The New CMS DAQ System for Run-2 of the LHC

Tomasz Bawej, Ulf Behrens, James Branson, Olivier Chaze, Sergio Cittolin, Georgiana-Lavinia Darlea, Christian Deldicque, Marc Dobson, Aymeric Dupont, Samim Erhan, Andrew Forrest, Dominique Gigi, Frank Glege, Guillelmo Gomez-Ceballos, Robert Gomez-Reino, Jeroen Hegeman, Andre Holzner, Lorenzo Masetti, Frans Meijers, Emilio Meschi, Remigius K. Mommsen, Srecko Morovic, Carlos Nunez-Barranco-Fernandez, Vivian O'Dell, Luciano Orsini, Christoph Paus, Andrea Petrucci, Marco Pieri, Attila Racz, Hannes Sakulin, *Member, IEEE*, Christoph Schwick, Benjamin Stieger, Konstanty Sumorok, Jan Veverka, and Petr Zejdl

***Abstract*—The data acquisition (DAQ) system of the CMS experiment at the CERN Large Hadron Collider assembles events at a rate of 100 kHz, transporting event data at an aggregate throughput of 100 GB/s to the high level trigger (HLT) farm. The HLT farm selects interesting events for storage and offline analysis at a rate of around 1 kHz. The DAQ system has been redesigned during the accelerator shutdown in 2013/14. The motivation is twofold: Firstly, the current compute nodes, networking, and storage infrastructure will have reached the end of their lifetime by the time the LHC restarts. Secondly, in order to handle higher LHC luminosities and event pileup, a number of sub-detectors will be upgraded, increasing the number of readout channels and replacing the off-detector readout electronics with a μTCA implementation. The new DAQ architecture will take advantage of the latest developments in the computing industry. For data concentration, 10/40 Gb/s Ethernet technologies will be used, as well as an implementation of a reduced TCP/IP in FPGA for a reliable transport between custom electronics and commercial computing hardware. A Clos network based on 56 Gb/s FDR Infiniband has been chosen for the event builder with a throughput of $\sim$ 4 Tb/s. The HLT processing is entirely file based. This allows the DAQ and HLT systems to be independent, and to use the HLT software in the same way as for the offline processing. The fully built events are sent to the HLT with 1/10/40 Gb/s Ethernet via network file systems. Hierarchical collection of HLT accepted events and monitoring meta-data are stored into a global file system. This paper presents the requirements, technical choices, and performance of the new system.**

***Index Terms*—Data acquisition, high energy physics.**

Manuscript received June 27, 2014; revised March 24, 2015; accepted April 10, 2015. Date of publication May 21, 2015; date of current version June 12, 2015. This work was supported in part by the Department of Energy (DOE) and the National Science Foundation (NSF) (USA).

T. Bawej, O. Chaze, C. Deldicque, M. Dobson, A. Dupont, A. Forrest, D. Gigi, F. Glege, R. Gomez-Reino, J. Hegeman, L. Masetti, F. Meijers, E. Meschi, S. Morovic, C. Nunez-Barranco-Fernandez, L. Orsini, A. Petrucci, A. Racz, H. Sakulin, C. Schwick, B. Stieger, and P. Zejdl are with CERN, 1211 Geneva, Switzerland (e-mail: hannes.sakulin@cern.ch).

U. Behrens is with DESY, 22607 Hamburg, Germany.

J. Branson, S. Cittolin, A. Holzner, and M. Pieri are with the University of California San Diego, La Jolla, CA 92093 USA.

G. Darlea, G. Gomez-Ceballos, C. Paus, K. Sumorok, and J. Veverka are with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

S. Erhan is with the University of California, Los Angeles, CA 90095 USA.

R. K. Mommsen and V. O'Dell are with FNAL, Batavia, IL 60510-5011 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNS.2015.2426216

I. Introduction

THE Compact Muon Solenoid (CMS) experiment [1], [2] at CERN's Large Hadron collider is one of two large general-purpose experiments exploring a wide range of physics at the TeV scale. Its on-line systems need to select around 1 kHz of interesting events out of the LHC bunch-crossing rate of nominally 40 MHz. At CMS, the event selection is done with only two trigger levels: a first-level trigger [3], based on custom electronics, reduces the rate to 100 kHz. The data acquisition (DAQ) system [4], [2] then reads out the detector and assembles full events at 100 kHz, passing the assembled events to the high-level trigger, a software system based on the full CMS reconstruction software, running on a farm of computers. At a nominal event size of 1 MB, the CMS DAQ system for Run-1 was designed to handle a throughput of 100 GB/s, making it the highest throughput DAQ system in high-energy physics to date. The DAQ system has performed successfully during Run-1 of the LHC from 2009 to 2013, acquiring physics data with an availability of 99.6% [5].

During its current first long shut down, the LHC is being upgraded to provide higher center-of-mass energy and higher luminosity during Run-2, from 2015 onwards. CMS therefore needs to prepare for higher pile-up and thus higher event size. Some of the CMS sub-systems are currently upgrading their detectors and/or on-line systems and are building new off-detector read-out electronics based on the µTCA standard. The DAQ system for Run-2 will need to be able to read out these off-detector electronics though a new optical readout-link at a higher bandwidth than in Run-1. Since many of the compute nodes and part of the networking infrastructure of the DAQ system for Run-1 are close to the end of their life cycle, we decided to completely re-design the DAQ system exploiting the considerable advances in computing and network technology to build a much more compact and even more powerful DAQ system. In this paper we present the final design of the new DAQ system that is currently being installed, and report on performance measurements including first measurements in the production system.

In Section II we give an overview of the main design parameters of the new CMS DAQ system. In Sections III, IV, V and VI, we then report in more detail on the frontend-readout optical link including a 10 Gb/s TCP/IP sender implemented in an FPGA and the SLINK-Express DAQ readout link, the data concentrator network based on 10/40 Gb/s Ethernet, the core

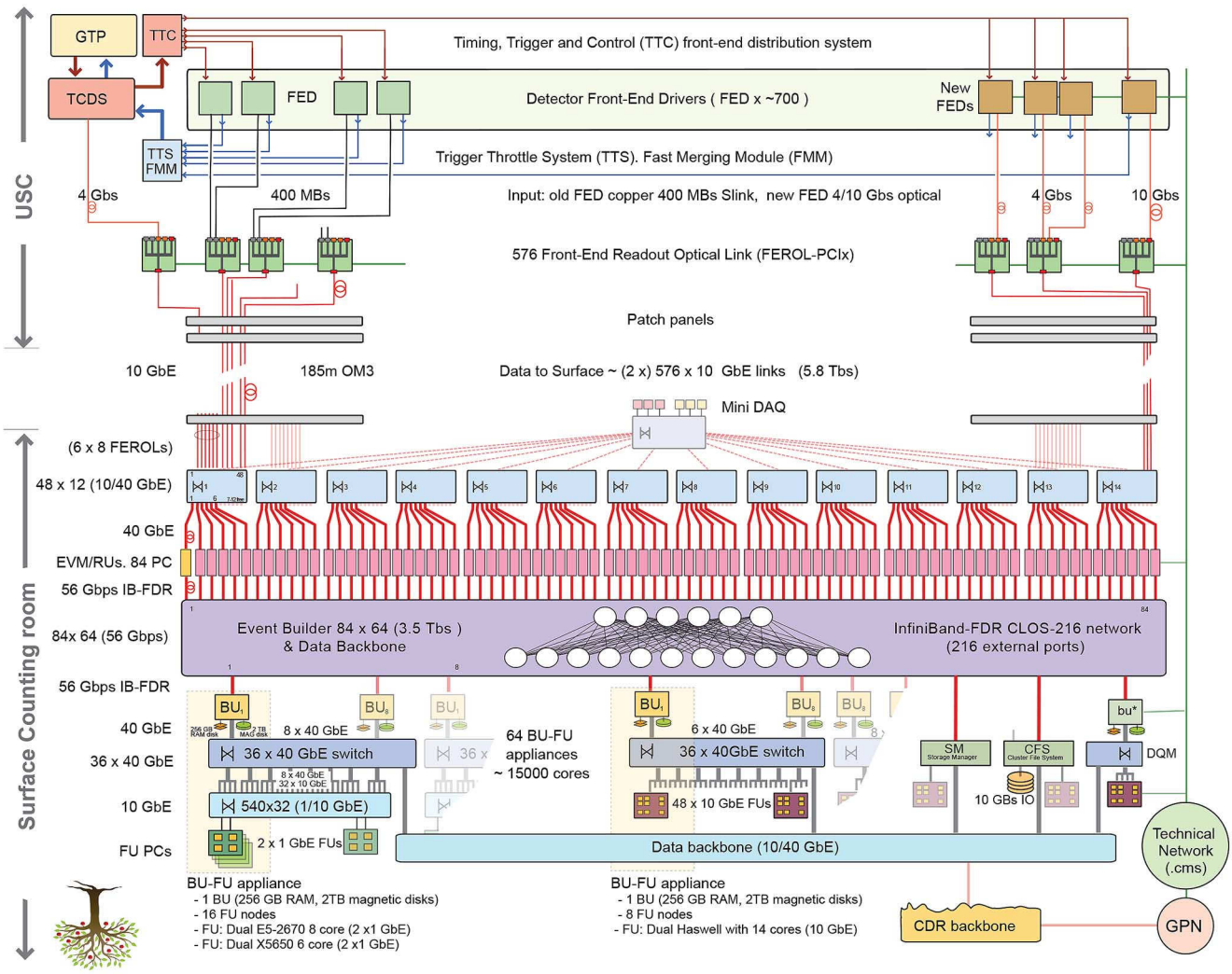

Fig. 1. Over-all architecture of the new CMS DAQ system for Run-2 of the LHC. See text.

event builder based on Infiniband and our file-based way of integrating the high-level trigger. In each of the sections we report on the requirements, design and on latest performance measurements.

II. Main Design Parameters

Table I compares the main design parameters of the CMS DAQ system for Run-1 (DAQ-1 [2]) and of the new CMS DAQ system for Run-2 (DAQ-2). The readout rate will remain at 100 kHz as it is limited by on-detector electronics of several CMS sub-systems. In order to cater for higher event sizes, fragments of up to 4 kB will be supported for legacy readout electronics based on SLINK-64 [6]. For readout electronics based on μTCA, the new readout-link standard SLINK-express [7] will be supported with fragment sizes up to 8 kB. The total bandwidth of the DAQ system will be doubled to 200 GB/s, allowing for event sizes up to 2 MB, which includes a large margin. As in Run-1, the new DAQ-2 system will build events in two stages (see Fig. 1). But unlike in Run-1, where a cost-effective switch with the required capacity was not available, advances in network technology made it possible to implement the stage-2 (core) event builder with a single network. The first stage of the DAQ-2 event builder is thus a pure data concentrator. The development of a custom 10 Gb/s network card sending TCP/IP from an FPGA made it possible to interface the custom electronics of the DAQ readout link to a commercial 10/40 Gb/s Ethernet network handling the data concentration. The high throughput per port in the event builder makes it impractical to combine Builder Units (BU) and Filter Units (FU) on the same machines as done in Run-1, since the throughput into FUs is limited by processing power on the FU. In Run-2, the high-level trigger will therefore run on dedicated FU machines connected to the BUs via 1/10/40 Gb/s Ethernet.

III. The New Front-End-Readout Optical Link

During Run-1, CMS sub-detectors were read out exclusively through the SLINK-64 DAQ readout link, an LVDS-based copper link capable of transferring up to 400 MB/s. One or two such links are received by the Frontend Readout Link (FRL), a custom Compact-PCI card, which forwarded the data to a commercial Myrinet NIC via an internal PCI-X bus. Super-fragments were then assembled using a commercial Myrinet network running custom firmware on the NICs to do the super-fragment building. Assembled super-fragments were

TABLE I
MAIN PARAMETERS OF THE CURRENT (DAQ-1) AND RUN-2 (DAQ-2) CMS DAQ SYSTEMS

	DAQ-1	DAQ-2
Readout rate	100 kHz	100 kHz
Front-End-Drivers, type of readout link, fragment size	640: SLINK-64 1 – 2 kB	640: SLINK-64, 1 – 4 kB 50: SLINK-Express, 2 – 8 kB
Total DAQ bandwidth	100 GB/s	200 GB/s
Event Builder Stage 1 Data concentration	2x 2.5 Gb/s Myrinet (commercial hardware, custom firmware & protocol)	10/40 Gb/s Ethernet (custom sender card, commercial TCP/IP protocol)
Number & type of readout-unit PCs	640 (8 x 80)	84
Event Builder Network	1x/2x/3x 1 Gb/s Ethernet	56 Gb/s FDR Infiniband
Number of parallel event builders (slices)	8	1
Event Builder topology	1 switch per slice	1 Clos network of 18 switches
Number of Builder PCs	1260 (8 x 158) builder + filter in one	64, builder only
Number of HLT cores	13000	~15000 initially
Interface to the HLT	Shared memory	Files
Storage	16 SAN systems	1 cluster file system
Bandwidth to storage	1.2 GB/s write + 1 GB/s read	2 GB/s write + 1 GB/s read
Storage capacity	~ 250 TB	~ 250 TB

then transferred into the memory of a Readout Unit (RU) PC via DMA.

In the Run-2 DAQ-system, the Myrinet NIC on the FRL is replaced by a custom developed PCI-X card, the Frontend Readout Optical Link (FEROL) [8]. The FEROL acts as a 10 Gb/s Ethernet NIC sending data to the event builder. It receives data either via the PCI-X interface from the FRL (from the bulk of the sub-systems that continue to use SLINK-64) or via optical SLINK-express inputs from new or upgraded sub-systems. Up to two SLINK-express inputs at 6 Gb/s or one input at 10 Gb/s will be supported. While SLINK-64 senders are mezzanine cards plugged onto the off-detector readout electronics of the sub-detectors, the SLINK-express sender is a firmware IP-core that is included in the FPGAs of sub-detector readout electronics. Many upgraded sub-detectors are planning to use a common μTCA readout-board, the AMC-13 [9], which includes this IP core. SLINK-express works with 8b/10b encoding at up to 5.0 Gb/s or 64b/66b encoding at 10.3 Gb/s resulting in an effective bandwidth of up to 4.0 Gb/s or 10.0 Gb/s, respectively. The data format (definition of headers and trailers) is identical to that of SLINK-64. SLINK-express is packet based and supports retransmission at the packet level. Packets have a variable size of at most 4 KiB. Fragments up to a size of 4 KiB are transferred in individual packets, while larger fragments are split across multiple packets.

At the output side, the FEROL sends data via TCP/IP employing a custom TCP/IP engine implemented in FPGA logic. This has been achieved though a simplification of the TCP/IP protocol for unidirectional use, which reduces the number of states from 11 to 3 [10]. TCP/IP streams (one per input) are sent via a commercial 10/40 Gb/s Ethernet switch to a commercial

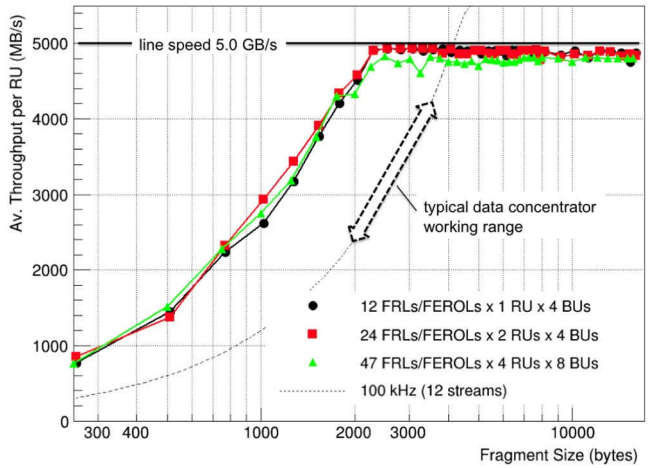

Fig. 2. Throughput per Readout Unit as a function of fragment size for fixed size fragments in saturation mode.

40 Gb/s Ethernet NIC in the Readout Unit (RU) PC where they can be received with the standard Linux TCP/IP stack. In a test setup sending data from a single FEROL through a switch to a receiving PC running Linux sockets with performance tuning as described in [8], a sustained point-to-point throughput of 9.7 Gb/s was shown to be achievable for fragments larger than 1 kB [8].

IV. DATA CONCENTRATION

As shown in Fig. 1, the data concentration layer consists of individual 10/40 Gb/s Ethernet switches (Mellanox MSX1024) receiving data at 10 Gb/s from up to 48 FEROLs each, sending the data at 40 Gb/s to up to 6 Readout Unit PCs. Readout links from legacy off-detector electronics will be typically merged by 12, allowing for up to 4 kB fragment size, while readout links from upgraded detectors will be merged less aggressively, allowing for larger fragment sizes up to 8 kB. In order to increase tolerance to failures of readout unit PCs, the switches will be interconnected with a small number of additional switches in a fat tree [11] topology so that data may also be sent to a Readout Unit on a neighboring switch. As RU PCs we are using the Dell PowerEdge R620 machines with dual 12-core E5-2670 CPUs and dual Mellanox ConnectX-3 NICs (one configured for Ethernet, one for Infiniband). The software receiving TCP streams from the FEROLs is based on Linux sockets. In order to achieve full throughput at 40 Gb/s, two threads are used for receiving data. CPU affinities of threads, interrupts and memory are fine-tuned manually [12].

Fig. 2 shows a measurement of the throughput of the RU as a function of the fragment size for fixed size fragments when the RU is merging fragments from 12 FEROLs sending 1 stream each. Setups with 1, 2 or 4 RUs connected to the same 10/40 Gb/s Ethernet switch, each RU receiving data from 12 FEROLs, show similar throughput per RU, demonstrating that the switch is non-blocking. The setups also included the full core event builder (see Section IV) with 4, 4 and 8 Builder Units, respectively. However, the core event builder has no influence on this measurement at fragment sizes up to 2 kB, and only a small influence above. In this measurement, FEROLs were sending fragments as fast as possible. Up to about 2 kB fragment size, the observed throughput is limited by the overhead of TCP/IP in the data concentrator, while at

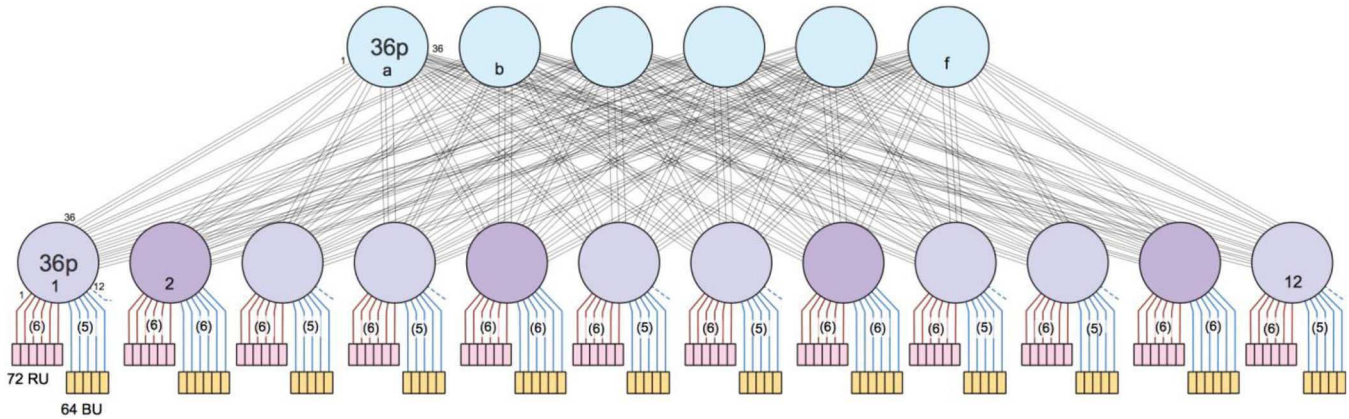

Fig. 3. Infiniband Clos network composed of 12 leaf and 6 spine switches with 36 ports each.

higher fragment sizes almost the full capacity of the 40 Gb/s link is used. The required rate of 100 kHz can be achieved for fragment sizes of up to 3.5 kB when merging from 12 sources (using 90% of the bandwidth). In the presented measurement, Ethernet pause frames are the main mechanism for flow control. Alternatively, FEROLs also support congestion control by means of a local TCP congestion window.

V. The Core Event Builder

The core event builder of the new DAQ system is based on 56 Gb/s Fourteen Data Rate (FDR) Infiniband. Infiniband is a high-speed, low-latency interconnect used in data centers. It is currently the most popular interconnect in the top-500 supercomputers. FDR refers to that fact that multiple (in our case 4) lanes at 14 Gb/s are combined to achieve the total bandwidth. Infiniband supports credit-based flow control, which enables it to efficiently avoid congestion without the need for large buffer memories in the switches. This feature also allows us to construct a large network in a cost effective way by combining small switches.

The core event builder of the DAQ-2 system consists of 84 Readout Unit (RU) PCs by 64 Builder Unit (BU) PCs. The switching fabric is composed of 18 individual 36-port switches, 12 used as leaves and 6 used as a spine arranged in a Clos [13] network as shown in Fig. 3. The switching fabric has a total bandwidth of 6 Tb/s full duplex out of which we are using 3.5 Tb/s from RUs to BUs. Leaf switches are shared between RUs and BUs allowing for more effective use of the leaf-to-spine connections. With 216 external ports, the switch fabric leaves room for future extensions of the event builder, should requirements change. As BUs we are using Dell PowerEdge R720 machines, equipped with the same CPUs and NICs as the RUs but with additional RAM (see Section VI). Our online-software is based on the Verbs API [14]. Data are transferred by RDMA [15] avoiding the need for copies. CPU affinities are fine-tuned manually for threads, interrupts and memory. While the event builder software for Run-1 was mostly single-threaded, we now use 4 threads on the RUs to pack data and 6 threads on the BUs to assemble and write data.

Very early measurements with a part of the installed network (up to 48 RUs by 48 BUs) show that for the typical expected super-fragment size of 16 to 32 kB, a per-node throughput above 4 GB/s can be achieved (Fig. 4). These measurements show the simultaneous sending and reception of streams without any event building at this stage. These measurements make us confident that the nominal overall event builder throughput of 200 GB/s can be easily achieved with this network.

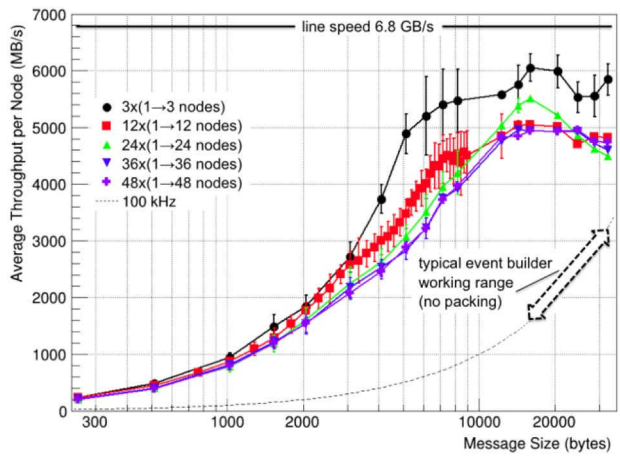

Fig. 4. Throughput per node on the Infiniband Clos network as a function of the message size for fixed size messages. Data are simultaneously streamed from n sources to n destinations with sources and destinations residing on separate leaf switches. No event building is performed.

VI. File-based Filter Farm

The CMS High-Level Trigger (HLT) performs online event selection using the full reconstruction framework CMSSW [16], which is primarily geared to off-line use, typically reading input data from files. For on-line use in Run-1, the event selection code has been integrated into an online-software application based on the XDAQ framework [17], which is responsible for data transport in the DAQ system. The two software frameworks therefore needed to be tightly coupled, forcing us to use the same compiler version and externals and to align our release schedules. This tight coupling also made debugging more complex, as in many cases knowledge about both frameworks was required.

For Run-2 we are aiming at completely separating the two frameworks [18]. Builder Unit PCs will write the assembled events to a local RAM-disk. The event-selection code will run on Filter Unit (FU) PCs connected to the BU PCs via a 1/10/40 Gb/s Ethernet network as shown in Fig. 1. FUs will be statically assigned to a certain BU, the ensemble of a BU and its FUs being called an appliance. FUs will mount the RAM disk of their BU via a network file system, so that CMSSW processes can run exactly as in off-line mode, reading input data

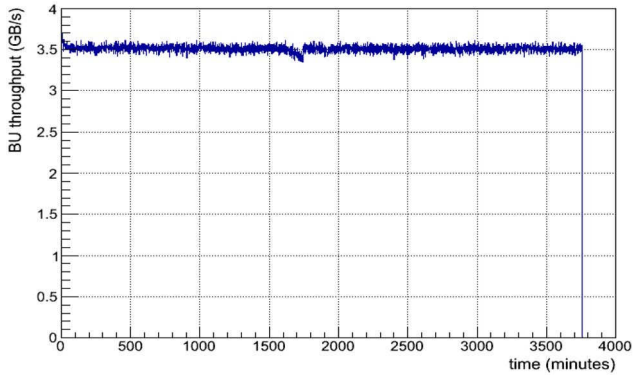

Fig. 5. Throughput of a BU over time in a test setup of a filter farm appliance.

from files. A RAM disk size of 256 GB per BU, corresponding to about 2 minutes of data, will allow us to decouple the event builder from the high-level trigger so that back-pressure can be avoided, especially during start-up of the HLT. With the current single-threaded version of CMSSW, we will start two processes per core (one per hyper-thread) to fully load the machines, as done in Run-1. A multi-threaded version of CMSSW and the corresponding HLT modules are being developed which will allow us to fully load the machines with fewer processes.

Fig. 5 shows a measurement of the BU throughput in a test setup in which 4 RUs are sending data to an appliance consisting of 1 BU and 8 FUs, the latter running 32 CMSSW processes each. BUs are writing all data to RAM disk. FUs have the RAM disk mounted via NFS-4 and are reading concurrently. A constant throughput of 3.5 GB/s was measured for one appliance, which will allow us to achieve over 200 GB/s with 64 appliances.

The filter farm will initially be made up of three generations of machines: the Westmere and Sandy Bridge based Dell PowerEdge C6100 and C6220 machines that were purchased in 2011 and 2012 and a new type of machine to be purchased in 2015. For the first two generations, two 1 Gb/s links will be sufficient to feed the FU. We will reuse the Force-10 1/10 Gb/s Ethernet switches used in the Run-1 event builder to connect these two generations of machines to the main 10/40 Gb/s switches. The third generation of machines will be directly connected with a 10 Gb/s link.

VII. Data Collection and Storage

The CMSSW processes will write their output files to the local hard disk of the FU, producing one file per process and per HLT stream (typically around 10 streams are used) in each luminosity section–a period of 23 s that is used as the quantum for data certification. These files will then be merged in two stages: The first merging stage will merge the output files from all processes running on a FU and copy them back to a hard disk on the BU. On the BU, the second merger stage will merge the output files of all FUs in the appliance and write the results to a storage system running a cluster file system. We are planning to use Lustre [19] as a cluster file system. With Lustre it will be possible to perform the final merge by simultaneously writing data to the same file from all BUs, so that the storage system will only need to sustain the write throughput of 2 GB/s in total from 64 BU nodes and a simultaneous read throughput of 1 GB/s for transfers to the CERN computing center. Measurements on an evaluation system at NetApp comprising 15 clients and 4 Object Storage Servers running Lustre 2.4 show that this technology can meet our requirements. The data collection and storage system will also collect and merge special streams and histograms for data quality monitoring.

VIII. Status and Outlook

The DAQ system of the CMS experiment has been re-designed during the ongoing first long shutdown of the LHC. By relying on modern networking technologies such as 10/40 Gb/s Ethernet and 56 Gb/s FDR Infiniband, the event builder now comprises an order of magnitude fewer machines with respect to the previous DAQ system, while its bandwidth was doubled to 200 GB/s. Installation of the event builder is complete and all front-end readout links have been switched over to the new event builder. First performance measurements of individual parts of the system indicate that the nominal performance can be achieved or even exceeded. The system will be commissioned and completed with a new generation of Filter Units over the remainder of 2014 and early 2015.

References

[1] The CMS Collaboration, *The Compact Muon Solenoid Technical Proposal*. Geneva, Switzerland: CERN, 1994, , CERN-LHCC-94-38.
[2] The CMS Collaboration: R. Adolphi *et al.*, "The CMS experiment at CERN LHC," JINST, vol. 3, 2008, Art. ID S08004.
[3] The CMS Collaboration, *CMS The TriDAS Project: Technical Design Report, Volume 1: The Trigger Systems*. Geneva, Switzerland: CERN, 2000, CERN-LHCC-2000-038.
[4] The CMS Collaboration, *The TriDAS Project, Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger*. Geneva, Switzerland: CERN, 2002, CERN-LHCC-2002-026.
[5] G. Bauer *et al.*, "Automating the CMS DAQ," *J. Phys. Conf. Ser.*, vol. 513, 2014, Art. ID 012031.
[6] A. Racz, R. McLaren, and E. van der Bij, "The S-Link 64 bit extension specification: S-LINK64," *CERN*, 2003 [Online]. Available: http://hsi.web.cern.ch/HSI/s-link/spec/
[7] A. Racz, "New CMS DAQ link development," presented at the xTCA Interest Group Meeting, TWEPP, Perugia, Italy, Sep. 2013 [Online]. Available: http://indico.cern.ch/event/228972/session/23/contribution/234
[8] G. Bauer *et al.*, "10 Gbps TCP/IP streams from the FPGA for the CMS DAQ event builder network," *JINST*, vol. 8, 2013, Atr. ID C12039.
[9] E. Hazen *et al.*, "The AMC13XG: A new generation clock/timing/DAQ module for CMS MicroTCA," *JINST*, vol. 8, Art ID C12036.
[10] G. Bauer *et al.*, "10 Gbps TCP/IP streams from the FPGA for high energy physics," *J. Phys. Conf. Ser.*, vol. 513, Jun. 2014, Art. ID 012042.
[11] C. E. Leiserson, "Fat-trees: Universal networks for hardware-efficient supercomputing," *IEEE Trans. Comput.*, vol. C-34, no. 10, pp. 892–901, Oct. 1985.
[12] T. Bawej *et al.*, "Achieving high performance with TCP over 40GbE on NUMA architectures for CMS data acquisition," *IEEE Trans. Nucl. Sci.*, 2015, submitted for publication.
[13] C. Clos, "A study of non-blocking switching networks," *Bell Syst. Tech. J.*, vol. 32, pp. 406–424, Mar. 1953.
[14] T. Bawej *et al.*, "Boosting event building performance using infiniband FDR for CMS upgrade," *PoS (TIPP2014) 190*, 2014.
[15] P. Culley, D. Garcia, J. Hilland, B. Metzler, and R. Recio, "A remote direct memory access protocol specification," *IETF RFC-5040*, 2007 [Online]. Available: http://www.ietf.org/rfc/rfc5040.txt
[16] C. D. Jones *et al.*, "The new CMS data model and framework," in *Proc. CHEP'06 Conf.*, 2007.
[17] G. Bauer *et al.*, "The CMS data acquisition system software," *J. Phys. Conf. Ser.*, vol. 219, 2010, Art. ID 022011.
[18] G. Bauer *et al.*, "Prototype of a file-based high-level trigger in CMS," *J. Phys. Conf. Ser.*, vol. 513, Jun. 2014, Art. ID 012025.
[19] The Lustre filesystem, [Online]. Available: http://www.lustre.org