Performance of the new DAQ system of the CMS experiment for Run-2

Jean-Marc André[¶], Anastasios Andronidis[†], Ulf Behrens[*], James Branson[§], Philipp Brummer[†], Olivier Chaze[†], Cristian Contescu[¶], Benjamin G. Craigs[†], Sergio Cittolin[§], Georgiana-Lavinia Darlea[‖], Christian Deldicque[†], Zeynep Demiragli[‖], Marc Dobson[†], Samim Erhan[‡], Jonathan Richard Fulcher[†], Dominique Gigi[†], Frank Glege[†], Guillelmo Gomez-Ceballos[‖], Jeroen Hegeman[†], André Holzner[§], Raúl Jiménez-Estupiñán[†], Lorenzo Masetti[†], Frans Meijers[†], Emilio Meschi[†], Remigius K. Mommsen[¶], Srecko Morovic[†], Vivian O'Dell[¶], Luciano Orsini[†], Christoph Paus[‖], Marco Pieri[§], Attila Racz[†], Hannes Sakulin[†], Christoph Schwick[†], Thomas Reis[†], Dainius Šimelevičius[**††], Petr Zejdl[¶††]

***Abstract*—The data acquisition system (DAQ) of the CMS experiment at the CERN Large Hadron Collider (LHC) assembles events at a rate of 100 kHz, transporting event data at an aggregate throughput of more than 100 GB/s to the High-level Trigger (HLT) farm. The HLT farm selects and classifies interesting events for storage and offline analysis at an output rate of around 1 kHz.**

The DAQ system has been redesigned during the accelerator shutdown in 2013-2014. The motivation for this upgrade was twofold. Firstly, the compute nodes, networking and storage infrastructure were reaching the end of their lifetimes. Secondly, in order to maintain physics performance with higher LHC luminosities and increasing event pileup, a number of sub-detectors are being upgraded, increasing the number of readout channels as well as the required throughput, and replacing the off-detector readout electronics with a MicroTCA-based DAQ interface.

The new DAQ architecture takes advantage of the latest developments in the computing industry. For data concentration 10/40 Gbit/s Ethernet technologies are used, and a 56 Gbit/s Infiniband FDR CLOS network (total throughput $\approx$ 4 Tbit/s) has been chosen for the event builder. The upgraded DAQ - HLT interface is entirely file-based, essentially decoupling the DAQ and HLT systems. The fully-built events are transported to the HLT over 10/40 Gbit/s Ethernet via a network file system. The collection of events accepted by the HLT and the corresponding metadata are buffered on a global file system before being transferred off-site. The monitoring of the HLT farm and the data-taking performance is based on the Elasticsearch analytics tool.

This paper presents the requirements, implementation, and performance of the system. Experience is reported on the first year of operation with LHC proton-proton runs as well as with the heavy ion lead-lead runs in 2015.

***Index Terms*—LHC, CMS, DAQ, HLT, Ethernet, Infiniband, event-building, network, cloud**

Manuscript received June 14, 2016.
This work was supported in part by the DOE and NSF (USA).
Corresponding author: Jeroen Hegeman (jeroen.hegeman@cern.ch)
[†]CERN, Geneva, Switzerland
[*]DESY, Hamburg, Germany
[¶]FNAL, Chicago, Illinois, USA
[‖]Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[‡]University of California, Los Angeles, Los Angeles, California, USA
[§]University of California, San Diego, San Diego, California, USA
[**]Also at Vilnius University, Vilnius, Lithuania
[††]Also at CERN, Geneva, Switzerland
978-1-5090-2014-0/16/$31.00

I. Introduction

The Compact Muon Solenoid (CMS) experiment at CERN's Large Hadron Collider (LHC) is one of two large general-purpose experiments exploring a wide range of physics at the TeV scale. A detailed description of the CMS detector can be found in Ref. [1].

The CMS online trigger-DAQ system needs to select around 1 kHz of interesting collision events out of the LHC bunch-crossing frequency of (nominally) 40 MHz. The first-level trigger, based on custom electronics, selects events at a rate of 100 kHz. The data acquisition (DAQ) system then reads out the full detector and assembles complete events. These events are transported to the High-Level Trigger (HLT) farm, where a software-based trigger reduces the output to a final 1 kHz of events for offline analysis.

During the first LHC data-taking run (run-1, 2009-2013) the CMS DAQ performed well, acquiring physics data with an overall availability of 99.6% [2]. In its first long shutdown, the LHC was upgraded to a higher center-of-mass energy and to provide higher luminosity. LHC data-taking run-2 started in 2015. The coming years various accelerator upgrades will continue to increase the luminosity by (at least) another factor two.

In order to maintain its physics performance with the increased pile-up, CMS is also performing a wide range of detector, trigger, and DAQ upgrades. The addition of new sub-detectors and the increased channel counts in upgraded detectors require an increase in the DAQ throughput. The introduction of new standards (e.g., MicroTCA) and of optical back-end links require an upgrade of the DAQ to accommodate these new technologies. A separate motivation for the CMS DAQ upgrade is to replace hardware nearing its end of life before entering run-2.

II. The CMS DAQ2 system

Following the requirements from physics performance and detector technology, the second generation CMS DAQ (DAQ2) was designed to handle an event size of 2 MB at a level-1 trigger rate of 100 kHz, resulting in a throughput requirement of 200 GB/s.

The DAQ2 architecture leverages the current state of the art in computing. The design is based on the use of high-performance nodes, connected in a flexible, high-performance layout that can be tuned to different data-taking use-cases. Nevertheless, the DAQ2 design still reflects the 'traditional' steps typically found in recent HEP DAQ systems: dedicated readout units take care of receiving and buffering incoming front-end data, followed by builder units assembling the data from all sub-detectors into a complete event.

The DAQ2 front-end readout link (FEROL) is a custom PCI Express board receiving data over optical links from upgraded sub-detectors. The FEROL interfaces to a legacy custom board, the FRL (also based on PCI Express), receiving data over copper links from legacy systems. The point-to-point links from sub-detector front-end drivers to the FEROL and FRL use a custom protocol. On its output the FEROL implements an in-FPGA, one-directional 10 Gbit/s TCP/IP connection [3], allowing connection straight into standard network devices from here on.

Data from the front-end readout links are routed to readout unit PCs (RUs) via a 10/40 GbE data-concentrator network, routing data from several FEROLs to each RU. The data-concentrator network uses a 'fat tree' layout with three switching layers in order to provide redundant routing to mitigate possible RU failures. For the event-building network, connecting the readout units to the event builder units (BUs), a 56 Gbit/s Infiniband FDR CLOS network with a total throughput of approximately 4 Tbit/s has been chosen.

The upgraded HLT processing chain is entirely file-based [4], providing a decoupling between the DAQ and the HLT. Whereas in the run-1 DAQ HLT processes were offline processes embedded in the online framework, DAQ2 treats HLT processes as independent offline jobs augmented with a few modules to facilitate online monitoring and book keeping. The file-based communication between DAQ and HLT is made possible by using a network file system (NFS4) in combination with a fast (10/40 Gbit/s) network. Each BU is equipped with a 250 GB RAM disk and a 2 TB magnetic disk. BUs and FUs are grouped in sub-clusters, and the FUs mount both BU disks over NFS. The BU writes completely-built events onto its RAM disk, from where they are picked up by the FUs and processed. Selected events are written to the FU's local disk, together with the corresponding event metadata. A hierarchical merging procedure combines all selected events and metadata from all filter processes on all FUs, and writes the results to a global, cluster file system (implemented using the Lustre file system on Netapp hardware [5]) awaiting transfer off-site to the CERN computing center.

The file-based nature of the HLT farm, and the availability of performance metrics in JSON format, allow the use of modern data analytics methods for cluster monitoring. The performance monitoring and diagnostics of the HLT farm are based on Elasticsearch. A dedicated set of nodes forms an online Elasticsearch cluster, allowing the near-real-time indexing of $O(10000)$ documents per second while maintaining a near-instantaneous query response time [6]. The possiblities of extending the JSON + Elasticsearch approach to the monitoring of the rest of the DAQ system are being investigated.

A description of the design of the CMS DAQ2 system was presented at the 2014 Realtime conference [7]. More details can be found in [4], [5], [6].

III. Event building performance

The performance of the DAQ2 event-building system has been evaluated both in a small-scale validation system and in the full-scale production system. Various system sizes (i.e., numbers of RUs and BUs) have been compared to study the scaling behavior of the system. Realistic system configurations (containing about 65 RUs and 65 BUs) deliver typical throughputs of about 3.5 GB/s per RU, for an aggregated throughput of $O(240\,\mathrm{GB/s})$. In order to achieve this performance the usage of the data-concentrator network was carefully optimized. Judicious port assignment ensures there are no hash collisions, guaranteeing a well-balanced traffic distribution over all network links. In addition the TCP stack has been tuned to the specific traffic conditions.

Two important optimizations required to achieve the above performance are highlighted in Figure 1. The left-hand side shows the improvement achieved by a careful grouping of related processes, memory structures, and interrupts and assigning these to CPU cores and memory banks using NUMA (Non-Uniform Memory Access) control commands. The right-hand side shows the effect of a custom routing scheme applied to the Infiniband event-building network. The data on this network is bursty, and for event building the slowest link becomes the bottleneck. The custom routing algorithm exploits the fact that the data flow is mostly one-directional to minimize the spread in the traffic across the spine switches, thereby reducing congestion.

The performance of the various components in the event-building chain have been extensively studied both in the validation system and in the full production system. An in-depth review of the results can be found in Ref. [8]. Figure 2 summarizes the performance of a DAQ configuration for proton-proton data-taking.

As can be seen from this figure, the presently installed system easily handles the current event size (1 MB) and trigger rate (100 kHz) requirements. In order to reach a total throughput of 200 GB/s the system can be scaled up by adding RUs and BUs.

IV. Storage and transfer systems

The original design for the storage system was to accommodate a 1 kHz HLT output rate with a (compressed) event size of 1 MB. Taking advantage of the possibility of the Lustre file system for parallel writing to the same file avoiding additional copies, an even higher throughput was achieved. Current data-taking conditions (see Figure 3) typically represent a throughput of ≈ 2 GB/s write, combined with ≈ 2 GB/s read to data quality monitoring and the transfer system. This has allowed to accommodate additional, specialized HLT output streams.

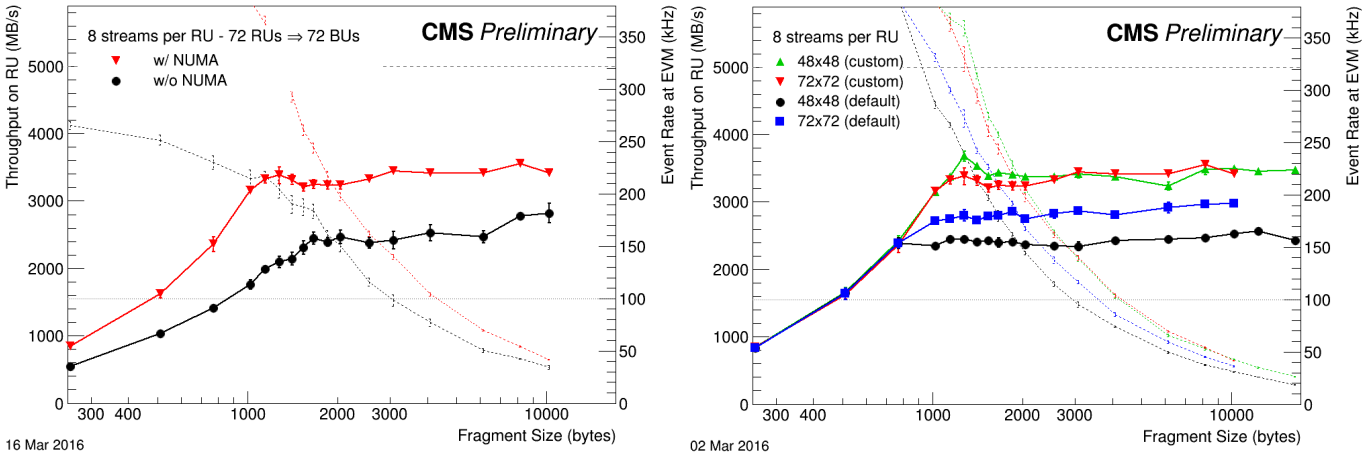

Fig. 1: Event-building performance measurements. In each figure the solid lines represent the measured throughput (corresponding to the left vertical axis) as a function of the size of the event fragments sent by the front-ends. The dashed lines represent the corresponding level-1 trigger rate (corresponding to the right vertical axis). Left: the effect of coordinated assignment of threads, memory and interrupts (top, red line) vs. no optimization (bottom, black line). Right: the effect of the custom Infiniband routing (top lines) vs. no optimisation (bottom lines) for two different event-building configurations. The red, downward-triangle markers in the left figure represent the same data points as the red, downward-triangle markers in the right figure.

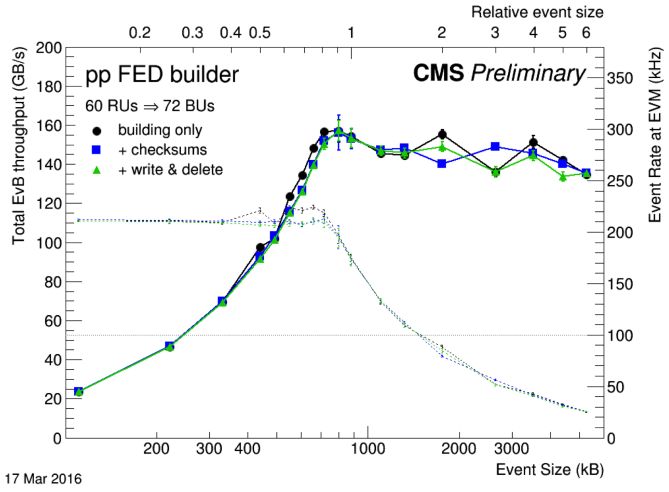

Fig. 2: Event-building performance curve, similar to Figure 1, but for a production DAQ configuration. The main difference is the fact that all front-end readout links now send different-size event fragments, leading to a very uneven distribution of load across the data-concentrator network. The CMS event size at the time of writing is approximately 1 MB. From the above curve it can be seen that there exists a 50% margin on the event size, at the 100 kHz event rate.

V. Overall DAQ2 performance

In 2015, the CMS experiment successfully collected a total of 4.4 PB of data (1.2 PB proton-proton, 1.9 PB lead-lead, and 1.3 PB auxiliary data). The central DAQ availability was $\gg 99\%$; the amount of luminosity lost due to central DAQ unavailability was negligible.

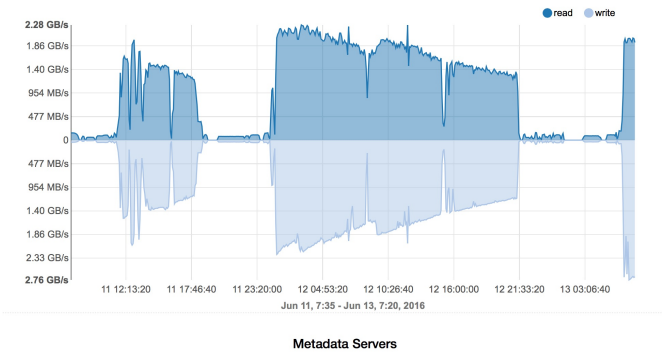

Fig. 3: Typical throughput of the Lustre global file system for proton-proton data-taking conditions. Upper part: output to data quality monitoring and the transfer system. Bottom part: input from High-Level Trigger farm. The period of June 11, 23:30 till June 12, 21:30 corresponds to a single LHC fill.

VI. Opportunistic offline usage of the HLT farm

The CMS HLT computer farm comprises 20000 cores (16000 cores until 2016 Q2), representing the same computing power as all CMS Tier-1 sites combined. Peak HLT performance is only required during high-luminosity data-taking. In order to benefit from the HLT infrastructure during e.g. maintenance periods or accelerator studies, an 'HLT cloud' overlay (based on OpenStack Grizzly) was developed. This cloud overlay allows the HLT farm to be used as Tier-1 computing site for CMS offline computing jobs. In 'cloud mode', virtual machines are deployed on a subset of the HLT compute nodes, maintaining a guaranteed level of HLT capacity for e.g., calibration runs. The HLT can host a total

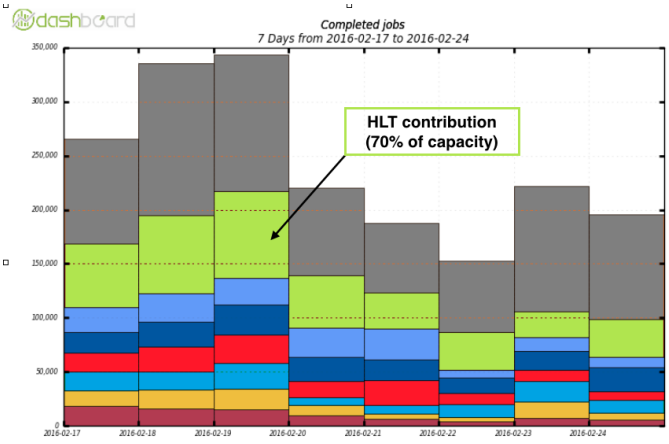

Fig. 4: Contribution of the HLT cloud to offline computing capacity during testing in 2015 [9]. Approximately 30% of the HLT capacity was kept for online purposes (e.g., calibration), the other 70% were assigned to the HLT cloud, representing the second largest contributor to the CMS offline computing capacity at the time.

of about 2000 virtual compute nodes. Figure 4 shows the contribution of the HLT cloud to the CMS offline computing capacity during a week in early 2016. The HLT contribution is only exceeded by the Fermilab Tier-1.

LHC statistics show the above approach could add about 100 days of HLT compute usage, about 30 of which were indeed recovered by CMS (excluding cloud commissioning time). A second cause of 'HLT free time' are the LHC inter-fill periods, when the accelerator is refilling and accelerating, but CMS is not taking collision data. Extensive caching and compression allows rapid deployment of the cloud overlay (cold start of 1000 VMs within 8 minutes), triggered automatically by the LHC end-of-fill signal. Once the accelerator approaches stable beams again, cloud jobs are automatically killed and compute nodes reassigned to the HLT. Careful selection of short jobs ensures this procedure remains efficient in the case of typical inter-fill periods (i.e., about six hours). A possible enhancement currently being studied is the possibility to hibernate VMs instead of killing them when data-taking starts, and then continuing the jobs in the next inter-fill period. Based on 2015 experience, CMS expects to add about 100 days of compute productivity from inter-fill cloud usage.

A more in-depth discussion of the CMS 'online cloud' design and performance can be found in Ref. [9].

VII. Summary and Outlook

In 2015, the upgraded CMS DAQ system has been fully commissioned, both in proton-proton and in lead-lead data-taking conditions. CMS DAQ2 has fulfilled all functional and performance requirements.

The DAQ2 design philosophy of a compact, highly-optimized system built from high performance components and fast and flexible interconnects has shown to work well for different data-taking configurations. This flexibility will prove invaluable with the addition of additional and upgrading sub-detectors to the CMS experiment.

Unexplored parameter space still exists for further tuning and optimization. The HLT farm has already been scaled up from 16k to 20k nodes in early 2016 to stay in step with the increasing LHC luminosity.

The first large extension of the DAQ2 system will come in the 2016-2017 year-end technical stop, in the form of the new CMS pixel detector. In preparation, the CMS DAQ group is developing an upgraded version of the FEROL front-end readout card capable of receiving up to four 10 Gbit/s front-end links, and optionally merging these data into a single 40 Gbit/s TCP/IP link. This 'FEROL40' card has a MicroTCA form factor, uses PCI Express for configuration and monitoring, and provides further integration of the DAQ systems with the upgraded Trigger Control and Distribution System (TCDS).

The flexible and scalable design of DAQ2 will play an important role in the extension of the CMS experiment over the coming years.

References

[1] S. Chatrchyan *et al.*, "The CMS experiment at the CERN LHC," *JINST*, vol. 3, p. S08004, 2008.

[2] G. Bauer *et al.*, "Automating the CMS DAQ," *Journal of Physics: Conference Series*, vol. 513, no. 1, p. 012031, 2014. [Online]. Available: http://stacks.iop.org/1742-6596/513/i=1/a=012031

[3] G. B. *et al.*, "10 Gbps TCP/IP streams from the FPGA for the CMS DAQ Eventbuilder Network," CERN, Geneva, Tech. Rep. CMS-CR-2013-416, Nov 2013. [Online]. Available: http://cds.cern.ch/record/1640924

[4] J.-M. Andre *et al.*, "File-Based Data Flow in the CMS Filter Farm," *J. Phys.: Conf. Ser.*, vol. 664, no. FERMILAB-CONF-15-601-E. 8, p. 082033. 8 p, 2015. [Online]. Available: http://cds.cern.ch/record/2134635

[5] J. Andre *et al.*, "Online data handling and storage at the CMS experiment," CERN, Geneva, Tech. Rep. CMS-CR-2015-074. 8, May 2015. [Online]. Available: http://cds.cern.ch/record/2016893

[6] J.-M. Andre *et al.*, "A scalable monitoring for the CMS Filter Farm based on elasticsearch." CERN, Geneva, Tech. Rep. CMS-CR-2015-060, May 2015. [Online]. Available: http://cds.cern.ch/record/2020877

[7] T. A. Bawej *et al.*, "The New CMS DAQ System for Run 2 of the LHC," CERN, Geneva, Tech. Rep. CMS-CR-2014-082, May 2014. [Online]. Available: https://cds.cern.ch/record/1711011

[8] K. Albertsson *et al.*, "A new event builder for cms run ii," *Journal of Physics: Conference Series*, vol. 664, no. 8, p. 082035, 2015. [Online]. Available: http://stacks.iop.org/1742-6596/664/i=8/a=082035

[9] J.-M. Andre *et al.*, "Opportunistic usage of the CMS online cluster using a cloud overlay," in *To appear in the proceedings of the International Symposium on Grids and Clouds (ISGC 2016)*, Academia Sinica - Taipei, Taiwan, 2016.