TWEPP 2013 - Topical Workshop on Electronics for Particle Physics

Contribution ID: **87**

Type: **Oral**

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

Thursday 26 September 2013 15:15 (25 minutes)

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 the TCP in order to simplify its FPGA implementation together with a set of firmware and hardware tests and performance measurements which were carried out with the current prototype. The measurements include tests of TCP stream aggregation and congestion control.

Summary

The CMS data acquisition (DAQ) collects data from more than 600 custom detector Front End Drivers (FEDs). In the current implementation data is transferred from the FEDs via 3.2 Gbps electrical links (SLINK) to custom interface boards, 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 face the new challenges expected after 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.

The interfaces to the FED readout links will be implemented with a custom board (FEROL) based on an Altera FPGA. The board supports two 10 Gbps and two 6 Gbps interfaces via four SFP+ cages.
One 10 Gbps interface implements Ethernet for connection to the new DAQ eventbuilder-network. Three interfaces are used to read out data from upgraded FEDs via a basic point-to-point protocol.

For a reliable data transmission into the eventbuilder network we chose to implement the TCP/IP protocol suite on top of 10Gbps Ethernet interface in the FPGA. TCP/IP is a well known, reliable and standard protocol suite already implemented in the all mainstream operating systems. TCP contains congestion control which allows us to efficiently merge several low bandwidth TCP streams to one faster interface in Ethernet switch. The stream merging greatly reduces the amount of the network equipment required for the new DAQ network.

To limit the implementation complexity we designed a simplified version of the TCP protocol. Several simplifications were possible because our data traffic flows only in one direction and because the DAQ network topology is fixed and designed with sufficient throughput to avoid packet congestion. But we preserved the full compliance with the RFC 793. Therefore we can use a PC with the standard Linux TCP/IP stack as a receiver.

The main simplifications includes:

1. We reduced the number of required TCP states from 11 to 3. The FEROL can open TCP connection and keeps the connection open until it is terminated. If an error is detected, the connection is terminated immediately.
2. The TCP complex congestion control was reduced to exponential back-off to decrease the throughput when temporary congestion is detected. A fast-retransmit algorithm is also implemented to improve the throughout in case a single packet loss is detected.

The current prototype board is equipped with low cost Altera Aria II GX FPGA, where 30% of the available resources are required for the TCP/IP and Ethernet interface. The board also contains 512MBytes of DDR2

memory for input and TCP socket buffer. Two TCP/IP engines were implemented allowing to open one or two simultaneous TCP streams.

The preliminary results show a maximum stable throughput of 9.7 Gbps for a direct connection between the FEROL prototype and a PC. With 16 TCP streams from 8 prototypes merged into one 40 Gbps interface via a switch we achieve 39.6 Gbps of stable data throughput.

We found that the maximum throughput is greatly sensitive to the receiver's PC configuration. In particular hyper-threading setting and network card IRQ and CPU affinity settings have to be tuned to achieve optimal performance.

Primary authors: SCHWICK, Christoph (CERN); GIGI, Dominique (CERN); ZEJDL, Petr (CERN)

Co-authors: HOLZNER, Andre Georg (Univ. of California San Diego (US)); PETRUCCI, Andrea (CERN); SPATARU, Andrei Cristian (CERN); Dr RACZ, Attila (CERN); DUPONT, Aymeric Arnaud (CERN); NUNEZ BARRANCO FERNANDEZ, Carlos (CERN); DELDICQUE, Christian (CERN); HARTL, Christian (CERN); PAUS, Christoph (Massachusetts Inst. of Technology (US)); WAKEFIELD, Christopher Colin (Staffordshire University (GB)); MESCHI, Emilio (CERN); STOECKLI, Fabian (Massachusetts Inst. of Technology (US)); GLEGE, Frank (CERN); MEIJERS, Frans (CERN); BAUER, Gerry (Massachusetts Inst. of Technology (US)); Dr POLESE, Giovanni (University of Wisconsin (US)); SAKULIN, Hannes (CERN); BRANSON, James Gordon (Univ. of California San Diego (US)); Dr COARASA PEREZ, Jose Antonio (CERN); SUMOROK, Konstanty (Massachusetts Inst. of Technology (US)); MASETTI, Lorenzo (CERN); ORSINI, Luciano (CERN); Dr DOBSON, Marc (CERN); PIERI, Marco (Univ. of California San Diego (US)); SANI, Matteo (Univ. of California San Diego (US)); CHAZE, Olivier (CERN); RAGINEL, Olivier (Massachusetts Inst. of Technology (US)); Dr MOMMSEN, Remi (Fermi National Accelerator Lab. (US)); GOMEZ-REINO GARRIDO, Robert (CERN); ERHAN, Samim (Univ. of California Los Angeles (US)); CITTOLIN, Sergio (Univ. of California San Diego (US)); MOROVIC, Srecko (Institute Rudjer Boskovic (HR)); BEHRENS, Ulf (Deutsches Elektronen-Synchrotron (DE)); O'DELL, Vivian (Fermi National Accelerator Laboratory (FNAL)); OZGA, Wojciech Andrzej (AGH University of Science and Technology (PL))

Presenter: ZEJDL, Petr (CERN)

Session Classification: Programmable logic, design tools and methods