Available on CMS information server

CMS CR 2006/034

CMS Conference Report

15 July 2006

R. Arcidiacono (on behalf of the CMS DAQ Group)

Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

The 2 Tbps ”Data to Surface” system of the CMS Data Acquisition.

Abstract

The Data Acquisition system of the CMS experiment, at the CERN LHC collider, is designed to build 1 MB events at a sustained rate of 100 kHz and to provide sufficient computing power to filter the events by a factor of 1000. The Data to Surface (D2S) system is the first layer of the Data Acquisition interfacing the underground subdetector readout electronics to the surface Event Builder. It collects the 100 GB/s input data from a large number of front-end cards (650), implements a first stage event building by combining multiple sources into larger-size data fragments, and transports them to the surface for the full event building. The Data to Surface system can operate at the maximum rate of 2 Tbps. This paper describes the layout, reconfigurability and production validation of the D2S system which is to be installed by December 2005.

Presented at *IEEE-NPSS Real Time*, Stockholm , June 2005

The 2 Tbps "Data to Surface" system of the CMS Data Acquisition.

R. Arcidiacono[8], V. Brigljevic[9], G. Bruno[5], E. Cano[5], S. Cittolin[5], S. Erhan[6], D. Gigi[5], F. Glege[5], R. Gomez-Reino Garrido[5], M. Gulmini[3,5], J. Gutleber[5], C. Jacobs[5], P. Kreuzer[1], G. Lo Presti[5], I. Magrans De Abril[5], N. Marinelli[2], G. Maron[3], F. Meijers[5], E. Meschi[5], S. Murray[5], A. Oh[5], L. Orsini[5], M. Pieri[7], L. Pollet[5], A. Racz[5], P. Rosinsky[5], C. Schwick[5], P. Sphicas[1,5], K. Sumorok[8], J. Varela [4,5]

[1]University of Athens, Athens, Greece

[2]Institute of Accelerating Systems and Applications, Athens, Greece

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

[4]LIP, Lisbon, Portugal

[5]CERN, Geneva, Switzerland

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

[7]University of California, San Diego, San Diego, California, USA

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

[9]Rudjer Boskovic Institute, Zagreb, Croatia

***Abstract*— The Data Acquisition system of the CMS experiment, at the CERN LHC collider, is designed to build 1 MB events at a sustained rate of 100 kHz and to provide sufficient computing power to filter the events by a factor of 1000. The Data to Surface (D2S) system is the first layer of the Data Acquisition interfacing the underground subdetector readout electronics to the surface Event Builder. It collects the 100 GB/s input data from a large number of front-end cards (650), implements a first stage event building by combining multiple sources into larger-size data fragments, and transports them to the surface for the full event building. The Data to Surface system can operate at the maximum rate of 2 Tbps. This paper describes the layout, reconfigurability and production validation of the D2S system which is to be installed by December 2005.**

I. Introduction

The CMS Trigger and Data Acquisition system is designed to collect and inspect the detector information at the LHC bunch crossing frequency of 40 MHz. It has also to select a maximum of 100 Hz events for offline analysis.

At the design luminosity of $10^{34}\ cm^{-2}s^{-1}$, the LHC rate of proton collisions will be around 20 per bunch crossing, producing approximately 1 MB of data in the CMS detector. The first level trigger is designed to reduce the incoming data rate to a maximum of 100 kHz, by processing fast trigger information coming from the calorimeters and the muon chambers, and selecting events with interesting signatures.

The Data Acquisition system (DAQ) [1] has therefore to sustain a maximum input rate of 100 kHz, for an average data flow of 100 GB/s coming from ∼650 data sources, and to provide enough computing power to reduce the rate of stored events by a factor of 1000. A block diagram of the CMS DAQ system is shown in Fig. 1.

The data flow is as follows. The various subdetector readout systems store data continuously in 40 MHz pipelined buffers. Upon arrival of a synchronous L1 trigger (3μs latency), the corresponding data are extracted from the front-end buffers and pushed into the DAQ system by the Front-End Drivers (FEDs). The event fragments are sent to the first stage of the event building, the FED Builder network, through a short distance link and a receiver interface card. This stage is in charge of transporting the fragments to the surface, and of assembling event fragments of variable size from 650 FEDs into 64 super-fragments of 16 kB average size. The super-fragments are then stored in large buffers in Readout Units (RU), waiting for the second stage of event building (RU Builder), implemented with multiple 64 x 64 networks. There will be up to 8 RU Builders, or "DAQ slices", connected to the FED Builders layer. Each FED Builder is in charge of distributing the super-fragments, on an event by event basis, to the existing RU Builders. Super-fragments corresponding to the same event are located in the same slice, and are read by one Builder Unit (BU) of the RU Builder network. The complete event is then transferred to a single unit of the Filter Farm. The High Level Trigger (HLT) algorithms will decide whether to store or reject the event.

This architecture has an innovative double-stage event building, implemented with two layers of switches, which optimizes the traffic load to the final event builder and allows for a progressive deployment of the full size system.

II. The Data to Surface system

The Data to Surface (D2S) system represents the first layer of the DAQ interfacing to the subdetector front-end readouts. It receives and transports the data to the surface. It also performs super-fragment building in the FED Builder networks. In order to reduce the number of inputs to the FED Builders and to better exploit the available bandwidth, it allows to merge up to four data sources into a single data fragment.

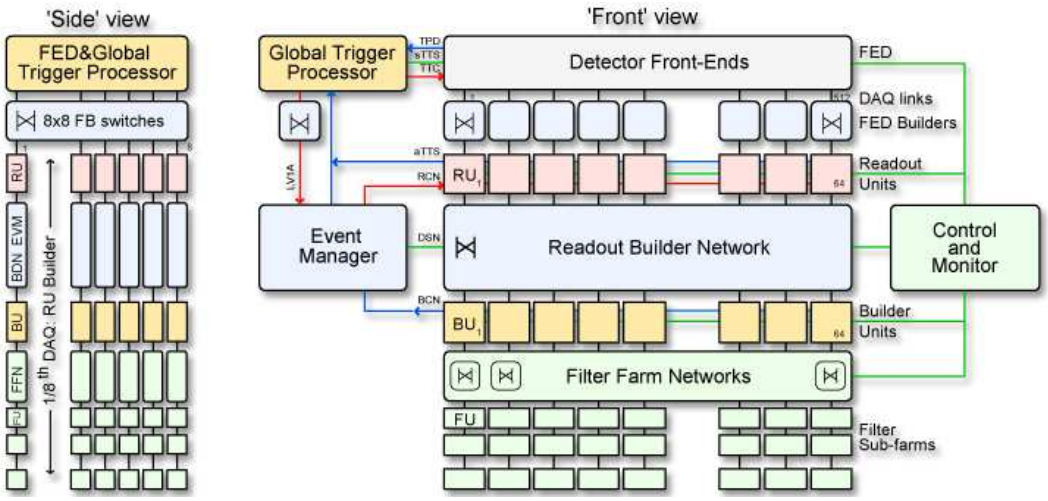

Fig. 1. Schematic view of the CMS Data Acquisition system.

The requirements which have driven the design implementation of this layer are:

- the need of a common interface to the subdetector specific front-end readout electronics, in order to reduce diversity and ease integration and maintenance operations
- a sustained data throughput of each D2S element of 200 MB/s over a distance of 200 m, with the capability of absorbing fluctuations of the event size
- the capability of providing balanced traffic to the event building stages, in order to optimize the building efficiency, and the capability of easy reconfiguration of the data sources routing.

Fig. 2 shows the physical implementation of one element of the D2S system. The architecture foresees a maximum of 512 of such elements. One element includes the following devices:

- one S-Link64 Sender card, housed by the subdetector Front-End Driver
- a short distance link implemented as LVDS copper cable, maximum 10 m long
- one Front-end Readout Link (FRL), housing up to four S-Link64 receivers
- two commercial Myrinet Network Interface Cards (NIC) [2], with two 2.0 Gbps data rate optical links (rails), used both in input and output of the FED Builder network
- eight network switches' ports (Myrinet technology)
- two 200 m long optical fibers making up the long distance link connecting underground to surface area, plus four short dual optical fibers to connect the two rails of the Myrinet NIC to the switch ports.

Fig. 3 displays a picture of a D2S element, from the S-Link64 card to the optical fibers which connect the Myrinet NIC to the FED Builder switch.

A. The Sender card

The S-Link64 Sender card is a Common Mezzanine Card (CMC) developed by the DAQ group. It receives data from the FED via an S-Link64 port [3]. The card is able to buffer up to 1.6 kB of data before generating back-pressure to the FED. The CMC has an LVDS converter to interface with the short distance link. The implementation of the S-Link64 port on the FEDs is part of the common functional block diagram followed by the subdetectors.

The default data transfer rate over the LVDS cable is 400 MB/s (50 MHz clock $\times$ 64 bits), double the average design speed of the DAQ system. The S-Link64 card has the possibility of being set at two other clock frequencies, 40

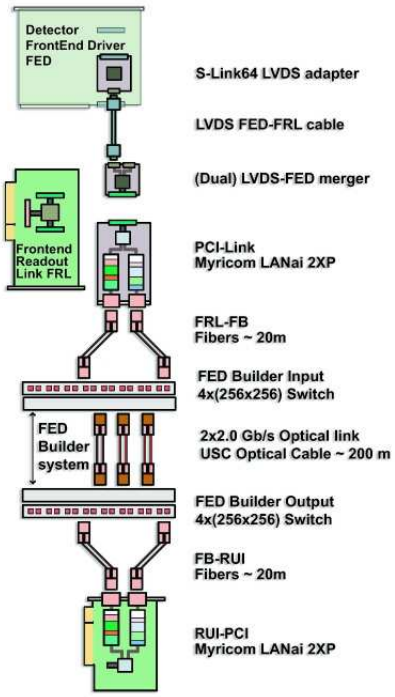

Fig. 2. D2S element physical implementation.

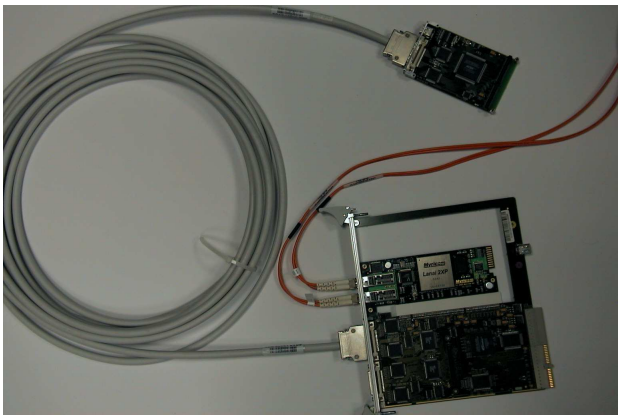

Fig. 3. Picture of a D2S element, from the S-Link64 sender card (up-right) to the long distance link (optical fibers).

MHz or 60 MHz, in case of less or more demanding front-end systems respectively. The maximum sustainable output rate is therefore 480 MB/s.

B. The FRL card

The FRL card is a CompactPCI card, developed by the DAQ group, which receives, buffers and optionally merges the event fragments, checks the payload CRC, and pushes the data to the Myrinet NIC in fixed size packets. The input buffer memories have a 64 kB size.

The FRL card has three different interfaces: an input interface which handles up to four LVDS cables; an output interface to the FED Builder implemented as a 64bit/66MHz PCI connector for the Myrinet NIC; a configuration and control interface which is a PCI bus interface connected to the CompactPCI backplane.

The FRL card also provides monitoring features, like the ability to spy on a fraction of events, via the CompactPCI bus, and to fill histograms of fragment sizes distribution. Each of the 512 FRL cards is expected to receive 2 kB event fragments on average. The FRL cards push the data to the Myrinet NICs via the internal 64bit/66MHz PCI bus. Measurements show that a speed very close to the PCI bus limit of 528 MB/s can be reached, depending on the FRL packet size and for event sizes above 1 kB.

C. The FED Builder

The FED Builder is based on Myrinet technology. Myrinet is a high performance interconnect technology for clusters, composed of crossbar switches and NICs, connected by point to point bi-directional links. It employs wormhole routing, and the delivery of packets is guaranteed by a flow control at the link level. The Myrinet switch supports hence back-pressure, which is propagated backwards up to the FEDs.

The design choice of having 8x8 logical networks in the FED Builders has been taken. One FED Builder is made of 8 input NICs, two layers of two 8x8 switch networks, 8 output NICs (see Fig. 4, where only one network layer is shown). Each rail of one Myrinet card is connected to an independent crossbar. The FED Builder input cards are programmed to read fragments from the FRL and to send them to the switch, via one of the optical links, with a destination port assigned on the basis of the fragment event number and a predefined lookup table. The FED Builder output cards, inputs of the RU Builders, concatenate fragments with the same event number from all the connected input cards, building the super-fragments.

The two switches layers are located one in the underground cavern, one in the surface DAQ room. They are implemented with large switches, 256 external ports each, allowing for a high reconfigurability of the FED-Builders' network by simple reprogramming of the switches' routing table.

The Myrinet card can transfer at 4.0 Gbps data rate over the two optical rails. As for the Myrinet switch performance, measurements done with different traffic conditions show that the 95% of the efficiency is reached with constant size events, while it drops to 50% when transporting variable size events.

Fig. 5 shows the FED Builder performance as measured in a test bench with an 8x8 network. Data are injected in the system by the FRL cards generating fragments in saturation

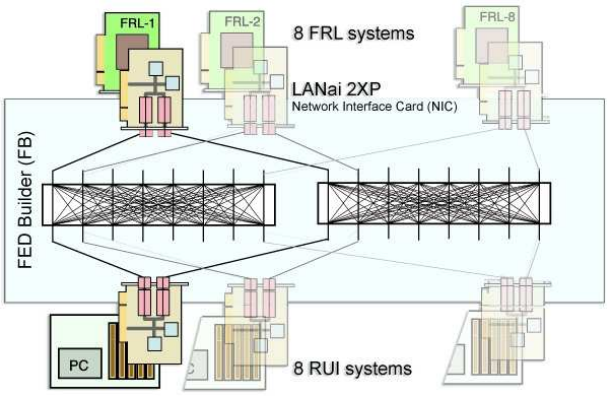

Fig. 4. 8x8 FED Builder with a two-rail network.

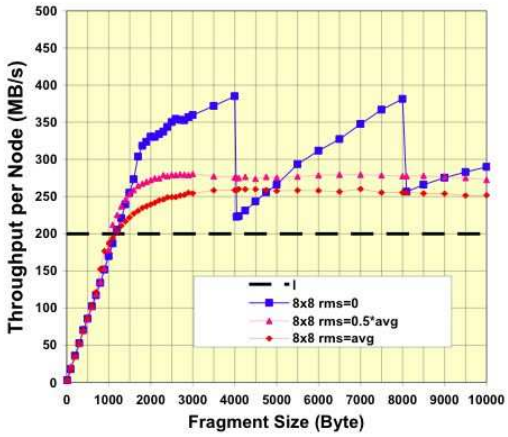

Fig. 5. Throughput per node versus fragment size in a 8x8 FED Builder.

mode. The packet size used in the FRL-NIC protocol is 4 kB. The throughput variation as a function of the average fragment size depends on the rms of the event size distribution. The best performance is obtained for constant event sizes. For the nominal average fragment size of 2 kB, the measured throughput per node is about 250 MB/s.

The sustained data transfer speed through the FED builders is 1 Tbps, given the design implementation choice of no traffic shaping through the Myrinet switches and the variable and unbalanced traffic conditions generated by the front-end data sources. This throughput meets the CMS DAQ requirements. A maximum peak bandwidth of 2 Tbps is theoretically possible, when fully exploiting the FRL and Myrinet NIC bandwidth, and with traffic shaping in the FED Builders.

III. D2S System Production and Validation

The D2S system is the first part of the DAQ to be installed in the experimental area. The delivery milestone is December 05.

Full poduction of the custom made components (S-Link64 and FRL cards) has been completed, and all the commercial equipment (Myrinet NICs and switches, LVDS cables, optical fibers) has been purchased and delivered.

All this material needs to be tested, classified in a database and validated before installation. Several testing procedures have been developed at CERN for this purpose. Tests have already started and will continue throughout the summer.

In the following subsections, the validation and quality control implemented for the custom made cards will be described.

A. FRL and S-Link64 Sender validation

The FRL and S-Link64 Sender cards are qualified using two different setups, both custom-built:

- Stage 1 - the first setup is where visual and electrical checks are performed and the card firmware is loaded.
- Stage 2 - the second runs several software tools in order to perform a full hardware integrity and operational check.

Stage 1 test bench is composed of a dedicated hardware station where the card under test can be plugged, inspected and measured by ad-hoc connectors. FPGA's firmware is programmed via the JTAG serial port, accessed through the PC parallel port. Operations and interface to the production database are controlled by a LabVIEW program.

The Stage 2 test bench is composed of a CompactPCI crate, controlled by a PC, reference LVDS cables, and a set of reference FRL cards (or S-Link64 cards, depending whether S-Link64 or FRL cards respectively are to be tested), which are used coupled with custom-built testing cards. Fig. 6 shows a diagram of this configuration.

The custom-built testing card is a GIII [4] PCI card, connected to the internal 64bit/66MHz PCI bus of the FRL card, and hosting a S-Link64 Sender card. This GIII card is able to generate pseudo-random or well-know pattern data. It emulates the FED data source. The data are then transmitted, via the S-Link64 Sender and the LVDS cable, to the input of the FRL. The FRL processes and pushes the data back to the GIII card, which emulates the Myrinet NIC-FRL protocol, and, at last, checks the integrity of the payload byte by byte.

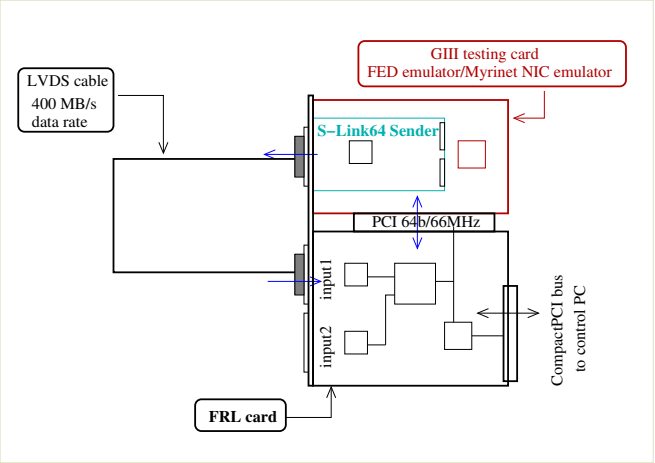

Fig. 6. Diagram of the FRL/S-Link64 test board used in the Stage2 setup.

The software controlling this setup performs first JTAG chain checks (via the CompactPCI bus), memory/register read/write checks, LVDS link autotests and eventually runs the closed-loop data transfer initiated by the GIII cards.

A linux-based Perl/Tk GUI controls the test procedure and the database interface.

B. Long duration stability tests

A third setup is used to qualify part of the D2S element (from S-Link64 card to Myrinet NIC), as a whole. It is equivalent to a Bit Error Rate Test (BERT) system implemented with D2S elements.

The test system is composed of one PC hosting 16 FED emulators, each with an S-Link64 Sender card, 16 LVDS cables and one CompactPCI crate hosting 16 FRL cards coupled this time with Myrinet NICs. The LVDS cables connect FED emulators to the FRLs. Fragments data from the FED emulators, along with the CRC, are transmitted to the FRL inputs, via the LVDS link, and pushed into the Myrinet NICs. The CRC is recomputed and compared to the original inside the FRL. Data corruption during the transfer is detectable by checking the CRC comparison.

The test runs for 24 hours. A total of 34.5 TB of pseudo-random data, at the speed of 400 MB/s, are injected in each link, leading to a BER upper limit of 1.3×10^{-14} at 90% C.L. per cable.

IV. Conclusion

The architecture of the D2S system is dictated by the requirements and constraints of the DAQ and Trigger systems.

The D2S has been designed to collect and concentrate data from all the readout sources, transport them on the surface, and feed balanced output streams to the event building process.

This is accomplished by providing a first stage in the event building which groups 8 FRL data sources in a single output

channel, and distributes it to different slices of the Event Builder. This scheme allows for a scalable design of the surface Event Builder, which can be procured and installed in phase with the requirements arising from the performance of the accelerator and the experiment itself.

The D2S components' development phase has been completed successfully. The FED builder based on Myrinet technology shows a throughput per node above the required 200 MB/s, for fragment sizes greater than 1 kB. The aggregate throughput of the D2S system, in standard running conditions, is 1 Tbps.

All the hardware components of the complete D2S system, custom made or commercial, have been produced or purchased. Validation procedures have been specifically designed for all the components, to qualify the system. The full production of FRL and S-Link64 Sender cards is currently being tested at CERN, while the LVDS cables, along with Myrinet NICs will undergo long duration tests during the summer. The procurement and validation of the D2S system is on schedule and the installation is foreseen for December 05.

REFERENCES

[1] CERN/LHCC 2002-26, CMS TDR, *Data Acquisition and High Level Trigger*

[2] Myricom, see http://www.myricom.com

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

[4] E. Cano, D. Gigi, "FEDKIT User's Manual and Programmer's Manual", v. 0.8d, CERN CMS internal note (2004). http://cano.home.cern.ch/cano/fedkit/fedkit-0.8d.pdf