Available on CMS information server

CMS CR 2007/016
FERMILAB-CONF-07-364-E

CMS Conference Report

15 May 2007

CMS DAQ Event Builder Based on Gigabit Ethernet

G. Bauer[9], V. Boyer[5], J. Branson[6], A. Brett[5], E. Cano[5], A. Carboni[5], M. Ciganek[5], S. Cittolin[5], S. Erhan[5,7], D. Gigi[5], F. Glege[5], R. Gomez-Reino[5], M. Gulmini[2,5], E. Gutierrez Mlot[5], J. Gutleber[5], C. Jacobs[5], J. C. Kim[4], M. Klute[9], E. Lipeles[6], J. A. Lopez Perez[5], G. Maron[2], F. Meijers[5], E. Meschi[5], R. Moser[5,1], S. Murray[8], A. Oh[5], L. Orsini[5], C. Paus[9], A. Petrucci[2], M. Pieri[6], L. Pollet[5], A. Racz[5], H. Sakulin[5], M. Sani[6], P. Schieferdecker[5], C. Schwick[5], K. Sumorok[9], I. Suzuki[8], D. Tsirigkas[5], J. Varela[3,5]

Abstract

The CMS Data Acquisition System is designed to build and filter events originating from 476 detector data sources at a maximum trigger rate of 100 KHz. Different architectures and switch technologies have been evaluated to accomplish this purpose. Events will be built in two stages: the first stage will be a set of event builders called FED Builders. These will be based on Myrinet technology and will pre-assemble groups of about 8 data sources. The second stage will be a set of event builders called Readout Builders. These will perform the building of full events. A single Readout Builder will build events from 72 sources of 16 KB fragments at a rate of 12.5 KHz. In this paper we present the design of a Readout Builder based on TCP/IP over Gigabit Ethernet and the optimization that was required to achieve the design throughput. This optimization includes architecture of the Readout Builder, the setup of TCP/IP, and hardware selection.

Presented by Matteo Sani at *IEEE NPSS Real Time 2007*, Fermilab, Batavia IL, USA, May 4, 2007

[1] Vienna University of Technology, Vienna, Austria

[2] INFN - Laboratori Nazionali di Legnaro, Legnaro, Italy

[3] LIP, Lisbon, Portugal

[4] Kyungpook National University, Daegu, Kyungpook, South Korea

[5] CERN, Geneva, Switzerland

[6] University of California, San Diego, La Jolla, California, USA

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

[8] FNAL, Batavia, Illinois, USA

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

1 Introduction

The Compact Muon Solenoid experiment (CMS) [1] is a general purpose detector which will operate at the Large Hadron Collider (LHC) [2], situated at the CERN laboratories in Geneva. The beam crossing rate at LHC will be 40 MHz and the events sizes will be approximately 1 MB, so it will be impossible to store all the interactions. This input rate must be reduced to of order 100 Hz, the maximum rate feasible for data storage and off-line processing. CMS has chosen to reduce this rate in two steps: a hardware Level-1 trigger which has a maximum accept rate of 100 KHz and a High Level software Trigger (HLT) with an additional rejection of a factor of 10^3. All events that pass Level-1 are sent to a computer farm (Filter Farm) which performs reconstruction and selection of events using the full data.

The DAQ system is designed to be modular in order to facilitate expansion as the luminosity increases and to retain the flexibility to change the implementation of parts of the system when technologies are available or new requirements are identified. The design of the CMS Data Acquisition System and of the High Level Trigger is described in detail in the DAQ Technical Design Report (TDR) [3].

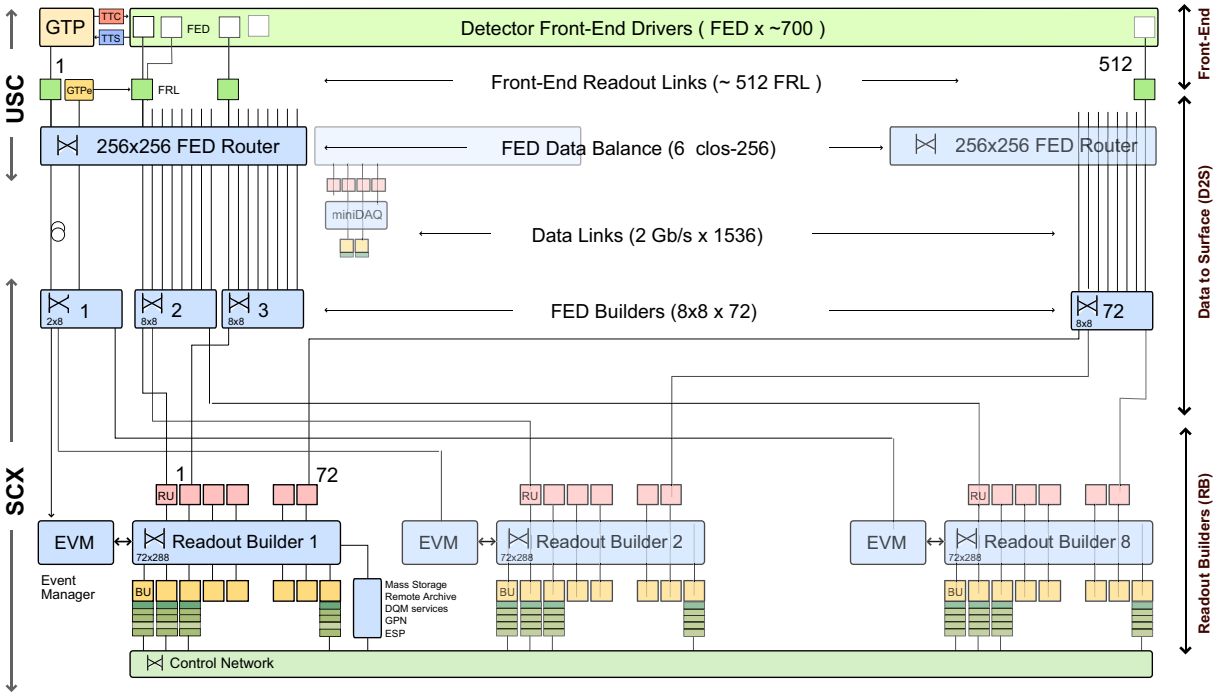

Figure 1: Simplified sketch of the CMS DAQ system.

An overview of the data flow within the CMS DAQ is shown in Fig. 1 proceeding from top to bottom. At the top are the Front-End Drivers (FEDs) which are the sub-detector specific data sources and which feed the 476 Front-End Readout Links (FRLs) which merge the data of up to two FEDs into one stream. The 2 KB outputs of the FRLs are then assembled into larger event fragments of 16 KB by the FED Builders and are distributed to up to eight independent Readout Builders. The Readout Builders consist of three types of software components running on Linux PCs. The Readout Units (RUs) receive the event fragments from the FED Builders and distribute them to the Builder Units (BUs) which assemble full events and pass them to the Filter Units (FUs). The FUs then have access to the full detector information for selecting events to send to the mass storage system.

The BUs and FUs run in a single PC and receive the data from the RUs via TCP/IP on Gigabit Ethernet (GbE). The configuration is referred to as the *Trapezoidal* configuration, because there are more BUs than RUs and a combined BU and FU PC is referred to as a BU-FU. A more detailed diagram of an individual Readout Builder is shown in Fig. 2.

With the event sizes of approximately 1 MB and each Readout Builder building events at 12.5 KHz, the data throughput of a RU PC will be approximately 200 MB/s (in and out). This paper describes measurements of the

Gigabit Ethernet network and PC throughput which demonstrate that the design requirements can be achieved.

Running the BU and FU components on the same PCs is a deviation from the original design described in the DAQ TDR [3], which had them on separate PCs connected by a second network. This modification, described in [4], removes the need for one network and one layer of PCs, but requires more TCP/IP sockets and larger switches.

2 Implementation Overview

Myrinet technology [5] is used in the FED builder and for the data transfer from the detector to the surface. The technology has a data link speed of 2 Gb/s, low latency, implements hardware flow control, and has an on-board CPU in the interface card. Furthermore, Myrinet provides a fiber optic solution for the 200 meter distance to surface at a relatively modest cost. Two parallel data paths are used to achieve 300 MB/s after taking into account efficiency in the Myrinet switches for input fragment sizes of 2 KB. Further details on the FED builder are described in [6].

The Readout Builder network, which connects the RUs to the BUs, is implemented with up to four parallel Gigabit Ethernet links, referred to as *rails*, to achieve an aggregate bandwidth of up to 490 MB/s, more than twice the design requirement.

The DAQ software is built on the XDAQ framework [7] running on commodity Linux machines.

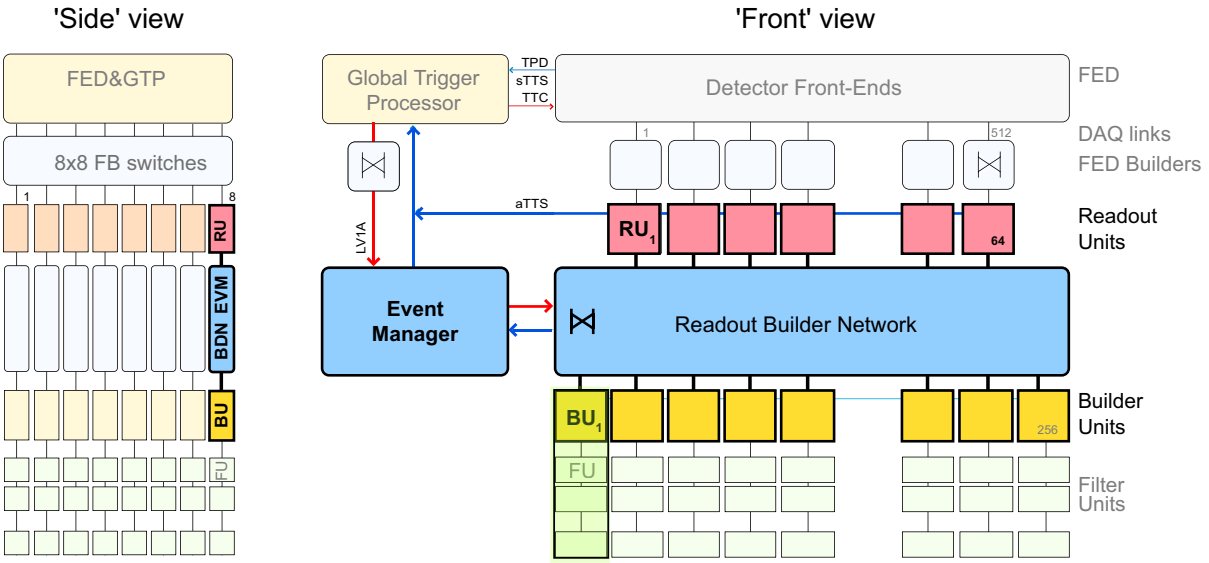

Figure 2: Simplified sketch of a Readout Builder in the Trapezoidal configuration.

3 TCP/IP setup

A transport software package called Asynchronous TCP/IP (ATCP)[1)] was developed to decouple the DAQ applications from the networking software. It also avoids blocking when more than one host is trying to send data to the same network interface of another host. ATCP puts all messages to be sent in different queues according to the destination and asynchronously processes them in another thread. It writes (reads) into (from) a given socket until it blocks and as soon as it blocks it passes to another socket and continues.

The choice of TCP/IP over Gigabit Ethernet has the advantage of using completely standard hardware and software. TCP/IP is a reliable protocol for which we do not need to worry about packet loss at the application level, which typically occurs when operating close to wire speed. The main drawback is the considerable usage of machine resources for its operation. In order to achieve good performance, the following design choices were made:

- We use Ethernet Jumbo frames. By increasing the maximum transmission unit (MTU) from the standard 1500 Bytes to 7000 Bytes we observe an increase in the performance of approximately 50%.

[1)] See *http://xdaqwiki.cern.ch/index.php/Ptatcp*.

- We implement multi-rail operation. This is done by using independent physical networks depending on the source and destination hosts. Fig. 3 shows a possible two rail configuration of the RU-BU communication.

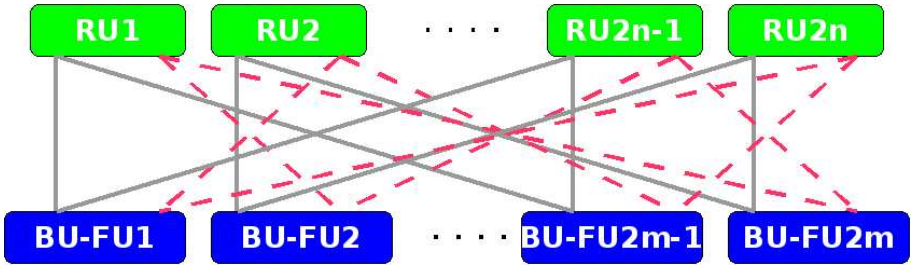

Figure 3: Example of two rails RU-BU communication. Even and odd RUs and BUs are connected with the following pattern: even to even and odd to odd network A (solid), odd to even and even to odd network B (dashed). Since every RU in a slice must write to every BU, half the data is on each network.

- Readout Builder control messages and data have different requirements in terms of latency and throughput. While the first must be delivered with low latency and have low throughput the latter require high throughput. In the configuration of TCP/IP in Linux the Nagle algorithm [8] can be turned on or off. The Nagle algorithm is a means of improving the efficiency of TCP/IP networks by reducing the number of packets that need to be sent over the network. When the Nagle algorithm is on and there are not many messages in the pipeline, messages may be delivered with a very large latency. On the other hand, when the Nagle algorithm is off, we observe an significant decrease of the throughput with time. The optimal Readout Builder performance is obtained by using different sockets for data and control messages and setting the Nagle algorithm on for data and off for control messages.

4 Tests of GbE Event Builder Architecture

The performance of the system as a whole and of individual components was measured using PCs with two single-core 2.6 GHz Xeon processors interconnected with a Force 10 E1200 switch and two Gigabit Ethernet rails for each PC.

With the prototype system of 16 RUs $\times$ 60 BU-FUs, corresponding to almost one quarter of a full-scale Readout Builder, we measure the throughput as a function of fragment size. Fig. 4 shows that for a fragment size of 8 KB network utilization is almost 100%. The fragment size of 16 KB produced by the FED Builders is sufficiently above this threshold, so even with fragments size fluctuations a high network utilization is maintained. The full line speed to the switch is achieved because the trapezoidal configuration has more reading than writing ports so the output buffers have low occupancy. For these tests, the FU software is a dummy and does not include the CPU and memory loads associated with the actual trigger algorithms.

To approach the full scale test in terms of the number of connections from RU to BU-FUs, we run multiple BU-FU applications (up to 4 per node). By doing this we obtain similar results up to a 16 RUs $\times$ 240 virtual BU-FUs configuration, where virtual refers to the fact that BU-FU PC ran multiple applications at the same time. This result shows that there is little or no decrease in performance when increasing the number of output sockets from RUs to BU-FUs, although the throughput is less than there will be in the full system, because there are only 16 RUs as inputs instead of 72.

For the design Filter Unit input of 50 MB/s, only 5-10% of the CPU is used to build events leaving the bulk of the processing power for the event selection algorithms.

5 Measurements using PCs with Two Dual-Core Processors

Several different PC architectures have been tested for the RU component. Based on our design requirements and an order tendering process, we have selected PCs equipped with two dual-core processors. In the following section, we present measurements using the selected two dual-core processor PC model. The PC is a Dell PE 2950 equipped with two Woodcrest Xeon 2 GHz processors (E5130), a 1.3 GHz front side bus, 2 PCI-X and 1 PCI-Express slots for full height expansion cards.

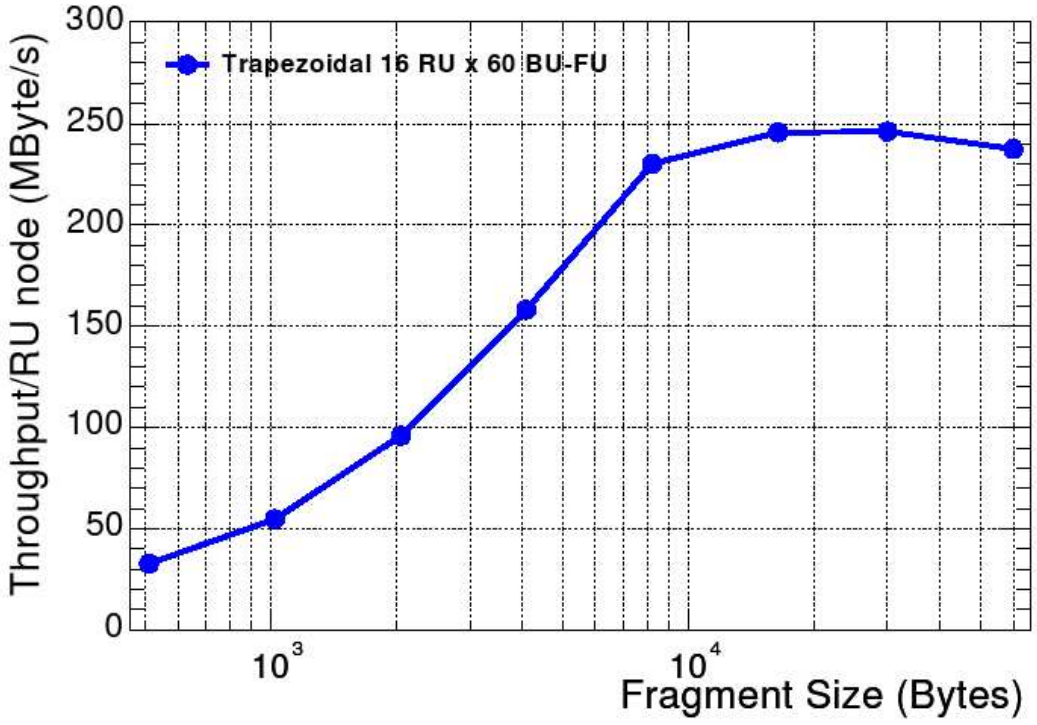

Figure 4: Throughput per node in the Trapezoidal configuration as a function of the fragment size.

5.1 Memory Throughput

In the RU configuration with Myrinet DMA input and Gigabit Ethernet TCP/IP output, performance was limited by memory throughput for earlier generation PCs. This is because the data are copied four times for each fragment: the Myrinet PCI-X DMA to memory, from user space memory to CPU then back to kernel space memory and finally data are sent by GbE PCI-E DMA transfer from memory. A more complex building algorithm in the RU might require an additional super-fragment copy in the PC thus moving the data six times.

In tests using a Dell PE 2850 with two single-cores with a 800 MHz front-side bus and non-fully buffered DIMMs, we find a memory bandwidth of 1900 MB/s and measure a throughput of 450-470 MB/s. This is consistent with being limited by the memory bandwidth which would predict 1900 / 4 $\simeq$ 475 MB/s. Adding an additional copy of the data we expect 320 MB/s and we measure 330 MB/s.

The newer Dell PE 2950 PCs described above, with fully buffered DIMMs, have a memory bandwidth above 4000 MB/s, so memory throughput is no longer a limitation.

5.2 RU and BU performance

Tests of the PCs as RUs have been carried out with one and two Myrinet inputs and six GbE rails. Measurements with two Myrinet inputs were made in order to evaluate the possibility of running two RU processes in a PC and in order to measure the fully saturated throughput of the PC. Fig. 5 shows the measured total throughput of the PC as a function of the fragment size from 2 KB to 32 KB in both configurations. The PC reaches the maximum throughput achievable with six rails and two Myrinet inputs of 640 MB/s for a fragment size of about 8 KB, not quite saturating 6 links. We find the PCs can fully saturate four Gigabit Ethernet links achieving a throughput of 490 MB/s.

5.3 Single Quad-Core Processor PC

The two dual-core processor PC was compared with a single quad-core processor PC. The purpose of this test was to explore the possibility of using a single quad-core processor and leaving a socket on the mother board free for future upgrades.

Comparing the memory throughput, we saw that the two dual-core processor PC was much faster than the single

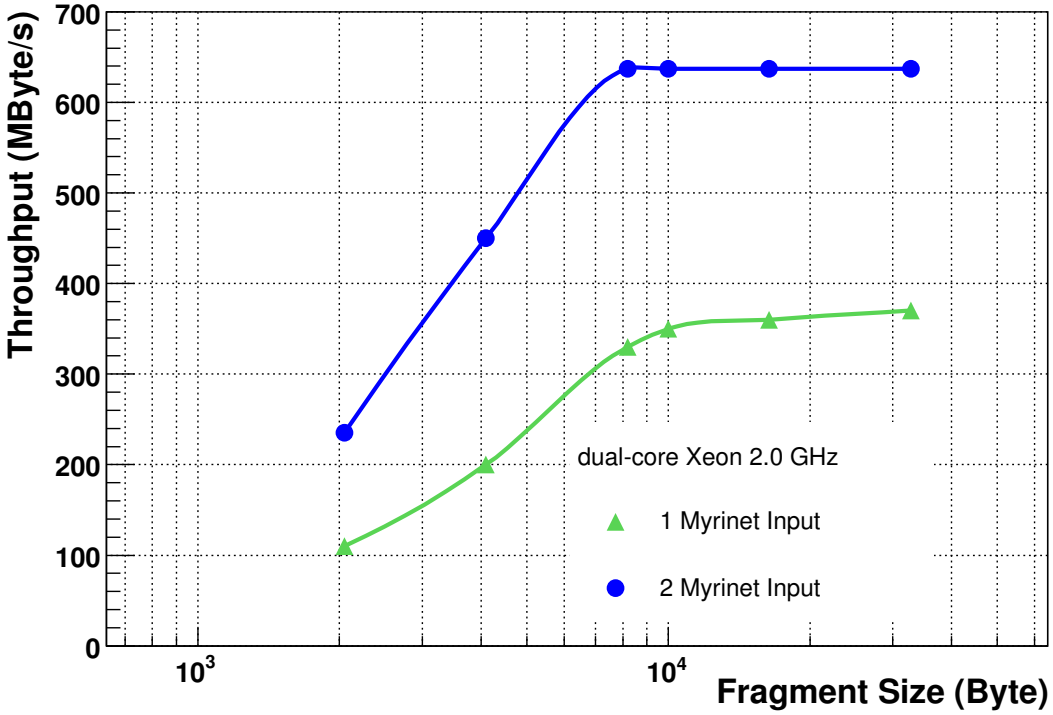

Figure 5: Total throughput measured on a PC equipped with two dual-core processor as a function of the fragment size. The results with one and two Myrinet inputs are shown.

quad-core which has only 2700 MB/s throughput. This memory bandwidth does not leave a sufficient margin for achieving the design throughput and reduces the flexibility for other potential modifications.

6 Event Builder and Filter Farm Configuration

6.1 PCs

We have now purchased 640 2u Dell PE 2950 PCs equipped with two dual-core processors that will be used in the long term as Readout Units. We are installing on each of them a Myrinet card and a 4 port Gigabit Ethernet card. The Myrinet card is a M3F2-PCIXE-2 with two fiber links and the Gigabit Ethernet card is a Silicom PEG4i with four links and based on the Intel 82571EB controller.

As it was shown in Section 5, we are in principle able to achieve a throughput of 490 MB/s per PC, although the throughput may be limited by components up and downstream of the Readout Builder.

6.2 Switches

The final choice for switches was to use Force10 E1200 switches equipped with 90 port line cards. Fully loaded, this switch accommodates 1260 Gigabit Ethernet ports. The connector density is achieved using mini-RJ21 connectors (one every six ports) which are then connected to the PCs using patch panels to standard RJ45 connectors located in the racks. The current implementation of these line cards is oversubscribed. Each group of sequential 24-24-22-20 ports can have a maximum aggregate throughput of 11 Gb/s in plus 11 Gb/s out. At the beginning of 2008 full throughput 90 port line cards are expected to be available. The chassis are compatible with both types of line cards.

7 Commissioning Configuration

All the scaling related tests described above were performed using an earlier generation of PCs and smaller configurations than the full-system. During the commissioning phase part of the 640 PCs will be used as BU-FUs. The baseline trapezoidal configuration with two slices of 72 RU and 248 BU-FU PCs should achieve a rate of at least 25 KHz. The other configurations that have been considered during the development will also be reevaluated

including the original DAQ TDR design. These configurations share common cabling and the difference is only a matter of software configuration. The choice of configuration for the first run may also depend on the High Level Trigger output requirements.

8 Future Deployment

The Event Builder and Filter Farm will be scaled to 50 KHz for the first LHC Physics run in 2008 at the nominal initial luminosity of $2 \times 10^{33} \mathrm{cm}^{-2} \mathrm{s}^{-1}$. This upgrade will involve the purchase and installation of additional filter nodes and line cards. At that time, full throughput line cards will be available. After the commissioning run, we will finalize the details of the Readout Builder architecture and decide whether to replace the oversubscribed line cards with full throughput line cards. We will then be able to use 4 slices of Readout units and have enough Filter Units to match the needs of the High Level Trigger for a L1 trigger rate of 50 KHz.

Full deployment to 8 slices and 100 KHz will be scheduled according to the requirements of the experiment. At that time, all the PCs we are currently installing will be used as Readout Units and at least 4 Force10 E1200 chassis will be needed to connect to of the order of 2000 filter nodes.

9 Summary

We have presented the design and prototyping of the second stage of the CMS Event Builder. This stage will be implemented using TCP/IP on Gigabit Ethernet. In order to achieve the design throughput of 12.5 GB/s in eight parallel slices for an aggregate bandwidth of 100 GB/s, the TCP/IP usage has been optimized and up to four parallel Gigabit Ethernet links are used per PC. The network will be interconnected using several high port density Force 10 switches.

In order to validate and optimize the design several measurements have been made. On individual PCs, we demonstrate a throughput of 490 MB/s. Using a quarter scale Readout Builder and lower performance PCs with only two links, we demonstrate a 240 MB/s throughput suggesting that the design scales well. Based on these results, we expect to be able to achieve the design throughput with the full system. This system is currently being installed with 640 PCs equipped with Myrinet and quad Gigabit Ethernet interfaces.

Acknowledgment

We acknowledge support from the DOE and NSF (USA), KRF (Korea) and the Marie Curie Program.

References

[1] The CMS Collaboration, *"The Compact Muon Solenoid Technical Proposal"*, CERN/LHCC 94-38 (1994).

[2] The LHC Study Group, *"The Large Hadron Collider Conceptual Design Report"*, CERN/AC 95-05 (1995).

[3] The CMS Collaboration, *"The Trigger and Data Acquisition project"*, CERN/LHCC 2002-26, 15 December 2002.

[4] M. Pieri et al., *"CMS DAQ Event Builder Based on Gigabit Ethernet"*, International Conference on Computing in High Energy and Nuclear Physics (CHEP 2006), Mumbai, Maharashtra, India, 13-17 Feb 2006.

[5] Myrinet products from Myricom, Inc. Arcadia, CA, USE, see *http://www.myri.com*.

[6] G. Bauer et al., *"The Terabit/s Super-Fragment Builder and Trigger Throttling System for the CMS Experiment"*, IEEE Real Time 2007, Batavia, Illinois, USA, 4 May 2007.

[7] V. Briglievic et al. *"Using XDAQ in Application Scenarios of the CMS Experiment"*, Computing in High Energy and Nuclear Physics, 24-28 March 2003, La Jolla, USA.

[8] Nagle J., *"Congestion control in IP/TCP internetworks"*, RFC 896, 1984.