Available on CMS information server

CMS CR -2013/416

The Compact Muon Solenoid Experiment

Conference Report

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

14 November 2013 (v2, 19 December 2013)

10 Gbps TCP/IP streams from the FPGA for the CMS DAQ Eventbuilder Network

Gerry Bauer[6], Ulf Behrens[1], James Branson[4], Olivier Chaze[2], Sergio Cittolin[4], Jose Antonio Coarasa[2], Georgiana-Lavinia Darlea[6], Christian Deldicque[2], Marc Dobson[2], Aymeric Dupont[2], Samim Erhan[3], Dominique Gigi[2], Frank Glege[2], Guillelmo Gomez-Ceballos[6], Robert Gomez-Reino[2], Christian Hartl[2], Jeroen Hegeman[2], Andre Holzner[4], Lorenzo Masetti[2], Frans Meijers[2], Emilio Meschi[2], Remigius K. Mommsen[5], Srecko Morovic[2,a], Carlos Nunez-Barranco-Fernandez[2], Vivian O'Dell[5], Luciano Orsini[2], Wojciech Ozga[2], Christoph Paus[6], Andrea Petrucci[2], Marco Pieri[4], Attila Racz[2], Olivier Raginel[6], Hannes Sakulin[2], Matteo Sani[4], Christoph Schwick[2], Andrei Cristian Spataru[2], Benjanin Stieger[2], Konstanty Sumorok[6], Jan Veverka[6], Christopher Colin Wakefiled[2], Petr Zejdl[2]

Abstract

For the upgrade of the DAQ of the CMS experiment in 2013/2014 an interface between the custom detector Front End Drivers (FEDs) and the new DAQ eventbuilder network has to be designed. For a loss-less data collection from more then 600 FEDs a new FPGA based card implementing the TCP/IP protocol suite over 10Gbps Ethernet has been developed. We present the hardware challenges and protocol modifications made to TCP in order to simplify its FPGA implementation together with a set of performance measurements which were carried out with the current prototype.

Presented at *TWEPP-13 Topical Workshop on Electronics for Particle 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 Institute Rudjer Boskovic, Zagreb, Croatia

Preprint typeset in JINST style - HYPER VERSION

10 Gbps TCP/IP streams from the FPGA for the CMS DAQ Eventbuilder Network

G. Bauer[f], T. Bawej[b], U. Behrens[a], J. Branson[d], O. Chaze[b], S. Cittolin[d], J. A. Coarasa[b], G.-L. Darlea[f], C. Deldicque[b], M. Dobson[b], A. Dupont[b], S. Erhan[c], D. Gigi[b], F. Glege[b], G. Gomez-Ceballos[f], R. Gomez-Reino[b], C. Hartl[b], J. Hegeman[b], A. Holzner[d], L. Masetti[b], F. Meijers[b], E. Meschi[b], R. K. Mommsen[e], S. Morovic[b,g], C. Nunez-Barranco-Fernandez[b], V. O'Dell[e], L. Orsini[b], W. Ozga[b], C. Paus[f], A. Petrucci[b], M. Pieri[d], A. Racz[b], O. Raginel[f], H. Sakulin[b], M. Sani[d], C. Schwick[b], A. C. Spataru[b], B. Stieger[b], K. Sumorok[f], J. Veverka[f], C. C. Wakefield[b] and P. Zejdl[b*]

[a] *DESY,*
Hamburg, Germany
[b] *CERN*
Geneva, Switzerland
[c] *University of California, Los Angeles,*
Los Angeles, California, USA
[d] *University of California, San Diego,*
San Diego, California, USA
[e] *FNAL,*
Chicago, Illinois, USA
[f] *Massachusetts Institute of Technology,*
Cambridge, Massachusetts, USA
[g] *Also at Institute Rudjer Boskovic,*
Zagreb, Croatia
E-mail: petr.zejdl@cern.ch

ABSTRACT: For the upgrade of the DAQ of the CMS experiment in 2013/2014 an interface between the custom detector Front End Drivers (FEDs) and the new DAQ eventbuilder network has to be designed. For a loss-less data collection from more then 600 FEDs a new FPGA based card implementing the TCP/IP protocol suite over 10Gbps Ethernet has been developed. We present the hardware challenges and protocol modifications made to TCP in order to simplify its FPGA implementation together with a set of performance measurements which were carried out with the current prototype.

KEYWORDS: Data acquisition concepts; Data acquisition circuits; Digital electronic circuits.

[*] Corresponding author.

Contents

1. Introduction

The central data acquisition system (DAQ) [1] of the Compact Muon Solenoid (CMS) experiment at CERN collects data from more than 600 custom detector Front End Drivers (FEDs). In the current implementation data are transferred from the FEDs via 3.2 Gbps electrical links (Slink64) to custom interface boards called Frontend Read-out Links (FRL), which transfer the data to a commercial Myrinet network based on 2.5 Gbps optical links.

During 2013 and 2014 the CMS DAQ system will undergo a major upgrade to address the obsolescence of current hardware and the requirements posed by the upgrade of the LHC accelerator and various detector components. Particularly, the DAQ Myrinet and 1 Gbps Ethernet networks will be replaced by 10/40 Gbps Ethernet and Infiniband. In addition, new μTCA based Front End Drivers (FEDs) will be in operation after the Long Shutdown 1 (LS1). These new FEDs will be read out via new point-to-point optical links (SlinkXpress) replacing the previous Slink64 links.

In order to accommodate both FED link types (legacy Slink64 links and new μTCA links) a new Frontend Read-out Link (FRL) hardware has been designed. The new hardware called FEROL provides a new link based on 10 Gbps Ethernet (10GE) and replaces the previously used Myrinet network adapters. The FED data are transmitted over this link using the simplified and unidirectional version of the TCP/IP network protocol. The new 10GE links provide a reliable connection between the front-end readout system in the underground and the eventbuilder network at the surface (Figure 1). More details on the new DAQ system can be found in [2].

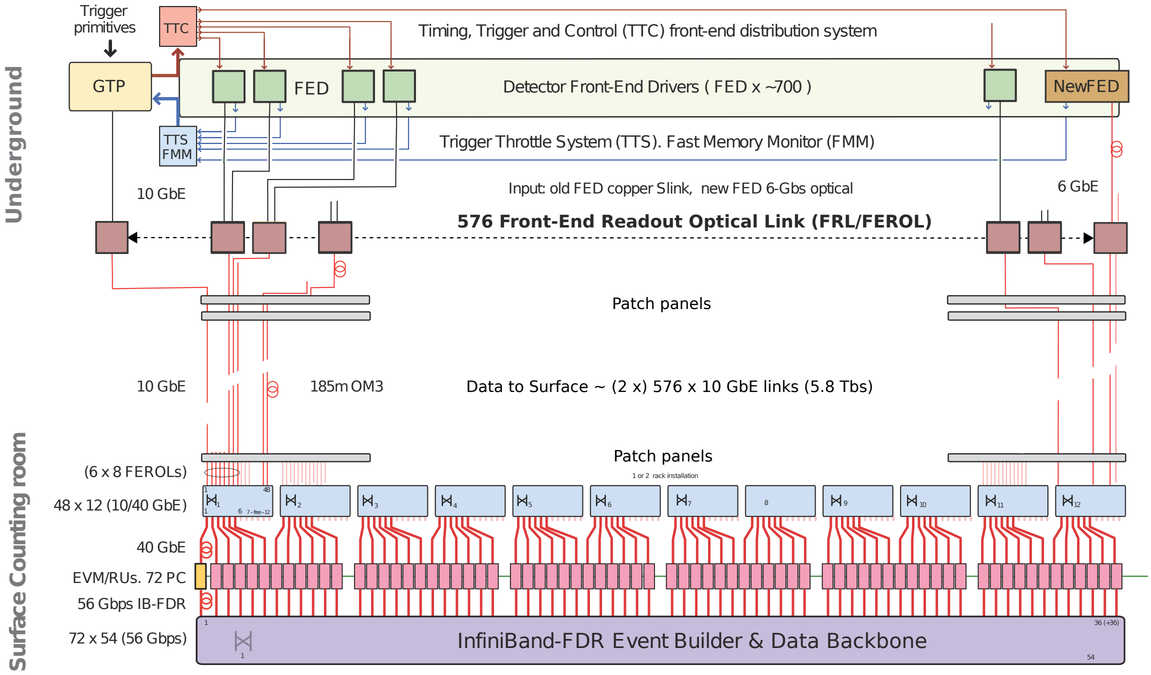
Trigger primitives
TTC
GTP
Timing, Trigger and Control (TTC) front-end distribution system
FED
Detector Front-End Drivers (FED x ~700)
NewFED
TTS FMM
Trigger Throttle System (TTS). Fast Memory Monitor (FMM)
Underground
10 GbE
Input: old FED copper Slink, new FED 6-Gbs optical
6 GbE
576 Front-End Readout Optical Link (FRL/FEROL)
Patch panels
10 GbE
185m OM3
Data to Surface ~ (2 x) 576 x 10 GbE links (5.8 Tbs)
Patch panels
1 or 2 rack installation
(6 x 8 FEROLs)
48 x 12 (10/40 GbE)
40 GbE
EVM/RUs. 72 PC
56 Gbps IB-FDR
72 x 54 (56 Gbps)
InfiniBand-FDR Event Builder & Data Backbone
Surface Counting room

Figure 1. A schematic diagram of the new DAQ readout system. The data flow from the top to the bottom. The FRL/FEROL hardware shown in the middle provides the interface between the custom links from the FEDs and commercial network used in the event building process.

2. Requirements

The following overall requirements are identified for the new Frontend Read-out Link hardware (FRL):

- Support two legacy Slink64 FED interfaces, with 3.2 Gbps throughput per interface.
- Support the new μTCA based FEDs with SlinkXpress interface up to 10 Gbps bandwidth.
- Ability to assert back-pressure to the FEDs when upstream congestion is detected.
- Provide reliable (loss-less) connection between FRLs (underground) and eventbuilder equipment (surface).
- Possibility of sending several TCP data streams over one 10GE link.
- Throughput close to the 10 Gbps bandwidth.

3. Front-end Readout Link hardware (FRL)

The FRL is a Compact PCI card providing an interface to the custom links from the FEDs. The FRL consists of two boards connected through a PCI-X interface. The base board provides two Slink64 interfaces to legacy FEDs. The upper board hosts a PCI-X card providing a network interface to the next level of DAQ system which are Readout Unit (RU) PCs receiving the FED data.

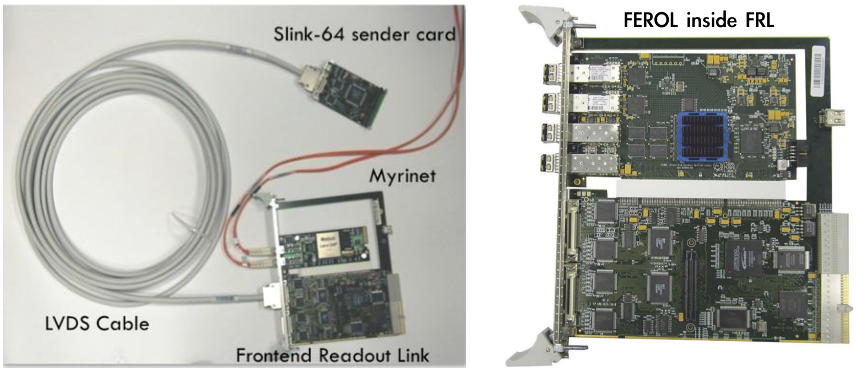

Figure 2. Compact PCI FRL card (left) with the embedded Myrinet PCI-X network card attached to the FED's Slink64 via a SLINK cable, and the new FEROL board inside the FRL (right).

Figure 2 shows the FRL with one Slink64 sender card connected and with a commercial Myrinet network interface card, which was used up to 2013 in the CMS DAQ system. For the new DAQ2 the Myrinet card is replaced with a new custom board called FEROL (Front-End Readout Optical Link).

The following interfaces to the FED will be used in the new DAQ:

Slink64 Interface The Slink64 interface implements an electrical LVDS link with 64 bits data width with maximum throughput 3.2 Gbps (400 MBytes). The Slink64 specification defines a set of connectors for sending and receiving data, along with mechanical constraints for the sender and the receiver cards. In addition, the specification also defines a FIFO based protocol for writing into the sender card and reading from the receiver card.

SlinkXpress Interface The new FEDs in CMS are being implemented as μTCA boards. Due to the small dimensions, the μTCA form factor does not allow a Slink64 sender card to be plugged. In addition, the future FEDs are expected to sent more than 3.2 Gbps to the DAQ.

SlinkXpress provides a reliable point-to-point optical connection for the new FEDs. Data are transmitted in 4kB maximum size packets using currently 8b/10b coding and packet framing and checksum similar to GbE. The transmitter keeps a short buffer of transmitted packets (currently 4) awaiting acknowledgement. Reliability is ensured by retransmission after timeout.

An SlinkExpress FPGA core is available for both Altera and Xilinx FPGAs. The current implementation allows to run at the link speeds up to 6.3 Gbps or at fixed 10 Gbps. The SlinkXpress core provides the same interface, protocol and data format used for the legacy Slink64. Therefore, the usage of the core in future FED hardware will be simple for the CMS subsystems.

4. Front-end Readout Optical Link (FEROL)

4.1 High-speed Interfaces

The FEROL supports two 10 Gbps and two 6 Gbps interfaces via four SFP+ cages. One 10 Gbps interface implements an Ethernet network interface for the TCP/IP connections to the new DAQ eventbuilder network. The other interfaces implement the SlinkXpress protocol. Either one 10 Gbps

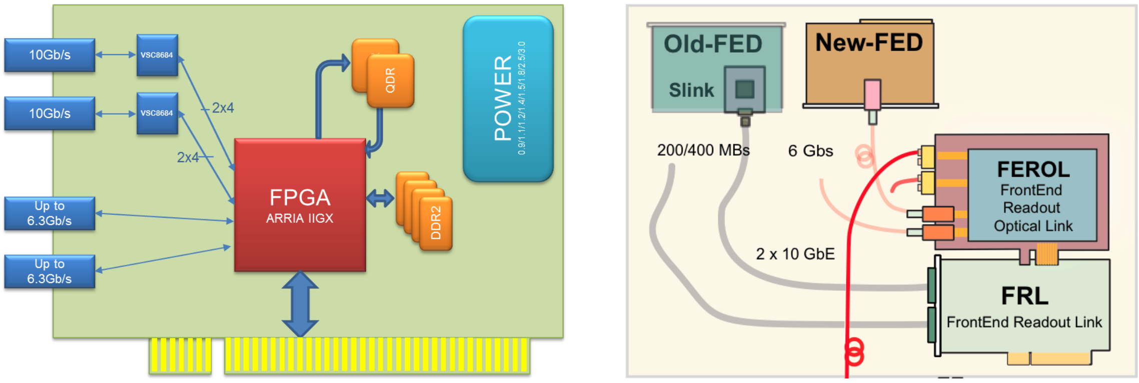

Figure 3. Block diagram (left) showing the FEROL hardware and interfaces, and a schematic diagram (right) showing the connection to the FEDs and DAQ Ethernet network.

SlinkXpress or up to two 6 Gbps SlinkXpress interfaces can be used to read out data from upgraded FEDs. The legacy Slink64 FEDs are read by the FRL and provided to the FEROL through the PCI-X interface. Figure 3 shows the FEROL block diagram and connections to the FED and DAQ network.

4.2 TCP/IP over Ethernet

TCP/IP provides a reliable and in-order data delivery featuring flow and congestion control. The Flow control controls the occupancy of the buffers in a receiving PC. In case of receiver's buffer overflow, the TCP will automatically decrease the throughput. The congestion control limits the rate of data sent to the network below a rate that would create a network congestion. As a result the congestion control allows transmitting multiple TCP streams on the same link and allows to merge several links, which are not fully utilized, together using a switch.

These features match the requirements specified in section 2. The link merging is also an important feature which greatly reduces the amount of network ports and equipment required for the DAQ network, therefore reducing cost. We expect to merge between 8 and 16 Ethernet 10 Gbps links from FEROLs into one 40 Gbps port.

Therefore, TCP/IP over the Ethernet network was chosen as the new readout link. In order to achieve the maximum data throughput, TCP has been implemented in the FPGA as opposed to running in an embedded CPU inside the FPGA. To limit the TCP implementation complexity and to minimize FPGA resource utilization, we designed a simplified and unidirectional version of the protocol.

Several simplifications [3] were possible because the data flow only in one direction from the FEROL to the receiving PC and because the DAQ network topology is fixed and designed with sufficient bandwidth to avoid packet congestion.

The main simplifications include:

1. Only the client part of TCP/IP is implemented. The FEROL (client) opens a connection to a receiving PC (server), sends the data and keeps the connection open until it is aborted. In addition, the server cannot send any user data back to the client. Only the acknowledgment packets are sent back, but these are part of the TCP protocol.

2. The complex TCP congestion control was simplified to an exponential back-off. This decreases the throughput when a temporary congestion is detected in order to avoid a congestion collapse. A fast-retransmit algorithm is also implemented to improve the throughput in case of single packet losses. In this case, the data are retransmitted immediately without waiting for a timeout.

The following standard TCP features are implemented:

- Jumbo frame support – up to 9000 bytes in one frame sent over an Ethernet network.
- Nagle's algorithm – merges data fragments to fully utilize the jumbo frame.
- Window scaling – supports TCP window sizes greater than 64KB.
- Silly window avoidance – stops sending data when the receiver is busy and requests a small amount of data.

The following TCP features are not implemented: timestamps, selective acknowledgments, out of band data.

4.3 TCP/IP Firmware Structure

Figure 4 shows the FEROL TCP/IP firmware structure. A packet arriving from a network enters the FPGA through a SERDES block. It is checked for validity (CRC, MAC and IP address match) and is decoded. If an action is required, an appropriate command is sent to the command FIFO. Simple commands like ARP or PING requests are processed by the protocol builder block, where a packet header containing a response is prepared and is sent to the packet header FIFO. The header is read

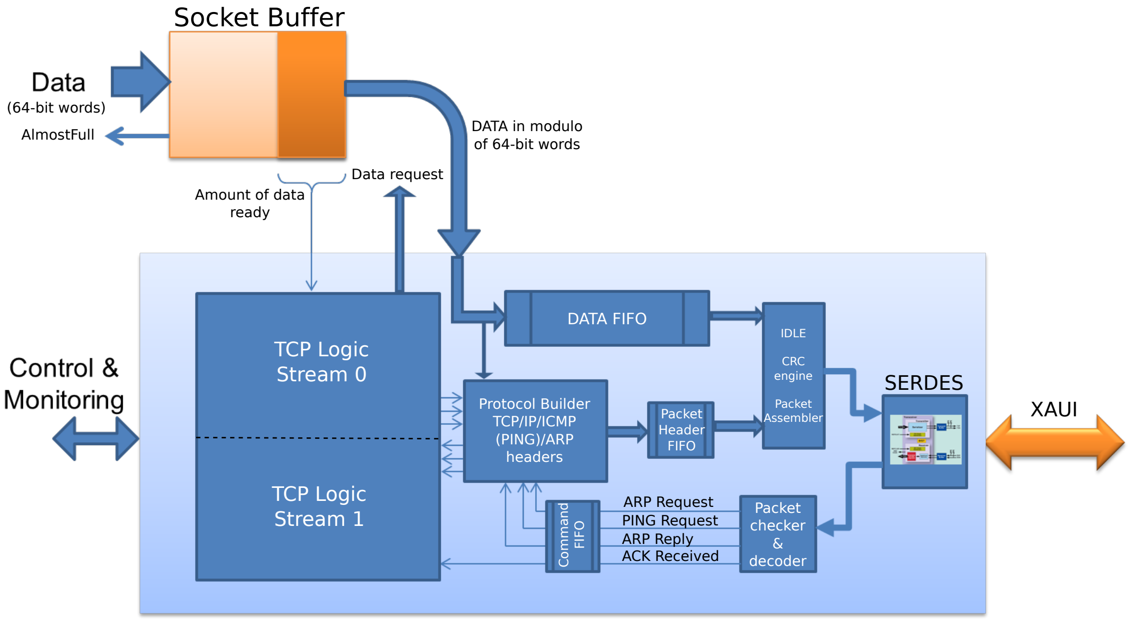

Figure 4. Block diagram showing the FEROL TCP/IP firmware structure.

by the packet assembler block, where the full Ethernet frame is prepared and the CRC is calculated. Finally, the packet containing the response is sent to the SERDES block and to the network.

The firmware contains two TCP cores implementing the simplified and unidirectional version of the protocol. Each block is able to open one TCP stream to the destination PC. The data coming from the detector are stored in a socket buffer. When a TCP packet has to be sent, the data are read from the socket buffer and temporarily stored in the DATA FIFO and a TCP header is prepared in the protocol builder block. The packet assembler block then merges the header with the data and sends the full Ethernet frame to the SERDES.

4.4 Resource Utilization

The core of the FEROL board is an Altera Arria II GX 125 FPGA. The board features two 6 Gbps interfaces, two 10 Gbps interfaces and a PCI-X interface. The 6 Gbps interfaces have direct connections to the FPGA. The 10 Gbps interfaces are connected through Vitesse transceivers which translate each 10 Gbps link into four links running at 3.125 Gbps (XAUI interface) directly connected to the FPGA. This solution allowed a cheaper FPGA (Altera Arria II GX) without 10 Gbps ports to be used. The board also contains 512 MBytes of DDR2 memory implementing the TCP socket buffer and 16 Mbytes of QDR memory as the secondary multipurpose memory.

Figure 5 shows the FPGA's resource utilization. The total design uses 54% of the FPGA. The logic implementing the TCP/IP protocol and Ethernet interface (red and black parts) together utilize 20% of the FPGA. The TCP core driving one stream (black part) utilizes 3% of the FPGA resources (∼3000 out of 99280 ALUTs / ∼1700 out of 49640 ALMs).

IP cores implementing TCP/IP at 10 Gbps wire speed are available commercially. The *Quick-*

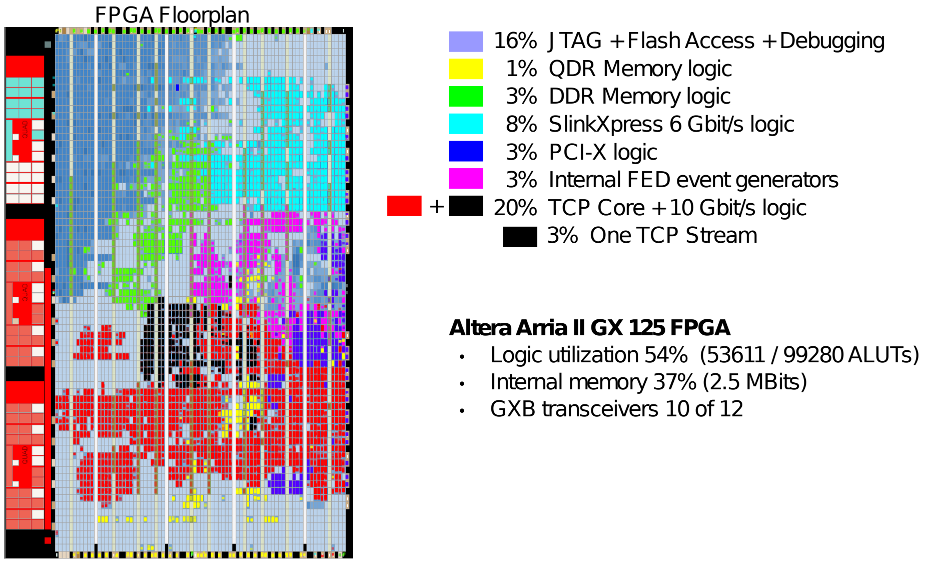
FPGA Floorplan
16% JTAG + Flash Access + Debugging
1% QDR Memory logic
3% DDR Memory logic
8% SlinkXpress 6 Gbit/s logic
3% PCI-X logic
3% Internal FED event generators
+ 20% TCP Core + 10 Gbit/s logic
3% One TCP Stream
Altera Arria II GX 125 FPGA
• Logic utilization 54% (53611 / 99280 ALUTs)
• Internal memory 37% (2.5 MBits)
• GXB transceivers 10 of 12

Figure 5. FPGA Resource Utilization

TCP core from PLDA [1] requires an Altera high-end FPGA STRATIX IV GX or STRATIX V GX. The *10G TCP Offload Engine* core from Intilop [2] can be used as a standalone IP core and consumes 30000 Altera ALMs. Both cores implement the full TCP/IP stack in hardware as opposed to the simplified and unidirectional TCP/IP presented here, but require high end FPGAs.

5. Measurements

The following measurements were performed with the FEROL hardware: Point-to-point measurement and link merging measurement.

Figure 6 compares the point-to-point throughput between a FEROL and a PC to a measurement between the two PCs using the Linux TCP/IP stack. The hardware implementation handles the small fragment sizes more efficiently due to smaller overhead in packet handling. The maximal throughput is 9.70 Gbps.

Figure 7 shows the throughput when 16x 10GE links are merged into one 40GE link using a 40GE Mellanox SX1024 switch. Up to 16 FEROLs send data using two TCP streams each. Starting from 10 streams (5 FEROLs) the transmitted data fully saturate one 40 GE link and a network congestion occurs. The congestion control limits the speed of the individual streams. Therefore, the streams equally share the 40GE bandwidth.

The PCs used in these tests are DELL PowerEdge C6100 with Intel Xeon X5650 CPU at 2.67 GHz equipped with Myricom 10GE network cards (point-to-point test) and DELL PowerEdge R620 with Intel Xeon E5-2670 CPU at 2.60 GHz equipped with Mellanox 40GE network card (link merging test). Running TCP streams at the maximum link speeds requires some performance

[1] http://www.plda.com

[2] http://www.intilop.com/

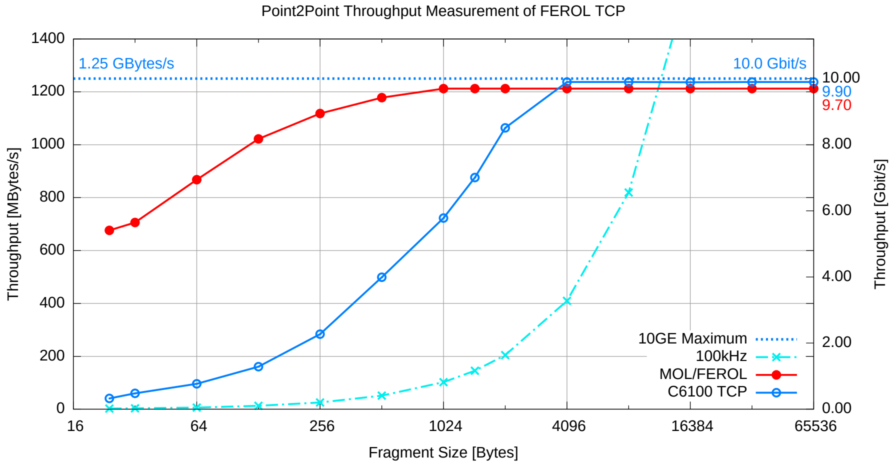

Figure 6. Linux TCP sockets compared to the FEROL hardware TCP implementation as a function of fragment size. The dash dotted line indicates the throughput corresponding to a 100 kHz fragment rate at the given fragment size.

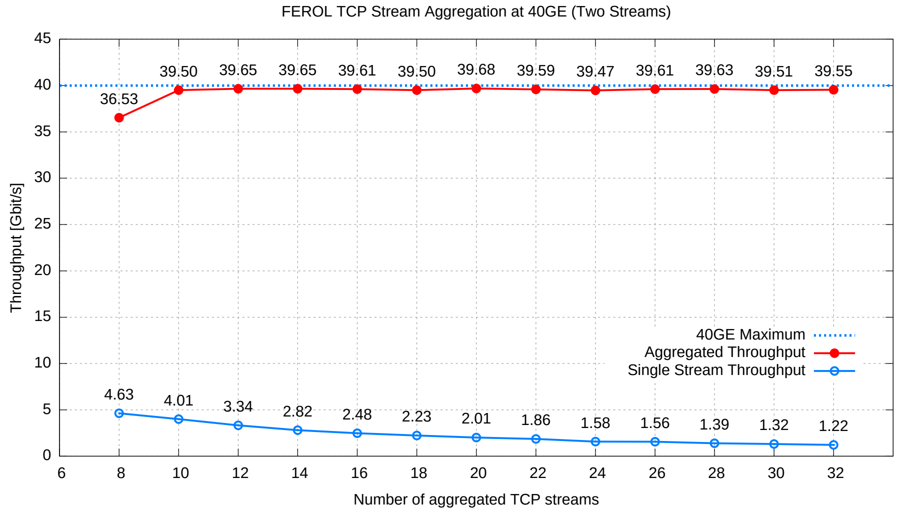

Figure 7. Up to 32 streams from 16 FEROLs are merged into one 40GE link. The red plot shows the aggregated throughput of the streams, the blue plot shows the throughput of one stream. The throughputs are plotted as a function of the number of streams.

tuning. In particular, the TCP socket buffer settings in the operating system have to be increased [4] and interrupt balancing (moving interrupt handlers between the available CPU cores) has to be disabled and manually fine-tuned. The maximum performance is also sensitive to the CPU hyper-threading settings found in the PC's BIOS.

6. Summary

We implemented a simplified and unidirectional version of TCP in the FPGA. The presented hardware can send up to two TCP streams with a maximum throughput of 9.7 Gbps. The congestion control and link merging were tested with 16 hardware senders and achieved a maximum useful throughput of 39.68 Gbps over a 40GE link without any significant amount of retransmissions.

The presented hardware is a key part of the new CMS DAQ system which will operate after the ongoing shutdown of LHC. About 600 hardware TCP senders will provide reliable 10 Gbps data connections between the front-end readout system in the underground and the eventbuilder network at the surface.

To our knowledge this is the first implementation of a simplified and unidirectional version of the TCP/IP protocol in an FPGA with moderate resources running at 10 Gbps in high energy physics experiments.

Acknowledgments

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

References

[1] Mommsen R et al., *The data-acquisition system of the CMS experiment at the LHC*, *J.Phys.Conf.Ser.* **331** (2011) 022021.

[2] Holzner A et al., *The new CMS DAQ system for LHC operation after 2014 (DAQ2)*, submitted to *J.Phys.Conf.Ser.*.

[3] Zejdl P et al., *10 Gbps TCP/IP streams from the FPGA for High Energy Physics*, submitted to *J.Phys.Conf.Ser.*.

[4] NASA, *TCP Performance Tuning on End Systems*. Available: `http://www.nren.nasa.gov/tcp_tuning.html`