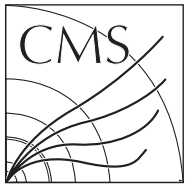
Available on CMS information server

CMS CR -2012/138



The Compact Muon Solenoid Experiment

Conference Report

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

12 June 2012 (v2, 14 June 2012)

Operational experience with the CMS Data Acquisition System

G Bauer[6)], U Behrens[1)], M Bowen[2)], J Branson[4)], S Bukowiec[2)], S Cittolin[4)], J A Coarasa[2)], C Deldicque[2)], M Dobson[2)], A Dupont[2)], S Erhan[3)], A Flossdorf[1)], D Gigi[2)], F Glege[2)], R Gomez-Reino[2)], C Hartl[2)], J Hegeman[2,a)], A Holzner[4)], Y L Hwong[2)], L Masetti[2)], F Meijers[2)], E Meschi[2)], R K Mommsen[5)], V O'Dell[5)], L Orsini[2)], C Paus[6)], A Petrucci[2)], M Pieri[4)], G Polese[2)], A Racz[2)], O Raginel[6)], H Sakulin[2)], M Sani[4)], C Schwick[2)], D Shpakov[5)], M Simon [2)], A Spataru[2)], K Sumorok[6)]

Abstract

The data-acquisition (DAQ) system of the CMS experiment at the LHC performs the read-out and assembly of events accepted by the first level hardware trigger. Assembled events are made available to the high-level trigger (HLT), which selects interesting events for offline storage and analysis. The system is designed to handle a maximum input rate of 100 kHz and an aggregated throughput of 100 GB/s originating from approximately 500 sources and 10^8 electronic channels. An overview of the architecture and design of the hardware and software of the DAQ system is given. We report on the performance and operational experience of the DAQ and its Run Control System in the first two years of collider runs of the LHC, both in proton-proton and Pb-Pb collisions. We present an analysis of the current performance, its limitations, and the most common failure modes and discuss the ongoing evolution of the HLT capability needed to match the luminosity ramp-up of the LHC.

Presented at *CHEP 2012: International Conference on Computing in High Energy and Nuclear Physics*

[1)] DESY, Hamburg, Germany

[2)] CERN, Geneva, Switzerland

[3)] University of California, Los Angeles, Los Angeles, California, USA

[4)] University of California, San Diego, San Diego, California, USA

[5)] FNAL, Chicago, Illinois, USA

[6)] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[a)] Now at Princeton University

Operational experience with the CMS Data Acquisition System

G Bauer[6], U Behrens[1], M Bowen[2], J Branson[4], S Bukowiec[2], S Cittolin[3], J A Coarasa[2], C Deldicque[2], M Dobson[2], A Dupont[2], S Erhan[3], A Flossdorf[1], D Gigi[2], F Glege[2], R Gomez-Reino[2], C Hartl[3], J Hegeman[2,7], A Holzner[4], Y L Hwong[2], L Masetti[2], F Meijers[2], E Meschi[2], R K Mommsen[5], V O'Dell[5], L Orsini[2], C Paus[6], A Petrucci[2], M Pieri[4], G Polese[2], A Racz[2], O Raginel[6], H Sakulin[2], M Sani[4], C Schwick[2], D Shpakov[5], S Simon[2], A Spataru[2], K Sumorok[6]

[1] DESY, Hamburg, Germany;
[2] CERN, Geneva, Switzerland;
[3] University of California, Los Angeles, Los Angeles, California, USA;
[4] University of California, San Diego, San Diego, California, USA;
[5] FNAL, Chicago, Illinois, USA;
[6] Massachusetts Institute of Technology, Cambridge, Massachusetts, USA;

Hannes.Sakulin@cern.ch

Abstract. The data-acquisition (DAQ) system of the CMS experiment at the LHC performs the read-out and assembly of events accepted by the first level hardware trigger. Assembled events are made available to the high-level trigger (HLT), which selects interesting events for offline storage and analysis. The system is designed to handle a maximum input rate of 100 kHz and an aggregated throughput of 100 GB/s originating from approximately 500 sources and 10^8 electronic channels. An overview of the architecture and design of the hardware and software of the DAQ system is given. We report on the performance and operational experience of the DAQ and its Run Control System in the first two years of collider runs of the LHC, both in proton-proton and Pb-Pb collisions. We present an analysis of the current performance, its limitations, and the most common failure modes and discuss the ongoing evolution of the HLT capability needed to match the luminosity ramp-up of the LHC.

1. Introduction

The Compact Muon Solenoid (CMS) [1] experiment at the Large Hadron Collider (LHC) is one of two large general-purpose detectors aimed at studying a broad range of physics at the TeV scale. The detector comprises about 55 million readout channels and is designed to study both proton-proton and heavy ion collisions produced at the LHC. Online event-selection is performed using only two trigger levels: a hardware-based first-level trigger that accepts up to 100 kHz of events and a software-based high-level trigger (HLT) that analyzes events using the full offline reconstruction software running on a farm of computers. While this concept has the advantage that the high-level trigger can work on full events, it poses a challenge for the experiment's Data Acquisition (DAQ) system [2]: it has to read out and assemble events at the level-1 trigger rate of 100 kHz, transporting around 100 GB/s of data.

[7] Now at Princeton University

CMS has successfully been recording proton-proton collisions at a center-of-mass energy of 7 TeV during 2010 and 2011, and at 8 TeV since the start of 2012. Over this time the instantaneous luminosity delivered by the LHC has been increasing by several orders of magnitude and is currently at around 60% of the LHC design luminosity of $10^{34}/(cm^2s)$. CMS also successfully recorded Pb-Pb collisions during one month in 2010 and one month in 2011. In the present paper we discuss the performance of the DAQ system for proton and for ion physics and we analyze whether the increased pile-up of events due to LHC operation at 50 ns bunch spacing rather than the foreseen 25 ns can be handled. We report on the extensions of the high-level trigger farm needed to match the luminosity ramp-up of the LHC and discuss the operational efficiency of the central DAQ system and of the overall CMS detector.

2. The CMS central DAQ system

2.1. DAQ hardware

The CMS central DAQ system is designed to read out event fragments of an average size of up to 2 kB from around 700 detector Front-Ends Drivers (FEDs) (figure 1) at the level-1 trigger rate of 100 kHz. In a first step, data are transferred from the sub-detector specific FEDs to common Frontend Readout Link (FRL) modules via the SLINK-64 [3] copper links. For FEDs with smaller fragment size, the FRL reads out two FEDs and merges the fragments in order to balance fragment sizes. Events are then built in two stages. In the first stage, a Myrinet [4] optical network is employed to build super-fragments from the output of 8 FRLs. FRLs send the data through Myrinet NICs housed on the FRL module using a custom protocol implemented on the RISC processor of the Myrinet NIC. Super-fragments belonging to one event are built in a set of Readout Unit PCs. Full events are then built on a set of Builder/Filter Unit PCs by a Readout Builder, which consists of the aforementioned PCs, an Ethernet switch with ~500 1 Gb/s ports and an Event Manager PC controlling the data flow. The CMS DAQ currently comprises eight Readout Builders, also called slices, each processing around $1/8^{th}$ of events working at an event-building rate of 12.5 kHz. Fragments are assigned to one of the 8 readout builders based on a look-up-table, which may be adjusted in order to accommodate Readout Builders with different performance. The Builder/Filter Unit PCs also run the high-level trigger software. Events accepted by the high-level trigger are transferred to Storage Manger PCs (two per slice), which store the data to a local Storage Area Network of a total capacity of 300 TB, from where they are transferred to the Tier-0 computing center at the main CERN site. Typically ~500 Hz of events are recorded in in the main data stream for proton physics, corresponding to a total throughput to disk of several hundred MB/s (data are compressed in the high-level-trigger). For special runs and for heavy-ion physics, recording rates of up to 2.8 GB/s can be achieved. The Storage Managers also serve events to consumers such as the Data Quality Monitoring (DQM) service. As discussed in section 6, the Readout Builders have been extended with additional Builder/Filter Unit PCs in order to increase HLT processing power.

2.2. DAQ software

Two main software frameworks have been developed for the CMS DAQ system. The Run Control System [5] provides the control hierarchy for the experiment. Based on Java and Web applications it allows for the definition of so-called Function Managers that control a part of the system and are in turn controlled by Function Mangers at a higher level, or by an operator through a web interface. For global operation, the DAQ operator controls the experiment through the top-level Function Manager. This Function Manager is at the top of a hierarchy of around 60 Function Managers that in turn control around 20000 worker nodes. It defines a global state machine for the experiment but also allows for individual control of the sub-systems. The top-level Function Manager is aware of the state of the LHC and of the Detector Control System and includes many built-in cross-checks to guide the operator.

The worker nodes are implemented using the C++ based XDAQ [6] framework. This framework provides libraries for hardware access and for data transport using various standard protocols. It provides a configuration mechanism through XML [7], communication through SOAP [8] and access to the applications via a web interface. XDAQ further provides services to collect monitoring information and error reports from all applications and to make the monitoring information available to a number of clients. The monitoring clients present the monitoring information either as high-level graphical summary or in detailed views that allow the operators to understand the performance of the system and to diagnose problems. A further monitoring client, the DAQ Doctor, constantly analyzes the monitoring data, performs checks and gives advice to the DAQ operator. It case of abnormal situations it alerts the shift crew via an audio alert system.

The event processor applications running on the Builder/Filter unit PCs include the full CMSSW [9] offline event reconstruction framework. One event processor is started per core or hyper-thread on the machine. In order to decrease the memory footprint, a prototype event processor is created and configured which then forks the other event processors as child processes so that they share large parts of the memory through the copy-on-write mechanism.

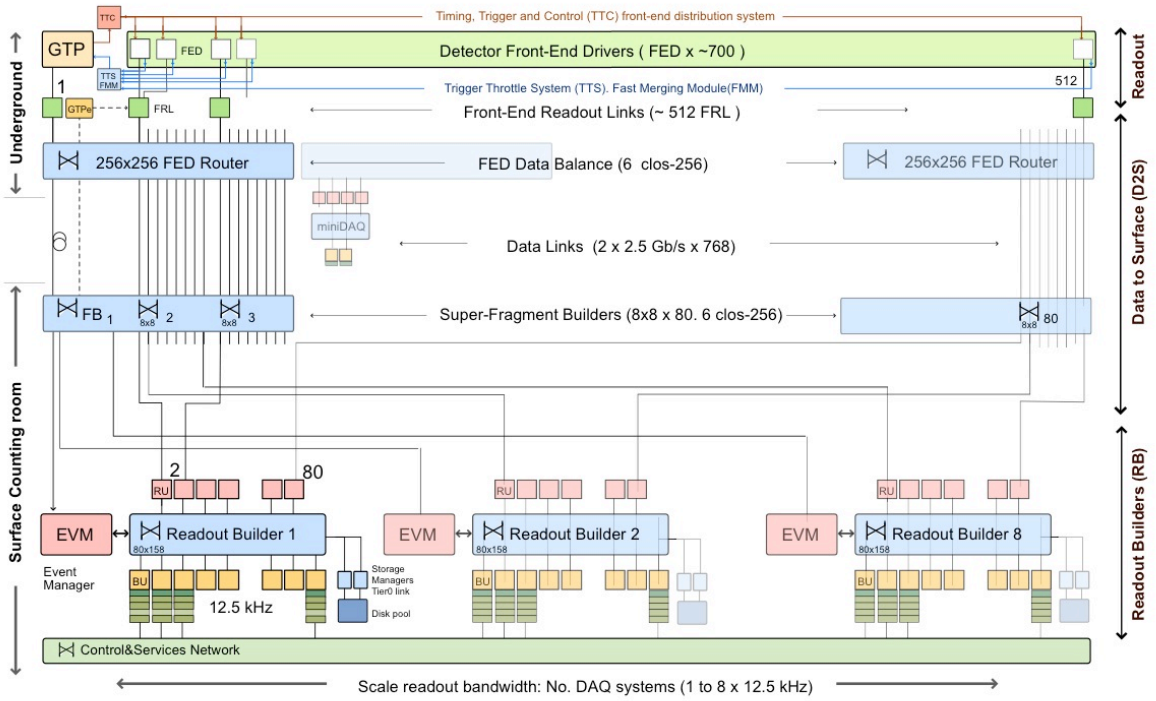


Figure 1. The CMS central DAQ system. Events are built in two stages. Super-fragments are built in the Readout Unit (RU) PCs, full events are built in the Builder/Filter Unit (BU) PCs. The BU PCs also run the high-level trigger software.

3. Operating conditions in 2011/12

Peak instantaneous luminosity provided by the LHC has been increasing throughout 2011 and 2012, and is expected to reach $7 \times 10^{33}/(\text{cm}^2\text{s})$ later in 2012. At the start of 2012 it was decided to continue to operate the LHC with a bunch spacing of 50 ns during 2012 rather than the foreseen 25 ns. Under these conditions the number of interactions piled up in a bunch crossing will reach 35, almost twice

the pile-up CMS was designed for [2]. In section 4 we analyze how this increased pile-up affects the data volume to be handled by the DAQ system. The increased pile-up conditions also affect the processing time needed for several algorithms in the high level trigger, and are dealt with in section 6.

4. Event size and throughput

The CMS event size as a function of the number of primary vertices has been studied during a special fill with high pile-up in 2011 (fill 2252) and is shown in figure 2. At the a pile of 35, which results in around 25 reconstructed primary vertices, the event size will still be well below the CMS design event size of 1 MB. So globally the required CMS DAQ throughput will stay well below DAQ design specifications. Bandwidth limitations may however be reached for individual Frontend-Readout Links.

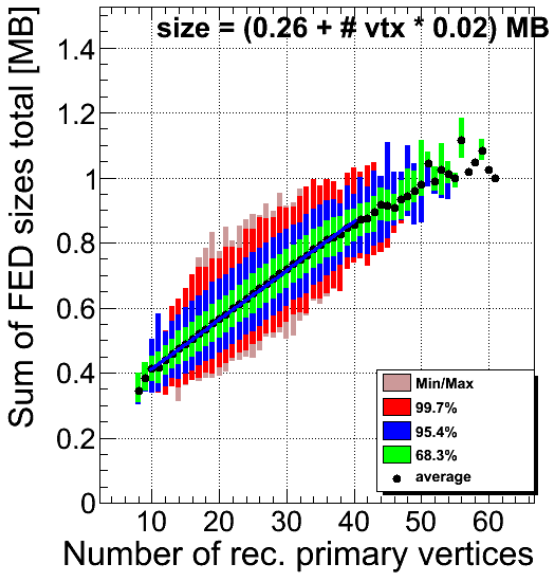


Figure 2. CMS total event size as a function of the number of reconstructed primary vertices in the event. Obtained from the analysis of events in a special LHC fill high intensity bunches (fill 2252).

At the nominal fragment size of 2 kB and a trigger rate of 100 kHz the throughput per input is 200 MB/s. There is plenty of headroom in the SLINK-64 and over the Myrinet fibres, which can handle 400 MB/s and 500 MB/s, respectively. The cross-bars in the Myrinet switches building the super-fragments however show limited throughput due to head-of-line blocking unless traffic shaping is employed. We studied the DAQ system performance using the Frontend-Readout links as data sources producing fragments of 2 kB averages size with a log-normal distribution of 2 kB standard deviation. Figure 3 shows the throughput per node in an 8x8 Super-Fragment Builder. The figure also shows the corresponding trigger rate. At a rate to 100 kHz, fragments of 2.5 kB can be handled resulting in a throughput of around 250 MB/s. The actual limit may be different as the throughput strongly depends on the fragment size distribution.

At a pile-up of 35, some Front-end Drivers in the Pixel sub-system are expected to send data at up to 280 MB/s, as measured during the special high pile-up fill in 2011. Depending on the fragment size distribution, the corresponding super-fragment builders may limit the DAQ throughput thus causing back-pressure and dead-time. Only 32 FEDs were found to reach a problematic throughput. In order to avoid any possible limitation during 2012 data taking, we decided to split the concerned super-fragment builders into smaller ones with only 5 or 6 inputs. To compensate for the increased use of DAQ resources, super-fragment builders of smaller sub-systems with only few inputs were combined. It should be noted that no out-of-time pile-up (pile-up due to events in neighboring bunch crossings) was generated in the special high pile-up fill, since only some of the LHC bunches were filled and these were spaced far apart. In some sub-systems that read out multiple bunch crossings, out-of-time pile-up is expected to cause an additional throughput increase.

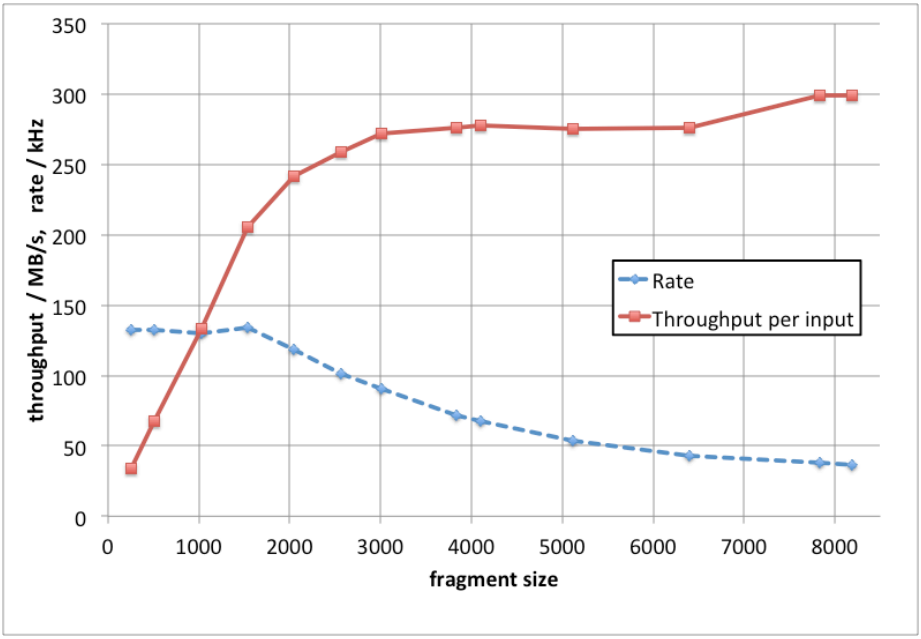


Figure 3. Throughput per input and data acquisition rate as a function of average fragment size for a DAQ system configuration with 8x8 super-fragment builders and 8 readout builders. Fragments are generated with the FRL emulator mode according to a log-normal distribution with standard deviation equal to the average fragment size.

5. Heavy ion operation

During LHC operation with heavy ions, readout requirements of CMS are different with respect to proton operation. Due to the much higher track multiplicity, hardware zero suppression is not used in the Si-strip tracker increasing the per-FED data size to 50 kB. After merging of 2 FEDs in the FRLs, fragments are of 100 kB size. The over-all event size is 20 MB. The required data acquisition rate on the other hand is only of the order of a few kHz since the interaction rate is much smaller. Measurements with generated fragments (events with log-normal event size distribution with a standard deviation equal to the average size) showed that by optimizing parameters of the DAQ system, a throughput of 350 MB/s per FRL can be reached for fragments of 100 kB average size. This corresponds to a data-acquisition rate of 3.5 kHz and to a throughput 1.7 times higher than the design throughput for proton physics. In the heavy-ion runs in 2011, a peak acquisition rate of 2.7 kHz was attained at the start of a fill.

6. Extension of the high-level trigger farm

The CMS event builder and high-level trigger farm are built using standard commercial PCs and networking equipment and are therefore easily extendable. In order to benefit maximally from the advances in processor technology, it is a CMS strategy to install the necessary computing power as late as possible. With increasing instantaneous luminosity, the high-level trigger needs to be more selective and algorithms require more processing power. Moreover, certain algorithms such as tracking are affected by the higher track multiplicity because of pile-up. The HLT farm has been extended twice so far, in May 2011 and in May 2012. Table 1 shows the parameters and counts of the deployed machines. The performance of the machines has been studied in a test setup, using playback of a set of recorded events that pass the Level-1 trigger. Results were obtained with a recent version of the CMSSW simulation software and with an HLT menu designed for a luminosity of 5×10^{33} /cm^{2}s.

The number of event processor applications run on the machines was varied between one and the number of cores (or the number of hyper-threads in case of the newer generations of machines) as shown in figure 4. For the HLT applications, hyper-threading gains around 30% of performance on the machines based on Westmere and Sandy Bridge processors. The over-all performance gain per motherboard with respect to the original 8-core Harpertown machines is 2.8 for the 12-core Westmere machines and 3.9 for the 16-core Sandy Bridge machines. Performance gains observed during 2011 on machines integrated into the Readout Builders / HLT farm were slightly lower. Based on more conservative performance gains, the available CPU budget per event (measured on one core of the original 8-core Harpertown machines) has been increased from originally around 50 ms to around 100 ms in May 2011 and to around 150 ms in May 2012.

Table 1. Parameters and counts of the machines in the high-level trigger farm

	Original farm (2009)	Extension 1(2011)	Extension 2 (2012)
CPU	Intel Xeon E5430 Harpertown	Intel Xeon X5650 Westmere	Intel Xeon E5-2670 Sandy Bridge
# cores per motherboard	2 x 4-core	2 x 6-core	2 x 8-core
Hyper-threading	No	Yes	Yes
Clock speed	2.66 GHz	2.66 GHz	2.6 GHz
RAM	16 MB	24 MB	32 MB
# motherboards	720	288	256
#cores (cumulative)	5.6k	9.1k	13.2k

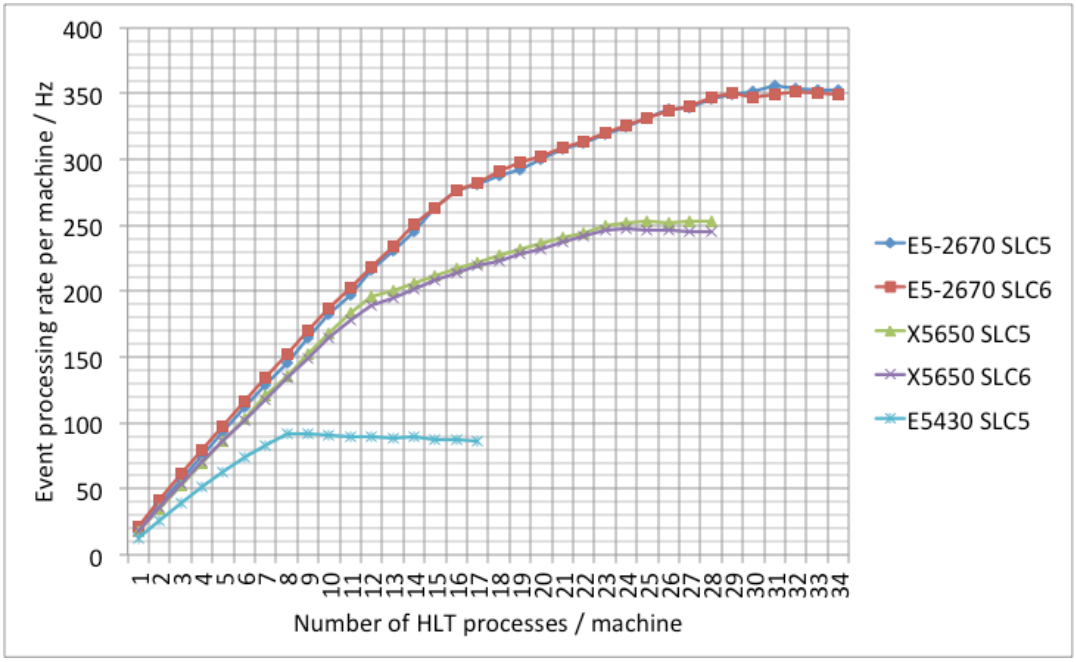


Figure 4. Event processing rate per machine as a function of the number of High-Level Trigger processes running on the machine for the three types of machines employed in the CMS high-level trigger farm. Measurements were done in a "playback" setup in which events are read from files. An HLT menu optimized for $L=5x10^{33}/(cm^2s)$ was employed. For the newer types of machines a performance gain of approximately 30% due to hyper-threading can be observed. Performance is similar under Scientific Linux CERN 5 (SLC5) and Scientific Linux CERN 6 (SLC 6).

7. Central DAQ operational efficiency

During 2011 proton-proton fills, the central DAQ system was available during 99.7 % of stable beam periods. The system was down during less than 4 hours. Around 3 hours of down-time were caused by software problems. These are defects that either surfaced due to changed operating conditions or that had been newly introduced. In general these defects were fixed as soon as they had been identified. Around 1 hour of down-time can be attributed to hardware failures which occurred in various parts of the system: a link failure in a Myrinet switch, a linecard failure in a Gigabit Ethernet switch and a failure of a control network switch and PC failures. In total, 203 PC failures were observed in 2011, mostly due to a problem with disk controllers, but most of them did not cause any down time due to the resilience features built into the CMS DAQ. Failures of PCs or applications handling the event building and storage are tolerated – only bandwidth is degraded in this case. Crashed event filter applications are restarted automatically. Crashes of PCs or applications controlling custom hardware are tolerated – in this case the run continues with reduced monitoring capabilities. Thanks to the DAQ slice concept, any problem (PCs, network, etc.) located in a single slice may be worked around by masking the corresponding slice. Data taking can then resume with $7/8^{th}$ of the performance. This action can be performed by the shift crew and just takes around 3 minutes.

In order to start a new DAQ session with all slices in the presence of failed PCs or network connections, a new DAQ configuration needs to be computed. For this purpose the CMS DAQ Configurator application has been developed. Based on a high-level description of the system and on Configuration Templates it computes the detailed configuration of the O(10000) applications in the DAQ system and their network connectivity. Initially several steps were needed to create a new configuration, a procedure that would take an on-call expert at least 10 minutes to complete. During 2010 the system was automated and a blacklist database and editor were integrated so that an expert could create a new configuration in about 2 minutes. Recently the system was integrated with the DAQ Doctor expert system, which now automatically creates a new configuration when it detects a crashed PC, taking about 40 seconds to do so.

8. CMS operational efficiency

In 2011 CMS recorded 91.2 % of delivered integrated luminosity in proton physics and 94.4 % of delivered integrated luminosity in heavy ion physics. Luminosity was lost due to down times (6 % during proton physics) and dead times (3 % during proton physics), the latter due to trigger rules, FEDs requiring to be resynchronized etc. Dead times were caused by infrastructure and miscellaneous problems (31%), by the trigger system (6%), by central DAQ (6 % - discussed above) and by sub-detector DAQ (57 %).

With increased instantaneous luminosity towards the end of 2011, sub-detectors observed an increased frequency of problems due to single-event upsets in the on-detector electronics. In many cases the recovery from these problems required the run to be stopped, the sub-detector to be reconfigured and a new run to be started. For 2012, a new recovery procedure under the control of the top-level run control has been designed which does not need the run to be stopped. Sub-detectors suffering from a single-event upset or other soft error notify the top-level run control node which invokes a recovery transition which also gives other sub-detectors a chance to reset/recover hardware in the background.

In general, the time spent on recovering from problems depends on the times needed by the various sub-systems to perform the state machine transitions needed to start a run. A lot of effort has gone into optimizing these. Currently the start of a data taking session, from a state where none of the applications are running, takes about 3 minutes; a stop and start of a run takes about 1 minute 15 seconds.

9. Summary

We have reported on various aspects of the operation of the CMS DAQ system in 2011 and on preparations for 2012 operation. We have studied whether the central DAQ system will be able to cope with the increased pile-up due to LHC operation with a bunch spacing of 50 ns and found that the system can be adapted to handle the increased throughput. We have reported on the operation of the DAQ system for heavy-ion physics in November/December 2011 where some of the inputs reached 50 times the nominal size but acquisition rates were smaller. By optimizing the DAQ system parameters, a throughput 1.7 times higher than the design throughput for proton physics could be reached. Moreover we have reported on the high-level trigger farm, which is designed to be extendable in order provide more CPU power as the LHC ramps up its luminosity. The farm has been extended in May 2011 and in May 2012, taking the total number of cores to 13000. The available CPU budget per event has been tripled by the two extensions.

The CMS central DAQ availability in 2011 proton operation was 99.7 % during stable-beam periods of the LHC thanks to built-in resilience features, optimized tools for configuration handling and the DAQ Doctor, an application giving advice to the operator based on continuous analysis of the monitoring data. In order to increase CMS over-all data taking efficiency, central DAQ transitions to start and stop data taking have been optimized. A new central recovery mechanism has been designed that allows recovery from single event upsets during an ongoing run.

References

[1] The CMS Collaboration, CMS Technical Proposal, CERN LHCC 94-38, 1994.
The CMS Collaboration (Adolphi R et al.), "The CMS Experiment at CERN LHC," *JINST* 3 S08004 p. 361, 2008

[2] The CMS Collaboration, CMS, The TriDAS Project, Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger, CERN/LHCC 2002-26, 2002

[3] Racz A, McLaren R, van der Bij E, *The S-Link 64 bit Extension Specification: S-Link64*, available at http://hsi.web.cern.ch/HSI/s-link

[4] Myricom, see http://www.myri.com

[5] Bellato M et al., "Run control and monitor system for the CMS experiment," presented at *Int. Conf. Computing High Energy and Nuclear Physics*, La Jolla, CA, March 24-28, 2003
Bauer G et al., "The run control system of the CMS experiment", *J. Phys.: Conf. Ser.* 119 022010, 2008

[6] Bauer G et al., "The CMS data acquisition system software," *J. Phys.: Conf. Ser.* 219 022011, 2010

[7] Boyer J, Canonical XML version 1.0, W3C Recommendation, 15 March 2001 (see also http://www.w3c.org/XML).

[8] Box D et al., Simple Object Access Protocol (SOAP) 1.1, W3C Note, 08 May 2000 (see also http://www.w3c.org/TR/SOAP).

[9] Jones C D et al., "The new CMS data model and framework," presented at *CHEP'06 - Int. Conf. Computing High Energy and Nuclear Physics,* Mumbai, India, February 13-17, 2006