Journal of Physics: Conference Series **513** (2014) 012023 doi:10.1088/1742-6596/513/1/012023

A PCIe Gen3 based readout for the LHCb upgrade.

M Bellato[1], G Collazuol[4], I D'Antone,[3] P Durante[2], D Galli[5], B Jost[2], I Lax[3], G Liu[2], U Marconi[3], N Neufeld[2], R Schwemmer[2] and V Vagnoni[3]

[1] INFN Sezione di Padova
[2] CERN European Centre for Nuclear Research
[3] INFN Sezione di Bologna
[4] Università di Padova and INFN Sezione di Padova
[5] Università di Bologna and INFN Sezione di Bologna
Production Editor, *Journal of Physics: Conference Series*, IOP Publishing, Dirac House, Temple Back, Bristol BS1 6BE, UK

E-mail: umberto.marconi@bo.infn.it

Abstract. The architecture of the data acquisition system foreseen for the LHCb upgrade, to be installed by 2018, is devised to readout events trigger-less, synchronously with the LHC bunch crossing rate at 40 MHz. Within this approach the readout boards act as a bridge between the front-end electronics and the High Level Trigger (HLT) computing farm. The baseline design for the LHCb readout is an ATCA board requiring dedicated crates. A local area standard network protocol is implemented in the on-board FPGAs to read out the data. The alternative solution proposed here consists in building the readout boards as PCIe peripherals of the event-builder servers. The main architectural advantage is that protocol and link-technology of the event-builder can be left open until very late, to profit from the most cost-effective industry technology available at the time of the LHC LS2.

1. Introduction

The LHCb experiment is designed to perform high-precision measurements of CP violation and search for New Physics by exploiting the decays of the beauty and charm hadrons copiously produced at the LHC. LHCb is expected to take in excess of 8 fb^{-1} by 2018 by recording data at a constant luminosity of $4. \times 10^{32}$ $\text{cm}^{-2}\text{s}^{-1}$ (twice the design luminosity and more than a factor four the average number of interactions per crossing at $\mu = 1.5$). During the period 2015-2018 the accelerator will increase the total center-of-mass energy to 13 TeV and will decrease the bunch spacing from the current 50 ns to 25 ns. Accordingly, the amount of beauty and charm quarks generated by LHC will double, while the pileup of the events will reduce. The prospect to augment the physics yield in the LHCb dataset looks therefore very promising.

Unfortunately, the LHCb capabilities are reduced because of the limited bandwidth available to the HLT, which is constrained by design to the hardware trigger maximum bandwidth of 1 MHz. This bandwidth limitation puts a hard limit of about 2 fb^{-1} that can be recorded per year. Additionally, the limited detector data available to the HLT trigger would limit the physics yield for hadronic decays even at higher trigger rates.

In order to remove these design limitations we plan to upgrade the spectrometer by 2018. The strategy for the upgrade consists of ultimately in removing the first-level hardware trigger

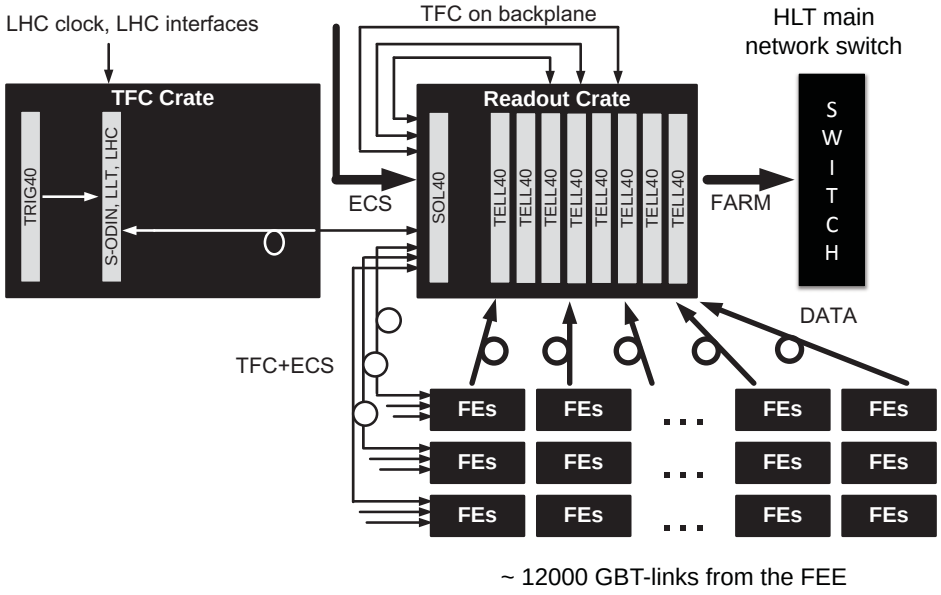

Figure 1. The ATCA-based readout system for the LHCb upgrade. A readout board (TELL40) is an ATCA compliant carrier-board, equipped with four active AMC40-card mezzanines.

to run the detector in a trigger-less mode. The Letter Of Intent [1] and the Framework TDR [2] document the plans for the upgraded detector. By running the detector at a leveled constant luminosity of $1 - 2 \times 10^{33}$ cm^{-2}s^{-1} the upgraded detector will enable the LHCb experiment to increase the yield of semi-leptonic decays with muons by a factor of 10, the yield of hadronic decays by a factor 20 and to record at least 50 fb^{-1} of data in total, .

The architecture is designed to allow for data transmission directly to the HLT computing farm synchronously with the LHC bunch-crossing at the full rate of 40 MHz. This requires in the order of 12.000 GBT optical links (available user bandwidth 3.2 Gb/s, radiation-tolerant, error-correcting link, developed for the LHC experiments by the CERN electronics group) [3] for a corresponding throughput of about 40 Tb/s. However, initially and during times of temporary congestion the data transfer rate will be tuned by means of a new Low Level Trigger (LLT), based on custom hardware, which will allow to vary the HLT input frequency in a range between 10 and 40 MHz.

In the proposed baseline readout architecture, shown in Figure 1, the readout boards (TELL40) act as the event buffers and data format converters for the injection of the event fragments into the HLT computing farm for event building and event selection. The TELL40 consists of an ATCA compliant carrier-board hosting up to four "AMC40-card" plugged onto the carrier board. Each AMC40-card is equipped with a single powerful FPGA (ALTERA Stratix V or newer) used for establishing high-speed serial connections and data processing. The proposed AMC40-card prototype provides 24 GBT-link input and 12 LAN-link output. The 24 inputs deliver a maximum amount of user-data of 77 Gb/s in the GBT standard mode and 115 Gb/s in wide mode. For the injection of the event fragments into the event-builder it has been foreseen

to implement a local area network protocol directly in the FPGA of the AMC40-card. The LAN candidate technology is 10 GbEthernet and the network protocol foreseen is UDP.

The LHCb experience with the UDP protocol over Ethernet is that it works but required a very careful tuning of the network devices. It has been demonstrated that this scheme can work reliably only when at least the first layer of switches, the one connected directly to the readout-boards, has very deep buffers (and hence these switches are expensive).

One can observe that the baseline implementation requires dedicated crates and that the implementation of the Ethernet protocol on FPGA is rather expensive, requiring the consumption of about 20% of the FPGA resources. In addition, the implementation of the network protocol requires buffering in the sender card and this implies the inclusion of a DDR3 memory interface to the FPGA. Modern LAN protocols use fast serializers (14 Gbit/s for FDR InfiniBand for instance) to reduce the number of lanes on the network. It can be expected that at the time-frame of the LHCb upgraded 25 Gbit/s serializers will be widely used in the LAN. However, for the majority of the serializers driving the GBT links, 6 Gbit/s is sufficient, and unfortunately, not all combinations of fast and slow serializers are readily available for any given device-family of FPGA.

2. An alternative readout using PCI Express.

PCI Express (PCIe) is the high-speed serial computer expansion bus standard designed to replace the older PCI bus [4]. PCIe devices communicate via point-to-point serial links between PCIe ports allowing both to send/receive requests and interrupts. The physical level of the link is composed of one or more lane. Each lane is composed of two differential signaling pairs: one pair for receiving the other for transmitting. Low-speed peripherals use a single-lane link, while, for instance, a graphics adapter board (GPU) typically uses a much wider 16-lane link. PCIe communication is encapsulated in transaction layer packets (TLP) and the Data Link Layer ensure reliable delivery of the TLP between two endpoints via an acknowledgement protocol. Like other high data rate serial interconnect systems, PCIe has a protocol and processing overhead due to the additional CRC and acknowledgements. The nowadays available standard PCIe generation 3 (PCIe-3) carries a bit rate of 8 Gbit/s per lane, with on overhead of about 2%, due to a 128b/130b encoding scheme.

The readout solution based on PCIe-3, currently under study as an alternative solution to the ATCA-based readout, consists of developing the LHCb readout boards as PCIe-3 standard boards, named PCIe40, which act as add-on cards in the motherboards of the HLT event-builder servers. In this approach, represented schematically in Figure 2, data from the front-end electronics are transmitted over the GBT-links directly to the event-builder PCs RAM via the PCIe40. Consecutive event fragments transmitted from the front-end electronics are received and buffered at the PCIe40 in Multi Event Packets (MEP) of suitable size and then copied into the event-builder server RAM by means of DMA through PCIe.

The proposed PCIe40 is equipped with 24 GBT-links, as the AMC40-card described above, and it is directly connected to the motherboard through 16-lane edge-connector. The PCIe readout requires therefore about 500 PCIe40 cards to read out the whole detector and the same number of event-builder servers, assuming to have just one PCIe40 card connected to one server.

Modern FPGAs, like for instance the above mentioned ALTERA Stratix V, offer several embedded PCIe-3 hard IP blocks to implement the protocol. The PCIe hard IP blocks available in the FPGAs are generally very efficient: one 8-lane block uses less than 1% of the resources. The Stratix V allows to instantiate two PCIe-3 8-lane hard IP devices that can be merged to one 16-lane interface to reach the theoretical transmission capability of 128 Gbit/s. The PCIe40 card logic scheme is shown in Figure 3. In order to provide the required input rate the PCIe40 can be equipped with 24 (or 36) input optical-to-electrical transducers, connected to a single FPGA (a newer ALTERA, Arria 10, for instance) and a PCIe switch chip, which is needed to

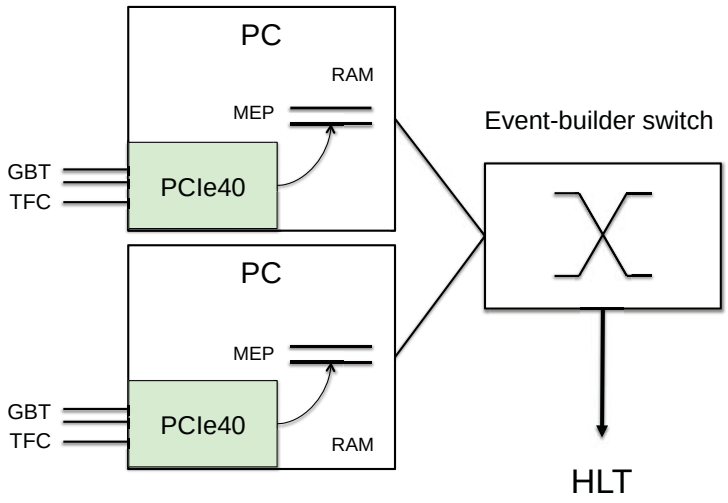

Figure 2. The PCIe based readout system. The PCIe40 readout boards are directly connected to the event-builder PCs through 16-lane PCIe edge-connector. The input rate to the server RAM through the PCIe40 board is of about 80 Gb/s.

merge two 8-lane PCIe-3 IP blocks to one 16-lane PCIe edge-connector. The PLX company offers a wide range of components (PXE 8733 chip as an example) to implement the required switch functionality at low cost and low power.

Apart from potential cost-savings there are several advantages with the PCIe proposed solution:

- PCIe reduces the FPGA firmware complexity of the readout board with respect to the AMC40. No higher level network and data transport protocol is needed. The network and higher level transport protocols are provided by the OS of the cards host system.
- PCIe solution increases the DAQ flexibility. Different DAQ schemes can be easily implemented in software compared to hardware description languages in FPGA (as it would be required with the ATCA based readout).
- Due to the buffering capacity and flexible DAQ scheme, many more choices for network technologies and devices are available to implement the DAQ network. We do not need to rely on the availability of expensive carrier-class routers (with deep buffers) and pay for many advanced features which are not needed in our system.
- PCIe is most likely the most long lived protocol, apart from Ethernet. In the timescale relevant for the project (until 2023) it automatically leverages any development, which will come up in the world of Intel computing servers, which is one of the most advanced and fast-moving technologies available with a huge market-base.

3. GPU emulation of the PCIe40 data traffic.

In order to assess the feasibility of the readout project based on the PCIe40 boards we aimed to measure the effective bandwidth available to PCIe boards writing data into the RAM of modern host PCs. To emulate a PCIe40 board we used PCIe-3 based GPUs equipped with

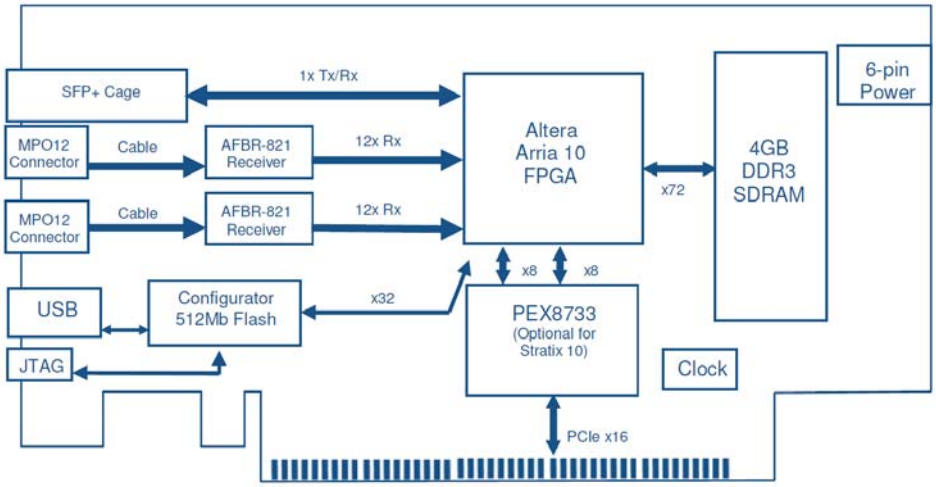

Figure 3. The PCIe40 readout board logic scheme.

16-lane edge-connectors. The assumption is that a GPU transferring data unidirectionally from its internal memory to the host CPU memory behaves similarly to a PCIe40. As a matter of fact they both use DMA memory-to-memory data transfer through the PCIe link. We measured the performance of several GPUs and motherboards combined systems, arranged in different setups. In any case, the hardware used for tests is the very first commercially available that support PCIe-3. We used the CUDA (Compute Unified Device Architecture by NVIDIA) API [5] to develop our own test programs, or to adapt preexisting ones, to vary the dimension of the record to be transferred, the number of active threads and the type of allocated memory. It is worth mentioning here that GPUs can directly access page-locked (pinned) host memory only (the host operating system never swaps pinned memory to disk) so that the best performance can be achieved by allocating pinned memory. Figure 4 shows the observed dependence of the bandwidth on the record size and the stability of the performance on long runs, lasting up to 10 hours. The record sizes in these measurements have been varied between 10 kB and 2 GB. We generally measured transfer rate above 100 Gb/s by transferring record sizes greater than few MiB. Table 1 shows the results achieved with different setups at the given reference points chosen for comparison.

On the basis of the results we can conclude that it is possible to reach more than 90% of the theoretical bandwidth of PCIe for a long time. The GPU uses 256 byte PCIe transfers instead of the maximum of 2 kB. More throughput might be achievable by reducing the overhead. The PC hardware can already now handle the expected bandwidth of 24 GBT-links, of approximately 77 Gb/s, with about 30% margin. CPU time spent in kernel space (the part that is actually responsible for handling the DMA transfers) was about 50% of one core. This should also decrease with bigger transfer sizes. We observed that locking the process to a particular set of

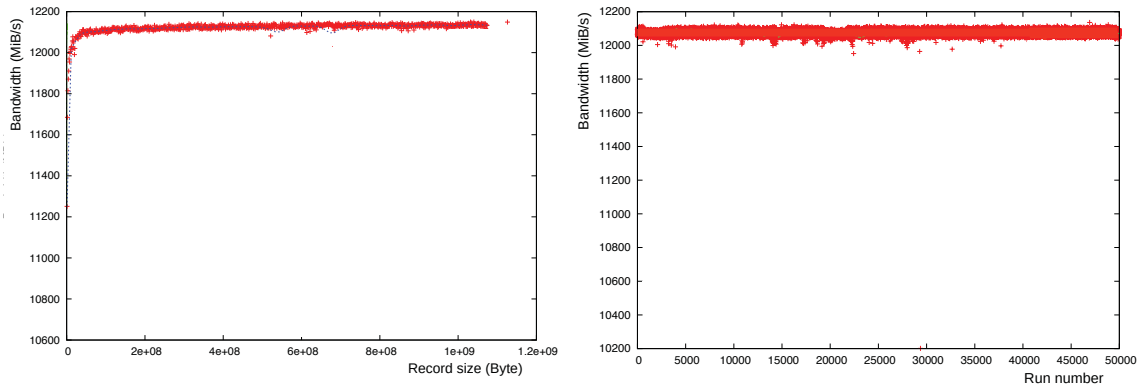

Figure 4. On the left, the measured bandwidth versus the record size. On the right, repeated measurements with the data record size set to 32 MiB (the entire measurement process lasted for about 6 hours). Both plots refer to the setup consisting of a GPU NVIDIA GTX770 connected to the motherboard Supermicro X9DRD-iF.

Table 1. GPU to host RAM data transfer rates comparison. Index in the first column distinguishes the various setup used to measure and compare the data transfer rates: 1) GPU NVIDIA GTX770, motherboard Supermicro X9DRD-iF; 2) GPU GTX Titan, motherboard ASRock - Z77 Extreme; 3) Twofold synchronous data transfer using identical PCIe-2 GPUs NVIDIA Tesla K20m, connected to the PCIe-3 motherboard slots, motherboard Supermicro X9DRG-HF.

Setup	Record Size (MiB)	Bandwidth (Gb/s)	σ (Gb/s)
1	32	101.3	0.1
2	32	98.5	0.1
3	32	100.2	0.1
1	1500	101.8	0.1
2	1500	100.4	0.1
3	1500	107.3	0.1

memories and CPUs is necessary for achieving stable throughput, avoiding spurious drops in performance. The average throughput does not depend on the locking, which shows that the QPI link between CPUs can also handle the necessary bandwidth.

We consider these achievements as an important step to demonstrate the feasibility of the PCIe based readout for the LHCb upgrade. We believe that the FPGA implementation of the PCIe protocol, provided as hard coded IP blocks (together with the DMA controller) by the FPGA producers (ALTERA for instance), is a guarantee of reliability. Evaluation boards are available on the market to test and evaluate the PCIe implementation performance.

In the PCIe approach data flows directly from the detector fronted electronics to the RAM of the event-builder servers. Control and management of the readout system is therefore facilitated, concerning the event-builder farm only.

The PCIe weak points are the event-builder servers, which of course could fail. It will be important to keep the server OS environment robust, relying also on elevated hardware

redundancy. In this respect we foresee to have a standby node within each rack, so that in case of a catastrophic failure the network cables simply need to be swapped rather than a cumbersome replacement of the PCIe40 card. We do not plan to go through a redistribution layer in the network, which would allow failing over between multiple paths from one data-source. This has not been a problem for LHCb in the past, and we are confident it will not be in the future since the network hardware is usually extremely reliable.

Throughput requirements for a single event-builder server are demanding, since it has to manage four data streams each of about 80 Gb/s: the input rate from the PCIe40 card to the RAM (77 Gb/s from the 24 GBT-links input); the output to and the input from the other servers through the switch for the event-building; finally, the output rate to the HLT (at the full rate of about 80 Gb/s in case no trigger selection is performed by the event-builder server). It has to be proven that these data throughputs can be maintained in a realistic prototype, by a server that while performing PCIe data transfers has to transmit and receive data over a commercial network (via InfiniBand and/or GbEthernet) using a standard protocol.

References

[1] The LHCb Collaboration. Letter of Intent for the LHCb Upgrade. CERN-LHCC-2011-001; LHCC-I-018. Available at "http://cds.cern.ch/record/1333091/files/LHCC-I-018.pdf".
[2] The LHCb Collaboration. Framework TDR for the LHCb Upgrade: Technical Design Report. CERN-LHCC-2012-007 ; LHCb-TDR-12 Available at http://cds.cern.ch/record/1443882/files/LHCB-TDR-012.pdf
[3] P Moreira, R Ballabriga, S Baron, S Bonacini, O Cobanoglu, F Faccio, T Fedorov, R Francisco, P Gui, P Hartin, K Kloukinas, X Llopart, A Marchioro, C Paillard, N Pinilla, K Wyllie, B Yu. The GBT project. Available at http://ds.cern.ch/record/1235836/files/p342.pdf
[4] PCI Express 3.0 specifications. Available at http://www.pcisig.com/specifications/pciexpress/
[5] CUDA. The Compute Unified Device Architecture by NVIDIA. Available at https://developer.nvidia.com