The readout system upgrade for the LHCb experiment

Paolo Durante, *CERN,* Jean-Pierre Cachemiche, *CPPM,* Niko Neufeld, *CERN,* Guillaume Vouters, *CNRS,*
Federico Alessio, *CERN,* and Renaud Le Gac, *CPPM*

***Abstract*—The LHCb experiment is designed to study differences between particles and anti-particles as well as very rare decays in the charm and beauty sector at the LHC. The detector will be upgraded in 2019 and a new trigger-less readout system has to be implemented in order to significantly increase its efficiency.**

In the new scheme, event building and event selection are carried out in software and the event filter farm receives all data from every LHC bunch-crossing. Another feature of the system is that data coming from the front-end electronics is delivered directly into the event builders memory through a specially designed PCIe card called PCIe40.

The PCIe40 board handles the data acquisition flow as well as the distribution of fast and slow controls to the detector front-end electronics. It embeds one of the most powerful FPGAs currently available on the market with 1.2 million logic cells. The board has a bandwidth of up to 490 Gbits/s in both input and output over optical links and up to 100 Gbits/s over the PCI Express bus to the CPU.

We present how data flows through the board and to its associated server during event building. We focus on specific issues regarding the design of the different firmwares being developed for the FPGA, showing how to manage flows of 100 Gbits/s, and all the techniques put in place when different firmwares are developed by distributed teams of sub-detector experts.

I. Introduction

The LHCb experiment at CERN will have to undergo a major upgrade in order to efficiently process the increased luminosity that the Large Hadron Collider will deliver during Run 3. A fundamental difference compared to the system currently in operation will be the elimination of the first-level hardware trigger, this leaves to the online system the onus to deliver the full event rate produced in the collider to the High Level Trigger cluster.

II. Readout system architecture

The architecture of the overall system is illustrated in Fig. 1, its main components can be described as follows:

A. Optical links

Data is received from the underground experiment over simplex optical links, covering a distance of about 300 meters [1]. Frontend electronics mount radiation-hardened optical modules developed at CERN [2], with a data rate of 4.8 Gbps.

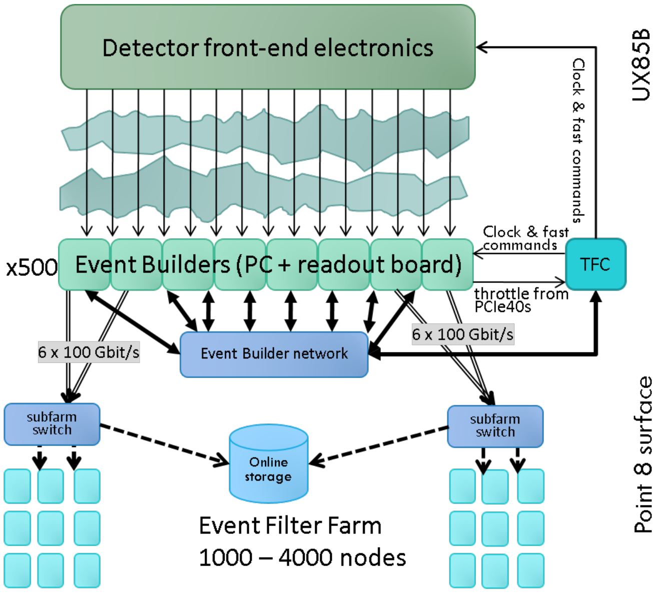

Fig. 1. High-level overview of the readout system

B. Timing and Fast Control

The TFC subsystem controls and monitors all frontends and all readout units using dedicated duplex optical links. At the same time, this subsystem provides the global timing reference that is used to keep the entire system synchronized with the LHC bunch structure. The same readout board used for data acquisition and event building is also used for the TFC.

C. Event builders

Each readout unit receives data directly from only a small fraction of the detector. In order to reconstruct the global state of a specific collision on a dynamically chosen readout unit, that node has to receive from all other even builders their local data fragments associated to the same collision. This distributed "event building" process occurs in real-time over a bidirectional network. Several technologies are currently under evaluation in order to implement this network at the required level of performance (100 Gbps per node in each direction).

D. Filter farm

Once events are fully assembled, they can be forwarded to the event filter farm where the actual high-level software trigger executes.

978-1-5090-2014-0/16/$31.00 ©2016 IEEE

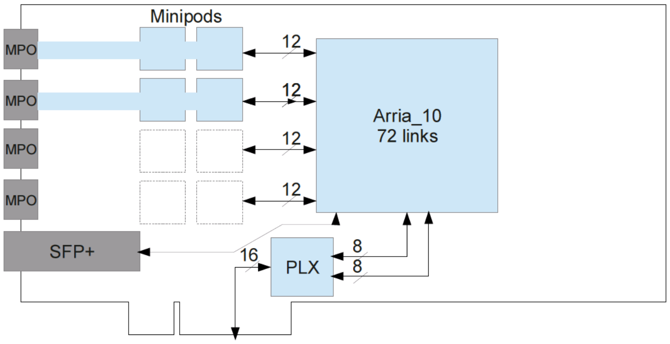

Fig. 2. High-level schematic of the PCIe40 board

III. The PCIe40 readout board

The diagram in Fig. 2 shows the main building blocks of the PCIe40 hardware platform. An Arria 10 FPGA provides over a million reconfigurable logic cells and 72 high-speed serializers. A variable number of these serializers (depending on board configuration) is connected to simplex or duplex optical links for the various subdetector frontends. At the opposite end of the board dataflow, 16 high-speed serial links implement a PCI Express Gen3 interface towards a COTS event building server.

Frontend optical links are implemented using Avago MiniPOD^{TM} transceiver modules. The PCI Express protocol on the backend is implemented directly by hardened logic inside the FPGA. As hardened logic is currently limited to 8 bidirectional PCIe lanes, implementing the 16 lanes necessary for the required I/O performance requires bonding two x8 interfaces. Link aggregation can be implemented by an additional discrete component on the board itself, or by enabling so called link “bifurcation” in the host root complex. The second solution has obvious advantages in terms of layout simplification, BOM reduction and power efficiency.

To conclude, an additional SFP+ optical interface is reserved for the bi-directional optical link used to propagate timing and fast commands through the TFC hierarchy.

As a consequence of the output channel between the board and the host being split into two logical PCI Express links, the dataflow inside the data acquisition firmware is also split into two simmetric halves. In this architecture, each half decodes, aggregates and optionally processes the data received from part of the optical input channels. In the end, the stream containing event fragments and timestamped with global synchronization information is transmitted to the host to initiate the distributed event-building process, through a high-performance DMA mechanism explicitly implemented for this application. The throughput across the PCI Express link, in the downstream direction (from the add-in card to the root complex), has been measured in excess of 111 Gbps.

Optimizing the utilization of the available I/O resources is critical for the feasability of the entire architecture and requires attention to all final consumers of the available bandwidth budget:

A. Data

Frontend data is received in different formats depending on its origin, in order to optimize the installed optical capacity while also trying to simplify the design of the different kinds of frontend circuitry. The readout board firmware has to synchronize and aggregate all the available input links, optionally perform some data processing and finally present a data stream at the output in a yet different format, this time taking into consideration the different set of constraints imposed by the PCI Express transport.

B. Metadata

As the event-building process consumes several times the amount of bandwidth required to simply read out the PCIe40, stream metadata should be separately provided by the firmware in order to accelerate stream parsing inside the event building application.

C. Slow control

Monitoring and configuration commands that are continously exchanged between the control system and the board over the same PCI Express link used for data acquisition must not negatively affect DAQ performance.

IV. Infrastructure

With the increase in the event rate and complexity to be processed by the high-level trigger, the datacenter infrastructure hosting it also requires an important upgrade, including a relocation from the current underground area to the surface. The new infrastructure that will be required to support such a design is currently undergoing an important review in order to finalize the final design during the course of this year.

V. Conclusions

The implementation of a full-software physics trigger at the scale required by the LHCb experiment and while leveraging the network and server technologies available on the commercial market is a complex technical task.

Current trends towards high-speed interconnects and high-density data-centres can be exploited to minimize cost and simplify the overall system layout.

The development of custom electronics in the online system is limited to a common high-performance hardware platform, the PCIe40.

Its generic design makes the PCIe40 suitable for a variety of fundamental tasks (data acquisition, slow and fast control), and can be applied as well outside of the LHCb experiment (as confirmed by its adoption by the ALICE experiment).

Development of hardware and software is advancing according to schedule in preparation for commissioning during the second long shutdown of the LHC.

References

[1] R. Schwemmer, J. Cachemiche, N. Neufeld, C. Soos, J. Troska, and K. Wyllie, “Evaluation of 400 m, 4.8 Gbit/s Versatile Link lengths over OM3 and OM4 fibres for the LHCb upgrade,” *Journal of Instrumentation*, vol. 9, no. 03, p. C03030, 2014.

[2] C. Sos, M. B. Marin, S. Dtraz, L. Olanter, C. Sigaud, S. Storey, J. Troska, F. Vasey, and P. Vichoudis, “The Versatile Transceiver: towards production readiness,” *Journal of Instrumentation*, vol. 8, no. 03, p. C03004, 2013.

[3] P. Durante, N. Neufeld, R. Schwemmer, U. Marconi, G. Balbi, and I. Lax, “100 Gbps PCI-Express readout for the LHCb upgrade,” *Nuclear Science, IEEE Transactions on*, vol. 62, no. 4, pp. 1752–1757, 2015.