

Studies of future readout links for the CMS experiment

Gerry Bauer⁷, Barbara Beccati², Ulf Behrens¹, Kurt Biery⁶, Olivier Bouffet², James Branson⁵, Sebastian Bukowiec², Eric Cano², Harry Cheung⁶, Marek Ciganek², Sergio Cittolin^{3,a}, Jose Antonio Coarasa², Christian Deldicque², Aymeric Dupont², Samim Erhan⁴, Dominique Gigi², Frank Glege², Robert Gomez-Reino², Derek Hatton¹, Andre Holzner⁵, Yi Ling Hwang², Lorenzo Masetti², Frans Meijers², Emilio Meschi², Remigius K Mommsen⁶, Vivian O'Dell⁶, Luciano Orsini², Christoph Paus⁷, Andrea Petrucci², Marco Pieri⁵, Attila Racz², Olivier Raginel⁷, Hannes Sakulin², Matteo Sani⁵, Philipp Schieferdecker^{2,b}, Christoph Schwick², Dennis Shpakov⁶, Michal Simon², Konstanty Sumorok⁷

¹ DESY, Hamburg, Germany

² CERN, Geneva, Switzerland

³ Eidgenössische Technische Hochschule, Zurich, Switzerland

⁴ University of California, Los Angeles, Los Angeles, California, USA

⁵ University of California, San Diego, San Diego, California, USA

⁶ FNAL, Chicago, Illinois, USA

⁷ Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

^a Also at University of California, San Diego, San Diego, California, USA

^b Now at University of Karlsruhe, Karlsruhe, Germany

E-mail: Sebastian.Bukowiec@cern.ch

Abstract. The Compact Muon Solenoid (CMS) experiment has developed an electrical implementation of the S-LINK64 extension (Simple Link Interface 64 bit) operating at 400 MB/s in order to read out the detector. This paper studies a possible replacement of the existing S-LINK64 implementation by an optical link, based on 10 Gigabit Ethernet in order to fulfil larger throughput, replace aging hardware and simplify an architecture. A prototype transmitter unit has been developed based on the FPGA Altera PCI Express Development Kit with a custom firmware. A standard PC has been acted as receiving unit. The data transfer has been implemented on a stack of protocols: RDP over IP over Ethernet. This allows receiving the data by standard hardware components like PCs or network switches and NICs. The first test proved that basic exchange of the packets between transmitter and receiving unit works. The paper summarizes the status of these studies.

1. Introduction

The existing Compact Muon Solenoid experiment (CMS) [1, 2] data acquisition (DAQ) system reads out data fragments from approximately 700 sub-detector specific Front End Drivers (FEDs), builds events from these data fragments using commercial network switch technologies, and provides these to the High Level Trigger (HLT) computing farm where they are analyzed to generate a final trigger decision.

The components of the central DAQ system have been installed, commissioned and operated successfully during the first year of data acquisition at the Large Hadron Collider (LHC) with both protons and lead ions collisions.

There are several reasons to improve an already properly functioning system. First, to provide higher performance in order to accommodate larger event sizes at higher LHC luminosities, resulting in larger throughput requirements. Second, to replace an aging hardware in order to maintain a high reliability of the system and to move away from vendor specific networking technologies. After several years (5-10 years) components used in the modules cannot be purchased anymore. Third, elimination of custom protocols, when possible.

The subject of the studies presented in this article is the new readout link implementation that provides a connection between each FED and the CMS event builder. The new readout link is a possible replacement of the existing Simple Link Interface 64 bit (S-LINK64) [3] implementation.

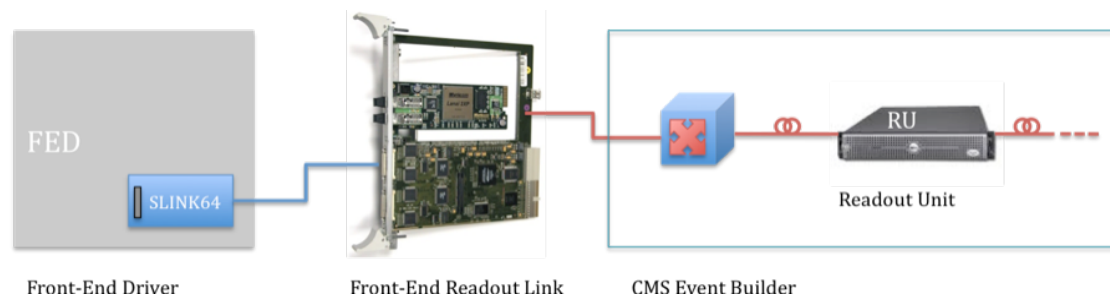


Figure 1. Current readout link implementation diagram.

2. Current readout link

As shown in Fig. 1 the current design consists of a custom developed Common Mezzanine Card (CMC) SLINK64 directly plugged into the sub-detector FED, a Low-Voltage Differential Signaling (LVDS) cable, and a Front-End Readout Link (FRL) [4] (Fig. 1, 3, 4). The interface that is used to connect the front-end electronics to the next layer of read-out is called S-LINK64. The S-LINK64 is a specification introduced by the CMS experiment that defines a protocol, electrical interface and form factor for a point-to-point, high-speed 64-bit data link. It does not describe the physical link itself. The physical link used in the current implementation is the LVDS cable containing several twisted-pair of copper wires over which data are sent. The existing CMS readout link can operate at 400 MB/s (3.2 Gb/s) when clocked at 50 MHz on the distance up to 11.5 meters. The FRL is a custom Compact PCI card that is used to pass data fragments from high-speed copper cable to the optical fiber network (Myrinet [5]). The FRL has the possibility to merge data from two links, check data integrity and also generate a fixed size data packet to avoid memory fragmentation at the receiving PC. It also provides extensive monitoring features.



Figure 2. Front-End Readout Link board.



Figure 3. Rack with Front-End Readout Link boards.

3. Possible future architectures

The readout link connects custom made components with commercial technologies and provides interfaces to all sub-detector systems. The foreseen upgrade for some sub-detectors will imply hardware modifications in their FEDs. They plan to replace the current FEDs with μ TCA based FEDs [6]. Therefore the link must support both legacy FEDs and newly designed FEDs that will be able to transfer substantially higher data volumes to the DAQ.

Fig. 4 shows two possible implementations of the new readout link. The interface between the FED and the link transmitter unit is not yet defined. The board can be developed as a custom made CMC card as in the present implementation or it can be developed as a μ TCA board itself. In the latter case FEDs would send the data to the readout link board over the back plane using one of the data transfer protocols supported by the μ TCA standard.

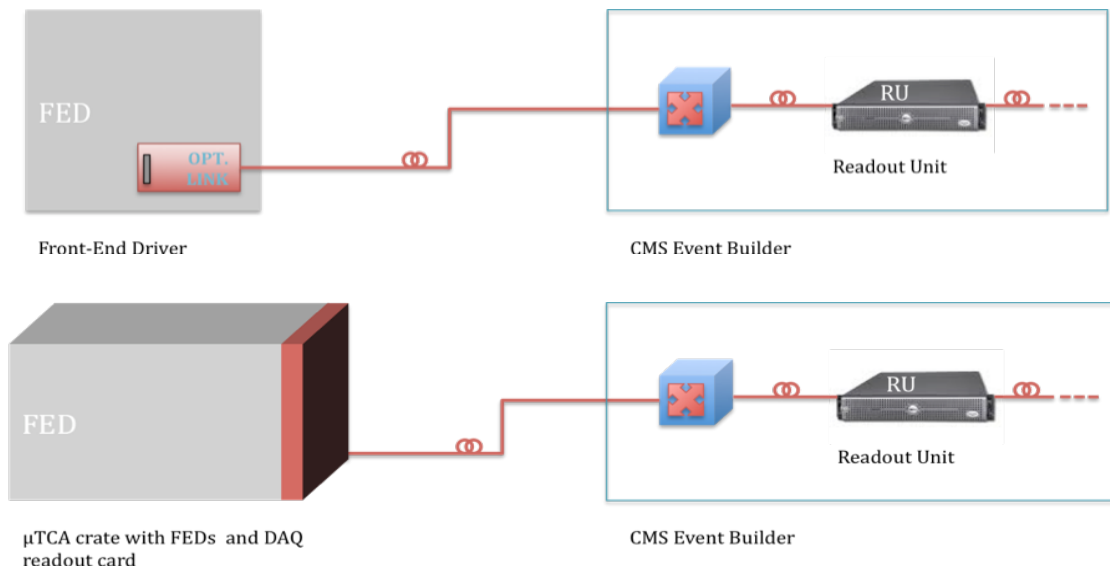


Figure 4. Diagram of the two possible implementations of the new readout link.

In the new design an optical signal can be sent directly from the link transmitter unit located underground to the surface. Thus, the current conversion of the electrical to optical signal underground

due to the long cable length to the surface is no longer needed. This allows simplifying the architecture by eliminating the FRL. In addition, the readout link board will provide monitoring functionalities and the possibility of injecting simulated data similar to those presently implemented in the FRLs.

The considered protocol for communication between transmitter and receiving units is Reliable Datagram Protocol (RDP). RDP is suitable for the application because it is lightweight; datagram oriented and provides reliable message service with sequence number and retransmission. Unlike TCP, RDP never splits the packets and it provides a sequence number related to the packet and not to the byte.

4. Development and studies of the upgrade link based on 10 Gigabit Ethernet

These studies encompass both the link transmitter unit and the link receiving unit. A link transmitter unit has been developed based on the FPGA Altera PCI Express Development Kit [7]. A standard PC acts as a link receiving unit. The software is developed in the C programming language on the Scientific Linux CERN operating system. The data transfer is implemented on a stack of protocols: RDP over IP over Ethernet.

4.1. Readout Link Test Setup

The test setup consists of three elements: PC, evaluation board and 1Gb/s switch. The evaluation board is plugged into the PC and sends data via optical fiber to the switch. The switch passes the data through Gigabit Ethernet (copper link) back to the server where the receiving software is installed. Data is passed through the switch to check if packets are built correctly and to ensure the switch does not drop them.

The correct packet format was checked using a network protocol analyzer Wireshark [8] and also looking at the switch statistics.

4.2. Evaluation board

The Altera PCI Express Development Kit, Stratix II GX Edition (Fig. 5) is used for these studies. For test purposes data is generated inside the FPGA. As shown in Fig. 6, the FPGA module uses two memories: the Data Memory stores blocks of data to be sent; the Control Memory stores auxiliary data. Upon acknowledgment, the data block is removed from the Data Memory. The development kit provides 6.5 Gb/s network interface but to be compatible with Ethernet standard for tests only 1 Gb/s is used.

The core is developed by PH-CMD group at CERN in order to provide high performance. Existing Soft IP doesn't meet the project requirements.

4.3. Implementation of the RDP protocol processing in the receiving PC

The RDP is designed to provide a reliable data transport service for packet-based applications. The protocol is intended to be simple to implement but still efficient in respect of physical bandwidth usage. It aims to provide a solution where User Datagram Protocol (UDP) does not meet the needs of reliability and Transmission Control Protocol (TCP) is too complex and causes too much overhead. It is specified as one of the encapsulated protocols in the Internet Protocol (IP), protocol number 27. A more detailed description of the RDP can be found in references [9, 10].

The receiver application is running in user space and is based on raw sockets. Sockets are the interface between the user processes and the networking system of the OS kernel. Raw sockets are provided by the Linux kernel and are available only to users with root-user authority. Sockets exist in communication domains. A socket domain or a protocol family is an abstraction that provides an addressing structure and a set of protocols. The PF_PACKET protocol family [11] was chosen as a communication domain because it supports SOCK_RAW type and it allows the application to manipulate the networking packets at lower levels of the OSI model [12] protocol stack. The user application can therefore efficiently implement custom protocols directly on top of the Ethernet data link layer, bypassing the kernel networking facilities. The receiver application directly receives and

transmits the Ethernet frames on the network interface. It formats, wraps and un-wraps the IP packets, the RDP packets and the data payload. As the studies are in the preliminary stage, the application currently provides the basic acknowledgment feature. The link transmitter unit sends a packet and waits for the acknowledgment before sending the next one. In case the acknowledgment does not arrive in a given timeout, the packet is sent again. Theoretically the smallest acknowledge packet could consist of the Ethernet header (14 bytes), IP header (20 bytes) and RDP header (18 bytes) for a total of 52 bytes. However, the size of the acknowledge packet must be of at least 64 bytes in order to prevent the switch from dropping the packets.



Figure 5. Board Diagram of the PCI Express Stratix II GX Development Board.

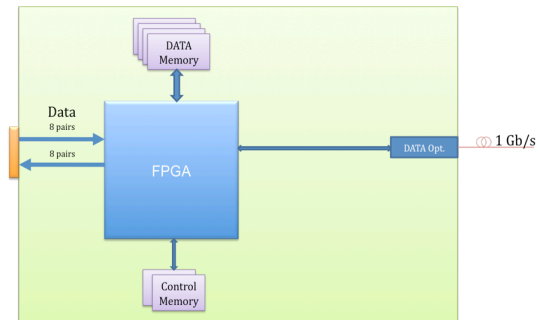


Figure 6. Evaluation board diagram of essential components (PCIe-link) to study RDP protocol.

5. Conclusion

The first version of the firmware of the evaluation board was created. It provides the core functionalities of sending and acknowledging the packets. A preliminary version of the software for the link receiving unit implementing the RDP protocol was developed and was proven to be able to communicate with the evaluation board. Transfer through a standard switch was successful and no packet was dropped.

Currently point-to-point architecture is used but the possibility of multi-destination data transfer will be investigated in the future.

Acknowledgment

This research project has been supported by a Marie Curie Early Initial Training Network Fellowship of the European Community's Seventh Framework Programme under contract number (PITN-GA-2008-211801-ACEOLE) and in part by the DOE and NSF (USA).

References

- [1] The CMS Collaboration, CMS Technical Proposal, CERN LHCC 94-38, 1994.
- [2] The CMS Collaboration (R. Adolphi et al.), "The CMS Experiment at CERN LHC", JINST 3 S08004 p. 361, 2008.
- [3] Racz A, McLaren R, Van der Bij E, The SLink64 bit extension specification: S-Link64, available at <http://hsi.web.cern.ch/HSI/s-link/>.
- [4] Gigi D, FRL (FED Read-out Link), <http://cms-frl.home.cern.ch/cms-frl/>.
- [5] Myricom, Myrinet Overview, <http://www.myri.com/myrinet/overview/>.
- [6] PCI Industrial Computer Manufacturers Group (PICMG), Open Modular Computing Specifications, <http://www.picmg.org/>.
- [7] PCI Express Development Kit, Stratix II GX Edition, http://www.altera.com/products/devkits/altera/kit-pciexpress_s2gx.html.
- [8] Wireshark, <http://www.wireshark.org/>

- [9] Velten D, Hinden R, Sax J 1984, Reliable Data Protocol, RFC-908 <http://tools.ietf.org/html/rfc908>.
- [10] Partridge C, Hinden R 1990, Version 2 of the Reliable Data Protocol (RDP), RFC-1151 <http://tools.ietf.org/html/rfc1151>.
- [11] Linux MAN Pages, <http://www.kernel.org/doc/man-pages/online/pages/man2/socket.2.html>.
- [12] Stevens W R, Fenner B, Rudoff A M, 2003 Addison-Wesley Professional, UNIX Network Programming: The sockets networking API, Chapter 1, Section 1.7, p. 18-19