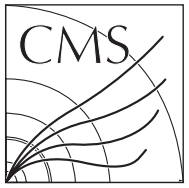
Available on CMS information server

CMS CR -2012/172



The Compact Muon Solenoid Experiment

Conference Report

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

10 July 2012 (v2)

Upgrade of the CMS Event Builder

G Bauer[6)], U Behrens[1)], M Bowen[2)], J Branson[4)], S Bukowiec[2)], S Cittolin[4)], J A Coarasa Perez[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 C Spataru[2)], K Sumorok[6)]

Abstract

The Data Acquisition system of the Compact Muon Solenoid experiment at CERN assembles events at a rate of 100 kHz, transporting event data at an aggregate throughput of 100 GB/s. By the time the LHC restarts after the 2013/14 shut-down, the current computing and networking infrastructure will have reached the end of their lifetime. This paper presents design studies for an upgrade of the CMS event builder based on advanced networking technologies such as 10/40 Gb/s Ethernet and Infiniband. The results of performance measurements with small-scale test setups are shown.

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, Princeton University, New Jersey,USA

Upgrade of the CMS Event Builder

G Bauer[6], U Behrens[1], M Bowen[2], J Branson[4], S Bukowiec[2], S Cittolin[4], J A Coarasa Perez[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], 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

[a] Now at Princeton University, Princeton University, New Jersey,USA

E-mail: Andrea.Petrucci@cern.ch

Abstract. The Data Acquisition system of the Compact Muon Solenoid experiment at CERN assembles events at a rate of 100 kHz, transporting event data at an aggregate throughput of 100 GB/s. By the time the LHC restarts after the 2013/14 shut-down, the current computing and networking infrastructure will have reached the end of their lifetime. This paper presents design studies for an upgrade of the CMS event builder based on advanced networking technologies such as 10/40 Gb/s Ethernet and Infiniband. The results of performance measurements with small-scale test setups are shown.

1. Introduction

The Compact Muon Solenoid (CMS) [1] is a general-purpose particle detector designed to study both proton-proton and heavy ion collisions produced at the Large Hadron Collider (LHC) [2] at CERN in Geneva, Switzerland. In CMS a rejection power of O(10^5), due to bandwidth and storage limitation, is required in order to reduce the event rate from the 40 MHz LHC beam crossing frequency to an acceptable rate of O(100) Hz for physics analysis. The detector comprises about 71 million readout channels. Online event-selection is performed using only two trigger levels: a hardware-based first-Level-1 Trigger (L1T) and a software-based high-level trigger (HLT). The L1T is implemented using custom electronics and reduces the initial event rate by a factor of 400 [3]. In the CMS Data Acquisition (DAQ) system [4] events are built in two stages: Front End Driver (FED) Builder and Readout Builder. The HLT is analysing full-granularity detector data using software reconstruction and filtering algorithms on a large computing cluster consisting of commercial processors. The selected events are forwarded to a storage manager (mass storage) and later sent to Tier-0. The Tier-0 processing system is the initial stage of the multi-tiered computing system and it is responsible for the first processing steps of data.

In the last 3 year of LHC operation CMS has successfully been recording proton-proton collisions at a center-of-mass energy of 7 TeV (2010 and 2011) and at 8 TeV (2012) with 50 ns bunch spacing. In

order to reach the energy of 14 TeV and a luminosity of 2 x 10^{34} $cm^{-2}s^{-1}$ the LHC machine needs to be upgraded and a long shutdown is planned during 2013-2014. In the years following the restart of LHC the pileup could reach 100 overlapping interactions and correspondingly larger event sizes will be generated due to the higher occupancy. During the long shutdown some CMS sub-detector front-end electronics and readout systems will be upgraded using μTCA-based [5] systems, and a new L1 trigger system will be deployed and operated in parallel to the existing system. The main motivations for the upgrade of CMS DAQ system are to accommodate sub-detectors with upgraded off-detector electronics and aging of existing hardware. PCs and networks are approximately five years old. The upgrade plans for the DAQ system are to replace FED Builder and RU Builder networks with more recent network technologies. The DAQ team has started feasibility studies on advanced networking technologies to identify the network technology to use for the upgrade of the event builder networks.

This paper presents the current architecture of CMS DAQ system, and the possible options for the upgrade of the system after the long shutdown. Preliminary results are included from performance tests of the 10-Gigabit-Ethernet (10 GE).

2. CMS Data Acquisition system

The online applications are based on the XDAQ [5] framework: a software platform designed specifically for the development of distributed data acquisition systems. The framework is a software middleware that eases the tasks of designing, programming and managing data acquisition applications by providing a simple, consistent and integrated distributed programming environment.
The architecture of the data acquisition (DAQ) system is shown in **Figure 1**. The 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) at the level-1 trigger rate of 100 kHz.

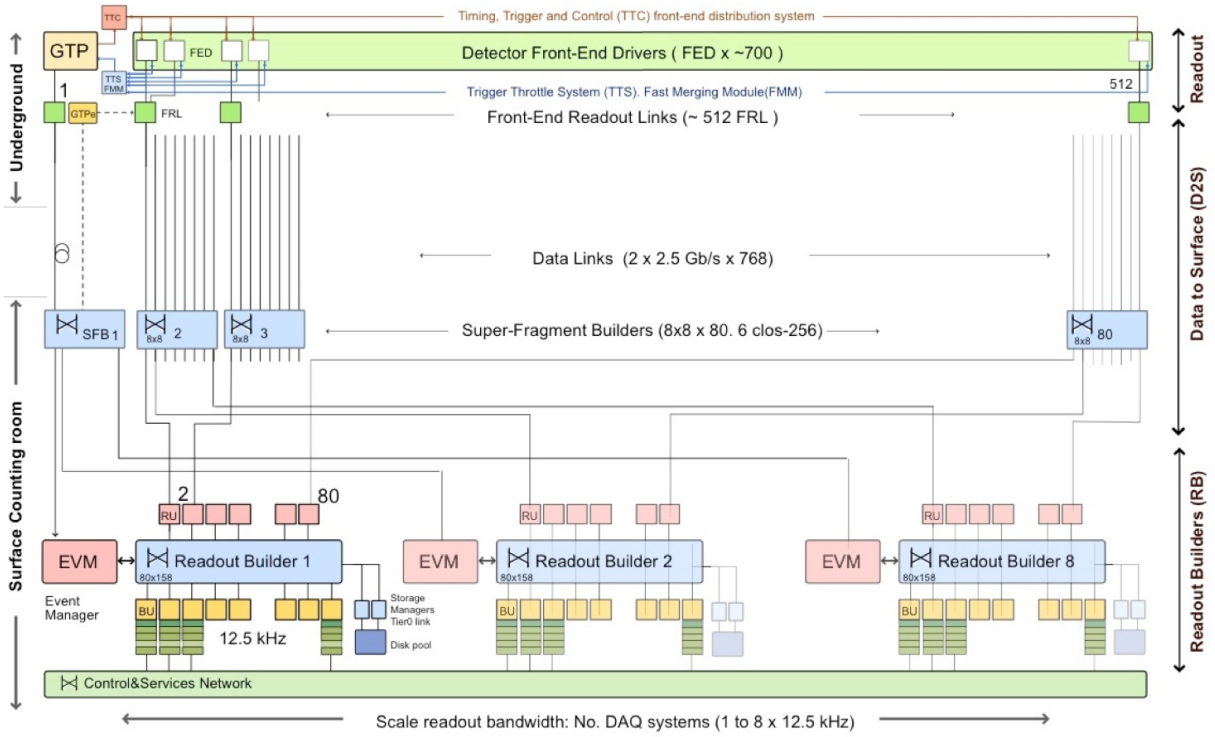


Figure 1. The CMS data acquisition architecture. Events are built in two stages: super-fragments are built in the Readout Unit (RU), full events are built in the Builder Unit (BU). The BUs also run the high-level trigger software.

Data are transferred using SLINK-64 [7] copper links from the sub-detector specific FEDs with SLINK-64 sender card to a common Frontend Readout Link (FRL) modules (see **Figure 2**). Some FRLs read out two FEDs with smaller fragment size in order to provide nominal fragment sizes. The total number of FRLs is about 435. The event building is performed in two different stages: FED Builder and Readout Builder. In the first stage, a super-fragment is built from the output of eight FRLs using an optical network based on Myrinet technology. The network is composed of crossbar switches and NICs connected by point to point bi-directional links. It employs wormhole routing and flow control at the link level. 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. The FED Builder output NICs are hosted by the Readout Builder PCs. In order to build the super-fragments the PCs are programmed to concatenate fragments with the same event number from all the connected input cards. The FED Builder is lossless due to a basic flow control and retransmission protocol, implemented on the RISC processor on the NIC.

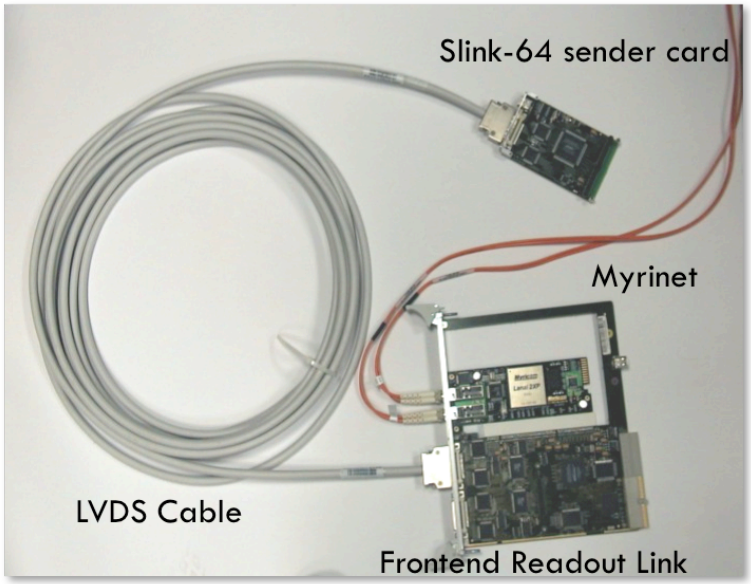


Figure 2. Photograph of S-link64 sender card, LVDS cable and compact-PCI FRL card with embedded Myrinet card. Connected to the NIC are fibres that go to the FED Builder input switch.

The second part the of event builder system assembles super-fragments into complete events and is composed of eight different Readout Builders, also called slices. Each Readout Builder must collect event super-fragments of average size about 16 kByte from about 60 data sources and build them into complete events at a rate of 12.5 kHz. A slice is made up of Read-out Units (RU), Builder Units (BU) and a single Event Manager (EVM) connected together by a Force10 Gigabit Ethernet switch using three NIC ports for the RU PCs and one or two NIC ports on the BU PCs. The EVM supervises the data flow in the RU Builder and receives a data record from the GTP via a dedicated FED Builder. The EVM allocates events on request to a BU, which subsequently collects the super-fragments from all RUs. Each of the Readout Builders operates independently from the others and fragments are assigned to one of the eight readout builders based on a look-up table, which may be adjusted in order to accommodate Readout Builders with different performance.

The Builder Unit PCs also run the high-level trigger software. An independent process, running the physics algorithms, subsequently analyzes each event. Accepted events are forwarded to the Storage Manager System (SM) for storage in a large disk pool. Stored events are transferred over a redundant 10 Gbps fiber optic connection to the CERN computer center (Tier-0), where they are processed for analysis and archived to a mass storage system.

3. Upgrade of the CMS DAQ

During the long shutdown CMS Hadron Calorimeter (HCAL) [8] front-end electronics and readout systems will be upgraded and a new L1 trigger system will be deployed and operated in parallel with the existing system. At the end of 2016 a new CMS Pixel [9] detector will be installed and readout systems will be upgraded. The upgrade plans for the DAQ system are:

- replace the current FED Builder based on Myrinet with 10 GE;
- upgrade the RU Builder networks with more recent network technologies (10/40 GE or Infiniband).

3.1. Readout link

The readout system for the HCAL, L1 trigger and Pixel will be based on µTCA with about 120 new FEDs (37 for L1 Trigger and HCAL - 80 for Pixel). Data will be transferred using the 10 Gbps optical link from sub-detector specific FEDs to common FRL modules. The expected maximum fragment size is up to 8 kB. Other sub-detectors will using the current, legacy FEDs with a total number of about 552 FEDs. The fragment size expected for the legacy FED ranges between 2 kB and 4 kB due to pile-up. The total number of FRLs will be around 480 with 360 FRLs for legacy FEDs with data transfer rates of 400 MB/s (LVDS cable) and 120 FRLs for new FEDs with data transfer rate of 800 MB/s. New DAQ readout system must be compatible with legacy data sources (FEDs/SLINK-64) and with the newly designed FEDs (µTCA/10 GE).

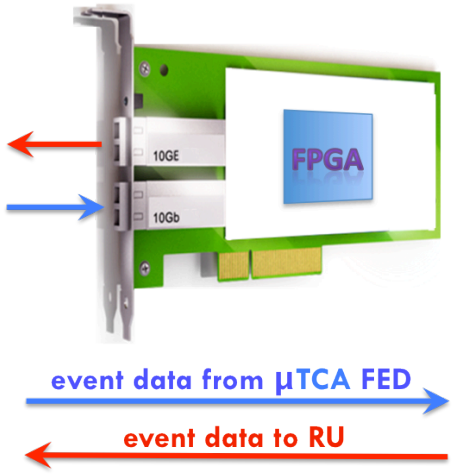


Figure 3. PCI card for existing FRL used to transfer fragments from µTCA FED or legacy FED to RU PCs.

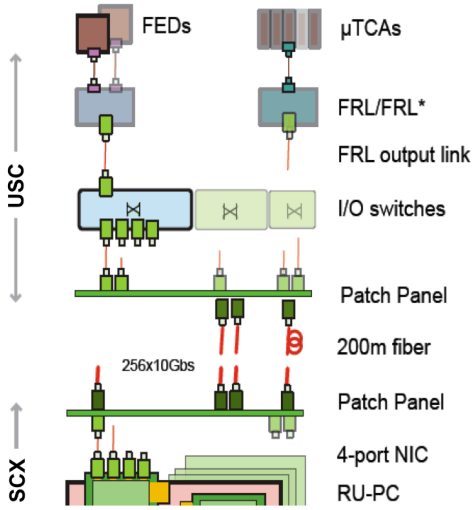


Figure 4. New Data-to-surface architecture with accommodating both types of FEDs.

A new PCI card will be built for the existing FRL, called Front-End Readout Optical Link (FEROL). The FEROL provides a 10 GE link to transfer fragments to the RU PCs and is able to receive data either via a fiber from newly built µTCA FEDs (see **Figure 3**) or from the FRL with the S-Link64 using PCI-x [10] bus. The architecture of the new Data to surface (D2S) is shown in **Figure 4**. The FEDs and FRL electronics are located in the underground electronics room (USC) instead the RU PCs are located in the surface building (SCX). The new FED Builder will use an optical network based on 10 GE with a point-to-point protocol. FRLs sources can be concentrated by I/O switches in USC or/and RU PC.

3.2. Event builder

The new event builder will be composed of the FED Builder and a single Readout Builder. In this scheme the data concentration of FRL sources in the FED Builder depends on the Readout Builder network technology. Two options are considered:

- Conservative: use 10 GE for Readout Builder network;
- Advanced: use Infiniband or 40 GE for Readout Builder network.

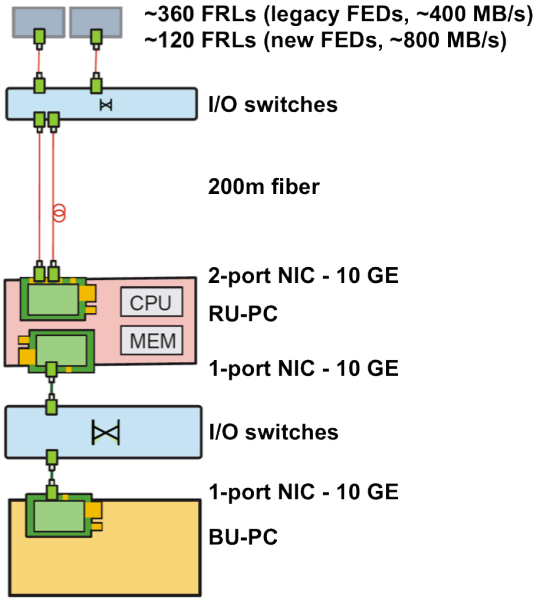


Figure 5. Conservative option for the FED Builder and RU Builder architecture where RU PC concentrates by factor two for legacy FEDs.

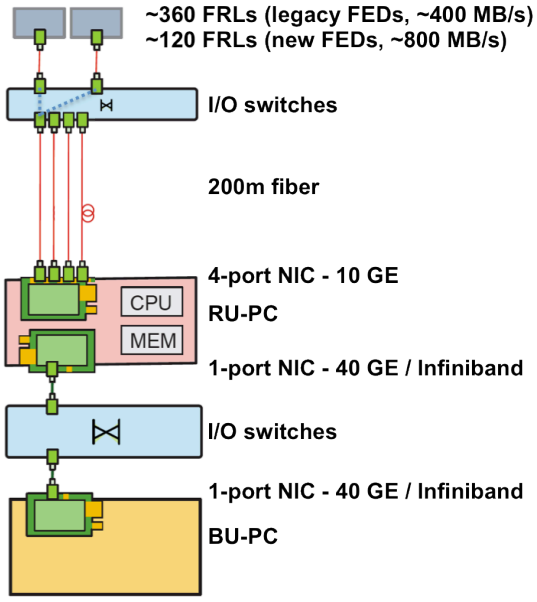


Figure 6. Advanced option for the FED Builder and RU Builder architecture where I/O switches concentrates by factor two for legacy FEDs and RU PC concentrates by factor four.

The conservative option, as outlined in **Figure 5**, concentrates two legacy FEDs in the RU-PC. The total number of RU-PCs should be about 300. The required input throughput per link in RU-PC for the two-port NIC at 10 GE is 100 kHz with a fragment size between 2 and 4 kB. In turn the required output throughput per link in each RU-PC is 100 kHz with a fragment size between 4 and 8 kB (from 400 MB/s to 800 MB/s). **Figure 6** shows the advanced option. The total number of RU-PCs will be about 75. The data concentration of FRLs is done in two steps: first I/O switches concentrate by a factor two for legacy FEDs and later RU PC concentrates by a factor four. The required input throughput per link in RU-PC for the four-port NIC at 10 GE is 200 kHz for legacy FEDs with a fragment size between 2 and 4 kB and 100 kHz for new FEDs with a fragment size between 4 and 8 kB. In turn the required output throughput per link in each RU-PC is 100 kHz with a fragment size between 16 and 32 kB (from 1.6 GB/s to 3.2 GB/s).

4. Feasibility studies on advanced networking technologies

In the last years, as the network speed has increased towards 10 gigabit per second, the CPU must spend more time working to service the network. To process network requests for the de facto networking standard, TCP/IP, the processor must dedicate a large number of cycles and resources to data transfers. To avoid this problem, technology such as Infiniband [11], iWARP [12], and RDMA over Converged Ethernet (RoCE) [13] have been developed that not only allow a very fast interconnect, but also provide a mechanism known as Remote Direct Memory Access (RDMA) [14] to

bypass the operating system and CPU in order to directly move data into application memory. The 10 GE link has been evaluated to investigate the feasibility of the new event builder based on the conservative option. For these studies the CMS event builder software has been used with TCP/IP as communication protocol. A new software relying on RDMA mechanisms based on the Direct Access Programming Library (DAPL) [15] is under development. A more detailed description and preliminary benchmarks of peer transport based on DAPL can be found in reference [16].

4.1. Cluster setup

To perform benchmark evaluation of the 10 GE link, a small cluster has been used. The setup consisted of 16 nodes of DELL PowerEdge R310 with dual sockets Intel Xeon X3323 4-core at 2.5 GHz and 4 GB of memory. The operating system running on the nodes was Scientific Linux CERN 5 (SLC5) with the 2.6.18-164.6.1.el5 kernel. Each node was equipped with a 10 GE adapter (Silicom - PE210G2SPI9-SR) and connected with 10 GE switch (Voltaire Vantage 6048).

4.2. Benchmarks

Two different benchmarks were performed to evaluate the 10 GE network technology:

- Measurement of the maximum throughput per node between one node to more nodes with multi-streams of I/O data;
- Measurement of the maximum throughput per node between N nodes to N nodes with event builder software.

In order to calculate the maximum throughput per node a XDAQ application (Multi-Stream I/O) has been implemented to send multiple streams data from one source to many destinations. As shown in **Figure 7**, throughput per node is measured sending continuously N messages to N receivers and time sampling is done at the receiver's side.

To perform the maximum throughput per node with the event builder application the CMS RU Builder [17] software has been used in emulation mode. RUs generate the event fragment data and BUs discard the event data once an event has been fully assembled. The L1 trigger is not emulated and all measurements correspond to the saturation limit. **Figure 8** shows the event builder protocol. With free capacity available, a BU requests the EVM to allocate it an event (step 1). The EVM confirms the allocation by sending the BU the event ID and trigger data of an event (step 2). These trigger data are the first super-fragment of the event. The BU now requests the RUs to send it the rest of the event's super-fragments (step 3). The BU builds the super-fragments it receives from the RUs (step 4) into a whole event within its resource table (step 5). Filter Units (FU) can ask a BU to allocate the events (step 6). A BU services a FU request by sending the FU a whole event (step 7). When a FU has finished with an event, it tells the BU to discard it (step 8).

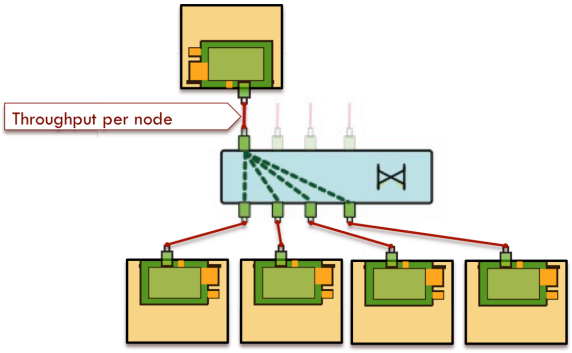


Figure 7. Diagram of the multi-stream I/O application.

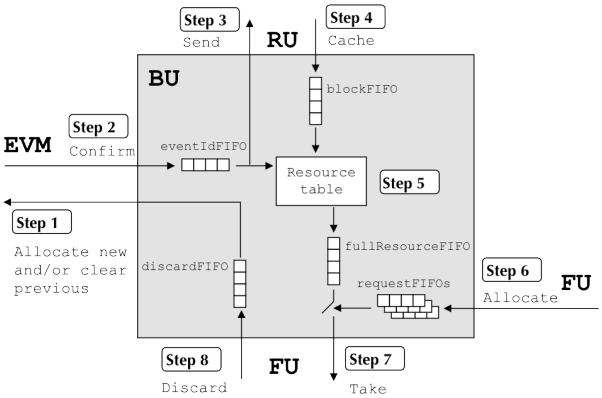
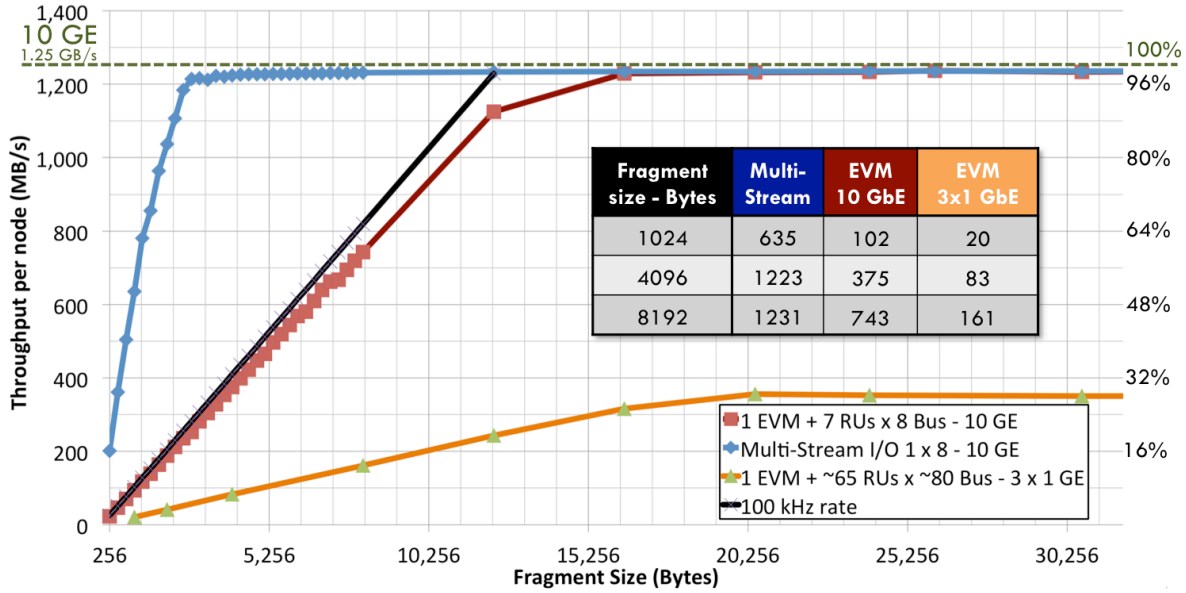


Figure 8. Event builder protocol.

A configuration with one sender and eight receivers has been tested using TCP/IP and the throughput per node as a function of fragment size is shown in **Figure 9** with the blue line (diamond points).



Fragment size - Bytes	Multi-Stream	EVM 10 GbE	EVM 3x1 GbE
1024	635	102	20
4096	1223	375	83
8192	1231	743	161

Figure 9. Throughput per node in MB/s versus fragment size in Bytes using multi-stream I/O (blue line – diamond points) and event builder applications (red line – square points for 10 GE and orange – triangle points for 3 x 1 GbE current system).

An event builder configuration with an EVM, seven RUs and eight BUs has been tested using 10 GE. The throughput per node as a function of fragment size is shown in **Figure 9** by the red line (square points). The performances of the 10 GE event builder are closed to the requirements (100 kHz rate line – cross points) of the conservative option, 400 MB/s for 4 kB fragments and 800 MB/s for 8 kB. The 10 GE link is approximately five times faster compared the current 3 x 1 GbE (orange line – triangle points). Based on these results, it is feasible to build the conservative option for event builder using 10 GE assuming the seven RUs by eight BUs scales to a 300 by 300 sized system.

5. Summary and further work

The preliminary work described in this paper has shown the possible options for the upgrade of the CMS DAQ system and the new readout link needed to accommodate sub-detectors with upgraded off-detector electronics. The results of performance tests with 10 GE are close to the requirements of the conservative option for the new event builder. New software to integrate the RDMA mechanism in on-line framework is under development. Continuation of the feasibility studies for the CMS event builder will include a larger cluster to check the scalability and test the RDMA technology using Infiniband and 40 GE. The new cluster is made up of 32 nodes of DELL PowerEdge C6220 with dual sockets Xeon E5-2670 8-core at 2.6 GHz and 32 GB of memory. Each node is equipped with a Mellanox - ConnectX-3 VPI adapter (MCX353A-FCBT) supporting 4x Fourteen Data Rate (FDR) connections with data rate of 54.4 Gbps and 40 GE. Tests of scalability, Infiniband and RoCE can be performed using the new setup.

Acknowledgement

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

References

[1] The CMS Collaboration, The Compact Muon Solenoid Technical Proposal, CERN/LHCC94-38 (1994)

[2] The LHC Study Group, The Large Hadron Collider Conceptual Design Report, CERN/AC95-05 (1995).

[3] Klabbers P. for the CMS Collaboration, Operation and performance of the CMS Level-1 Trigger during 7 TeV Collisions, Technology and Instrumentation in Particle Physics, 2011.

[4] The CMS Collaboration, The Trigger and Data Aquisition project, CERN/LHCC 2002-26, 15 December 2002.

[5] MicroTCA® specifications, PICMG, http://www.picmg.org/v2internal/specifications.cfm

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

[7] Attila Racz, Robert McLaren & Erik van der Bij, The S-LINK 64 bit extension specification: S-LINK64, EP Division, CERN

[8] D. Baden et al. "Developments for the upgrade of the CMS HCAL front-end electronics", 2010JINST, Topical Workshop on Electronics for Particle Physics 2010 (TWEPP-10), http://iopscience.iop.org/1748-0221/5/11/C11005

[9] W. Bertl, The upgrade project for the CMS pixel detector, The 20th Anniversary International Workshop on Vertex Detectors (Vertex 2011), 19th and 24th June 2011 in the village Rust at Lake Neusiedl, Austria.

[10] PCI-X specifications, PCISIG, http://www.pcisig.com/specifications/pcix_20/

[11] InfiniBand Architecture Specification, InfiniBand Trade Association, October 2004. URL http://www.infinibandta.org/specs/

[12] H. Shah et al. (October 2007). "Direct Data Placement over Reliable Transports". *RFC 5041*. http://tools.ietf.org/html/rfc5041

[13] InfiniBand Trade Association, *InfiniBand™ Architecture Specification Release 1.2.1 Annex A16: RoCE*, InfiniBand Trade Association, April 2010. http://members.infinibandta.org/kwspub/spec/Annex_RoCE_final.pdf

[14] Jeff Hilland, Paul Culley, Jim Pinkerton, and Renato Recio. RDMA Protocol Verbs Specification, April 2003. http://www.rdmaconsortium.org/home/draft-hilland-iWARP-verbs-v1.0-RDMAC.pdf

[15] uDAPL API Spec Version 2.0, http://www.datcollaborative.org/uDAPL_v20.zip

[16] G. Bauer et al., A Comprehensive Zero-copy Architecture for High Performance Distributed Data Acquisition over Advanced Network Technologies for the CMS Experiment, Proceedings of the RT2012: 18th IEEE NPSS Real Time Conference, 11-15 Jun 2012, Berkeley, CA

[17] G. Bauer et al., The CMS event builder and storage system, 17th International Conference on Computing in High Energy and Nuclear Physics (CHEP 09), DOI: 10.1088/1742-6596/219/2/022038