racks, so that if the network or the power fails in one rack the service is still provided.

Having several computers serving the same services allows balancing the load among them. So far, two "*load sharing*" techniques are used:

- DNS Round Robin;
- Segregating the computers of the cluster in groups that point to different primary servers for a particular service, as depicted in figure 4. This allows to have a load shared, controlled and reproducible environment while still keeping redundancy for the services. In case of failure of the primary server, in this schema, only the affected group of computers will have to default to the secondary server.

In table 1 a summary of the services, the replication method used for the configuration data, and the load balancing technique used, can be found.

Table 1. The CMS IT structural, installation and configuration services.

Service	**Replication through**	**Load Balancing through**
DNS	named	Explicit segregation
DHCP	in-house scripts	No. First who answers
Kerberos	in-house scripts	Explicit segregation
LDAP	slurpd	Explicit segregation
NTP	-	Explicit segregation
syslog	-	None. Single server, only used for stress test purposes
Nagios monitoring	-	Explicit segregation
PXE boot/TFTP	No replication needed. NFS mount from the NAS.	Not yet
Kickstart Installation server		DNS Round Robin
yum repository Installation server		DNS Round Robin
Quattor repository Installation server		DNS Round Robin
Quattor cdb configuration server		Not yet

3.5. The CMS installation, management and configuration infrastructure

Management of big clusters requires remote control and a configuration management system on top of the infrastructure services described previously. Their implementation is detailed bellow.

The remote control to the CMS cluster is provided using tools for the *Intelligent Platform Management Interface* (IPMI) [12]. The IPMI allows interacting with the remote computer through a dedicated network interface to perform operations like: boot up, power off, access the console, etc. In addition it allows reading environmental conditions of the computer like: ambient temperature, rotation speed of fans, etc.

As a configuration management system, Quattor (QUattor is an Administration ToolkiT for Optimizing Resources) [13] was chosen. All linux computers, including the infrastructure servers, or the Quattor servers themselves, are configured using Quattor modules and packages (RPMs) distributed with Quattor. As an example, the BIOS parameters, the IPMI interface parameters, and all the networking parameters (routing tables and IP addresses for the data networks) are configured in the computers at boot time through shell scripts packaged in RPMs. CMS online software is also packaged in RPMs and distributed through Quattor.

The installation and configuration management infrastructure is based on: PXE [14], TFTP, the anaconda Redhat kickstart [15] and Quattor. All these services follow the same model as the structural services which have been described above (table 1). The services are DNS aliased and they are load shared among 3 servers using DNS Round Robin. In this case, the configuration data do not need to be replicated because they are kept on the NAS, where all of the servers have access. The master server, defined via a DNS alias, is given write access to the configuration data.

During the installation phase the computers read their kickstart file through TFTP, start the anaconda Redhat installer, install the Quattor client packages through yum in the postinstall section of the anaconda installer, and apply the configuration information found in the Quattor description templates and reboot.

During the installation or reconfiguration phase of a large number of computers the network traffic to the NAS is limited, due to the fact that the necessary packages are read only once by the repository servers and kept in cache memory while distributed to the clients. This way the quality of the NAS service is guaranteed.

The CMS production Quattor implementation is based on Quattor 1.3: cdb 2.0.4, swrep 2.1.38, PANC 6.0.8. A hierarchy of configuration templates and a set of policies have been developed for the CMS cluster along with tools that ease the administration of the configuration of the computers. In particular, they ease the upgrade of the online software, make the procedure less error prone and accessible to users without knowledge or privileges to the Quattor system. These tools include:

- A tool lists the groups of computers with the same versions of specific online software. These lists allow automating other administration tasks, such as running a specific command only on a specific group of computers;
- A dropbox for RPMs allows the users to deploy an RPM in a directory to upload it to the Quattor repository for later distribution;
- A template updater can automatically update the templates of the computers;

The CMS Quattor implementation currently manages more than 2000 Scientific Linux CERN 4 computers of more than 50 types. A compilation of all templates takes approximately 3 minutes in the 8 core 2.66 GHz computer hosting the active *master* Quattor server. Fully reinstalling one computer (with full formatting of the disk) takes between 4 and 9 minutes depending on how much software is loaded. The existing infrastructure allows reinstalling 1000 computers from scratch in less than an hour and fifteen minutes. It is presently limited by the network traffic to the repository servers.

3.6. CMS Monitoring infrastructure

Monitoring is the key element in being able to diagnose and solve a problem. Especially in the case of having so many instances of services, correlating the different failures to the real cause can take quite some time if not automated.

The CMS monitoring infrastructure is based on Nagios [16]. The CMS monitoring Nagios infrastructure is still being developed. Apart from including the standard Nagios tests (ping, ssh accessibility, disk usage...) specific modules have been developed to check:

- If the Myrinet links are functional;
- If a configuration and installation finished properly in the computers;
- If the IPMIs are reachable;
- If the storage manager components are operational.

The development of more modules still continues to improve the cluster diagnosis.

The CMS Nagios monitoring infrastructure performs presently about 20000 service checks. The current effort aims at improving the performance of the service by distributing it on multiple hosts. The goal is to reduce the time necessary for a complete system check to less than one minute (currently 15 minutes are necessary) and to make it more scalable such that future expansions (more services or more computers) can be accommodated.

3.7. CMS IT Security

The control of the CMS experiment is hosted in networked computers. This demands high levels of IT security.

On the other hand, the development of applications controlling hardware is more difficult when security has to be observed. At the same time, applying the security process (follow the security news, apply the patches and so forth) to a large cluster of 2000 computers demands for man power and can be quite disturbing for the CMS data taking.

These are the reasons why the CMS internal networks are private (figure 1). The CMS computers hosting the control and DAQ system are not only behind the CERN firewall, that already protects against access to all non published services, but also inside a private network not even reachable from the CERN network unless the specific service is proxyed.

The need to be able to control parts of the experiment from the outside is fulfilled by gateway computers that allow the users to login into the CMS network from the CERN cluster. These computers are subject to tighter security following the CERN security policies [17].

In special cases (online monitoring, data quality monitoring, etc.), internal selected http traffic is proxyed through a reverse proxy, only accessible to the CERN network and subject to tighter security measures too (authentication is always required).

4. Summary

The CMS Online Cluster has been working successfully since 2007 during the commissioning phases and cosmic runs of the CMS detector.

It consists presently of more than 2000 computers that run applications implementing the control of the experiment, the event building, the high level trigger, the online database and the control of the buffering and transferring of data to Tier 0 at CERN.

The design of the cluster and networking allows the CMS DAQ system to handle unprecedented data volumes up to 100 Gbytes/s.

The chosen IT infrastructure and services provide high availability and allow autonomous and independent running of the cluster.

The design allows for easy scalability to accommodate further expansions.

Acknowledgements

This work was supported in part by DOE and NSF (USA) and the Marie Curie Program.

References

[1] The CMS Collaboration (Adolphi R et al.) “The CMS Experiment at CERN LHC”, *JINST* **3** S08004 361, 2008.
The CMS Collaboration, “The Compact Muon Solenoid Technical Proposal”, CERN/LHCC 94-38, 1994.

[2] The LHC Study Group, “The Large Hadron Collider Conceptual Design Report”, CERN/AC 95-05, 1995.

[3] The CMS Collaboration, “The CMS Physics Technical Design Report, Volume 1”, CERN/LHCC 2006-001, 2006. CMS TDR 8.1.
The CMS Collaboration, “The CMS Physics Technical Design Report, Volume 2”, CERN/LHCC 2006-021, 2006. CMS TDR 8.2.
The CMS Collaboration, “High Density QCD with Heavy Ions; Physics Technical Design Report, Addedum 1”, CERN/LHCC 2007-009, 2007. CMS TDR 8.2-Add1

[4] The CMS Collaboration, “Trigger and Data Acquisition Project: Level-1 Trigger,”, CERN/LHCC 2000-38, Vol. I, 2000.
The CMS Collaboration, “The Trigger and Data Aquisition project”, CERN/LHCC 2002-26, Vol II, 2002.

[5] Mommsen R K et al. “The CMS event builder and storage system”, 2009 Prepared for

International Conference on Computing in High Energy and Nuclear Physics (CHEP 2009), Prague, Czech Republic, 21-27 Mar 2009.

[6] http://www.myricom.com/myrinet/overview/

[7] The CMS Collaboration, "The CMS High Level Trigger", CERN/LHCC 2007-021, LHCC-G-134 29, June 2007.

G. Bauer et al. "CMS DAQ event builder based on Gigabit Ethernet", IEEE Trans.Nucl.Sci.**55**:198-202,2008.

[8] http://www.oracle.com/technology/products/database/clustering/index.html

[9] Bauer G et al. 2008, "The run control and monitoring system of the CMS experiment", *J. Phys. Conf. Ser.* (CHEP 07, Victoria, BC, Canada, 2-7 Sep 2007) 119 022010

[10] http://www.force10networks.com/

[11] http://www.netapp.com/

[12] http://www.intel.com/design/servers/ipmi/index.htm

[13] Quattor: Tools and Techniques for the Configuration, Installation and Management of Large-Scale Grid Computing Fabrics. *Journal of Grid Computing* Vol. 2/4. Springer Verlag, December 2004.

http://quattor.org/

[14] "The Preboot Execution Environment specification v2.1" published by Intel & Systemsoft. http://download.intel.com/design/archives/wfm/downloads/pxespec.pdf

[15] http://www.redhat.com/docs/manuals/linux/RHL-7.3-Manual/custom-guide/ch-kickstart2.html

[16] http://www.nagios.org/

[17] http://security.web.cern.ch/security/