Trigger-less readout architecture for the upgrade of the LHCb experiment at CERN

F. Alessio¹

CERN, Geneva, CH-1211, Switzerland

E-mail: Federico.Alessio@cern.ch

ABSTRACT: The LHCb experiment has proposed an upgrade of its detector in order to collect data at ten times its initial design luminosity. The current readout architecture will be upgraded by removing the existing first-level hardware trigger whose efficiency is limited for hadronic channels at high luminosity. The new readout system will record every LHC bunch crossing and send data to a trigger selection process performed entirely by software running in a computing farm. Therefore, the new readout system must cope with higher sub-detector occupancies, higher rate and higher network load. In this paper, we describe the architecture, functionalities and technological challenges of such an upgraded system.

KEYWORDS: Data acquisition concepts; Trigger concepts and systems (hardware and software); Front-end electronics for detector readout; Detector control systems (detector and experiment monitoring and slow-control systems, architecture, hardware, algorithms, databases)

¹On behalf of the LHCb collaboration.
1 The upgrade of the LHCb experiment

The LHCb experiment [1] is a high-precision experiment at the LHC devoted to the search for New Physics by precisely measuring its effects in CP violation and rare decays. By applying an indirect approach, LHCb is able to probe effects which are strongly suppressed by the Standard Model, such as those mediated by loop diagrams and involving flavor changing neutral currents. This is a powerful approach as it allows accessing larger mass scales than those reachable by direct searches at the other LHC experiments. These apply a more direct approach and are bounded to the center-of-mass energy scale accessible at the LHC.

In the proton-proton collision mode, the LHC is to a large extent a heavy flavor factory producing over 100,000 bb-pairs every second at the nominal LHCb design luminosity of $2 \times 10^{32}$ cm$^{-2}$s$^{-1}$. Given that bb-pairs are predominantly produced in the forward or backward direction, the LHCb detector was designed as a forward spectrometer with the detector elements installed along the main LHC beam line, covering a pseudo-rapidity range of $2 < \eta < 5$ well complementing the other LHC experiments ranges.

LHCb proved excellent performance in terms of data taking [2] and detector performance, operating with close to 99% of working detector channels and recording more than 90% of the data delivered by the LHC adding up to a total of about 3 fb$^{-1}$ of data saved to disk over the period 2010–2012. More than 99% of the data recorded online was already good for offline analyses, benefiting from the excellent background conditions at the LHC. This corresponds to about $60 \times 10^{11}$ cc-pairs and $26 \times 10^{10}$ bb-pairs on disk. The experiment was able to record such a high number of events by operating at more than a factor of two beyond its nominal design luminosity and by more than a factor four in terms of average number of interactions per crossing $-\mu = 1.5$ instead of the nominal $\mu = 0.4$. LHCb proved itself to be the Forward General-Purpose Detector at the LHC well complementing the physics results from ATLAS and CMS.

After the LHC Long Shutdown I (2013–2014), the LHCb experiment will keep recording data, improving the statistics of its measurements. It is expected to take in excess of 5 fb$^{-1}$ by 2018,
recording data at a constant luminosity of $4 \times 10^{32} \text{cm}^{-2}\text{s}^{-1}$. Hence, the total LHCb dataset will add up to $\sim 8 \text{fb}^{-1}$ by 2018. During the same period, however, the LHC accelerator will increase the total center-of-mass energy to 13 TeV and it will change the bunch spacing from the current 50 ns to the nominal 25 ns. This is expected to allow the LHC to reach its nominal designed luminosities of about $2 \times 10^{34} \text{cm}^{-2}\text{s}^{-1}$ in ATLAS and CMS while keeping the complexity of events low as $\mu$ will decrease. In LHCb, the increased energy and reduce bunch spacing after the Long Shutdown I means that the amount of beauty and charm decays generated will essentially double, while reducing the complexity of events by a factor two. As the current LHCb detector is already performing very well in conditions harsher than nominal and the LHC accelerator will keep improving its performance, the prospect to augment the physics yield in the LHCb dataset seems very attractive. However, the LHCb detector is limited by design in terms of data bandwidth — 1 MHz instead of the LHC bunch crossing frequency of 40 MHz — and physics yield for hadronic channels at the hardware trigger, therefore it will not be possible to increase the physics yield during the period 2015–2018 even though LHCb will see the amount of decays increase and its complexity decrease. Figure 1 illustrates the physics yield as a function of the instantaneous luminosity at the LHCb first-level hardware trigger. At a luminosity of $4 \times 10^{32} \text{cm}^{-2}\text{s}^{-1}$ the physics yield for hadronic channels is essentially half of the one for muonic channels. A Letter Of Intent [3] and Framework TDR [4] document the plans for an upgraded detector which will enable LHCb to increase the yield in the decays with muon by a factor of 10, the yield for hadronic channels by a factor 20 and to collect $\sim 50 \text{fb}^{-1}$ at a leveled constant luminosity of $1-2 \times 10^{33} \text{cm}^{-2}\text{s}^{-1}$. This corresponds to ten times the current design luminosity and increased complexity (pileup) of a factor 5.

2 Strategy for the upgrade of the experiment

In order to remove the main design limitations of the current LHCb detector, the strategy for the upgrade of the LHCb experiment essentially consists of ultimately removing the first-level hardware trigger entirely, hence to run the detector fully trigger-less. Currently, the Calorimeters and Muon detectors provide $p_T$, $E_T$ and muon information to the first-level hardware trigger. Based on
that information, the rate of accepted events is reduced from the LHC bunch crossing frequency of 40 MHz to about 1 MHz. Events are then built and transmitted through a $\sim 1$ Tb/s readout network to a processing farm consisting of $\sim 30000$ cores. A software trigger running on each core of the processing farm analyses the events and overall selects about 5 kHz of them to be saved to disk.

Figure 2 illustrates the main differences between the current readout architecture and the proposed upgrade. By removing the first-level hardware trigger (L0) in the upgraded scenario, LHC events are recorded and transmitted from the Front-End electronics (FE) to the readout network at the full LHC bunch crossing rate of 40 MHz, resulting in a $\sim 40$ Tb/s network. All events will therefore be available at the processing farm where a fully flexible software trigger will perform selection on events, with an overall output of about 20 kHz of events to disk. This will allow maximizing signal efficiencies at high event rates.

The direct consequences of this approach are that some of the LHCb sub-detectors will need to be redesigned to cope with an average luminosity of $2 \times 10^{33}$ cm$^{-2}$s$^{-1}$ and to be equipped with new trigger-less Front-End electronics, and the entire readout architecture must be redesigned in order to cope with the upgraded multi-Tb/s bandwidth and a full 40 MHz dataflow [5]. The main technological choices and solutions for the upgraded readout architecture are presented below.

3 The upgraded LHCb readout architecture

In the upgraded readout architecture (figure 3), all FE electronics record and transmit data continuously at 40 MHz to the Readout Boards. The expected non-zero suppressed event size would result in a very large number of links between the FE and the new Readout Boards (here Back-End or BE). In this regard, it has been shown [5] that almost a factor of ten could be gained by sending zero-suppressed data already at the FE, reducing the number of links from $\sim 80000$ to $\sim 12500$. The zero-suppression will thus be performed in radiation-tolerant FE chips. A possible consequence of zero-suppression is a varying latency of data transmission. Data are essentially
transmitted asynchronously to the Readout Boards which must be able to absorb the asynchronous data-driven readout and variable latencies across data links. Synchronicity between the FE and BE is ensured by tagging each event with a unique identifier, the Bunch Crossing ID (BXID). Once all event fragments corresponding to a particular BXID are received by the Readout Board, the board will compile a packet of events (Multi-Event Packet, MEP) and send it across the event-building network (DAQ) to a particular processing node. Enough memory must be envisaged at the Readout Boards to allow for latencies of the order of a $\mu$s.

Synchronicity across the entire system is ensured by the upgraded Timing and Fast Control System (TFC) [6]. This system will distribute timing information, clock and fast commands to the Readout Boards as well as to the FE electronics, by profiting from the capabilities of the CERN GigaBitTransceiver (GBT) [7] and by being interfaced to the LHC accelerator timing and clock system [8]. FE chips will not receive triggers, but rather commands associated to a particular BXID (e.g. calibration commands, reset commands, periodic commands). These commands will be computed centrally well in advance by the TFC system based on configurable parameters and the LHC filling scheme. The TFC system will also be responsible to handle back-pressure (throttle). Back-pressure can happen due to high readout load or slow processing time at the Readout Board or at the processing node, keeping the processing farm busy and not available to accept more events. The TFC system will monitor the availability of processing nodes by collecting requests of events from each processing node and selecting to which processing node an event must be sent.

It should be noted that although the final system will ultimately be fully trigger-less, a first-level hardware trigger based on the current L0 trigger will be maintained. This is commonly referred to as Low Level Trigger (LLT or Trigger Processing in figure 3) and its main purpose is to allow a staging of the DAQ network, gradually increasing the readout rate from the current 1 MHz.
2013 JINST 8 C12019

Figure 4. Generic FE module for the LHCb upgrade.

to the full and ultimate 40 MHz as more network nodes are added. The final decision whether to accept an event or not will be taken by the TFC system basing its decision on the LLT output, on the network load and on processing nodes availability. However, while the current first-level trigger selects which events to retain, in the upgraded architecture it will mostly act as a rejection mechanism, suggesting which events to reject in case of back-pressure or low processing availability.

Finally, the new TFC system will also profit from the bidirectional capability of the CERN GBT by interfacing the FE electronics to the global LHCb Experimental Control System (ECS). This will be done by transmitting configuration data to the FE electronics and receiving monitoring data from the FE electronics and relaying it back to the global LHCb ECS by means of a TFC+ECS Interface card, called SOL40.

3.1 Trigger-less Front-End electronics

Figure 4 illustrates a generic FE module compatible with the requirements of the LHCb 40 MHz trigger-less readout architecture. A generic FE chip will convert analog signals into digital signals; it will zero-suppress (or compress) events and it will pack them onto the data link via the use of a derandomizing buffer. LHCb FE electronics will extensively profit from the GBT chipset capabilities to transmit data to the Readout Boards.

While compression or zero-suppression mechanisms are essential to reduce the total number of optical links from the FE to the Readout Boards, an efficient packing mechanism has been envisaged to allow different length of packets to be sent over the link in an efficient way.

The idea is illustrated in figure 5, where events of different lengths and belonging to consecutive LHC crossings are packed dynamically across the GBT frame, which has a fixed length. The general rule is that the average event size must match the link bandwidth in order not to overflow the output buffer of the FE module. The depth of the output buffer can vary from one detector to another one and must be able to absorb fluctuations in occupancy so that buffer overflow happens for less than 0.01% of the events. In addition to the BXID event identifier, each event must contain
the length of the packet so that the Readout Boards can decode each event packet and find the next header for the next event. This type of mechanism is extremely useful in the case of detectors with low occupancy (∼1–3%). One GBT link can be connected to hundreds of channels and events can be transmitted without incurring in fragment truncation if an isolated event is too big to fit in the fixed GBT width. Other specific information can be added to the header of an event if needed.

The drawback of this solution is that the quantity of resources used in the encoding and decoding chips is bigger than in the case of fixed packing mechanisms due to the dynamic packing algorithms. Simulation efforts are on-going in order to find the best trade-off point between resource utilization and bandwidth exploitation and to meet each sub-detector’s requirements in terms of occupancy, number of optical links and data transmission robustness.

Due to the complexity of the algorithms to be implemented in the FE, both for compression and zero-suppression logic and for the dynamic packing, a campaign of studies has started in order to evaluate the use of Field Programmable Gate Arrays (FPGA) in the relatively modest radiation environment of LHCb. First tests were performed on Actel A3PE1500 FPGAs [9] and Altera Arria GX FPGAs [10] for integrated doses up to ∼50 krad concluding that the cutting value for current FPGAs technologies is at around 30 krad. The trade-off between FPGA flexibility and robustness has to be evaluated carefully.

3.2 Fast and slow control to FE

Due to the huge amount of data which are transmitted from the FE to the BE, it was chosen to separate the links between controls and data. By utilizing the capabilities of the GBT chip, it would be possible to transmit data, fast commands, clock and slow control all on the same bidirectional optical link. However, a simple calculation showed that the number of fibres would be reduced if data was sent over simplex 4.8 Gb/s data links (∼12500 of them) from the FE to the BE, and clock, fast and slow controls were sent and received over 4.8 Gb/s duplex links (∼2500 of them). This is illustrated in figure 6. The fast control information (TFC in figure 6) transmitted to the FE is the same for the whole FE electronics and can be fanned out more efficiently if it is sent over separate links driven by the same source rather than across all data links. Moreover, the on-detector GBT chip associated to the control links — commonly referred to as Master GBT — is also used to generate the clock to be used at the FE electronics to record data and drive the on-detector data GBTs. Therefore, it is logical to use a single instance of the Master GBT to drive many data GBTs.

It was estimated that about 1 Gb/s of bandwidth per link is required to control fully the FE electronics by means of commands and resets. Slow control configuration data instead is shipped to
Figure 6. Data is transmitted from the FE to the BE over simplex 4.8 Gb/s optical links. Clock, fast and slow control are transmitted and received via duplex 4.8 Gb/s optical links.

GBT-SCA [7] chips via e-links from the GBT chipset onboard the FE. Monitoring data is received upstream via the same duplex link, from the FE to the BE electronics. Each duplex link will have a total bandwidth of $\sim$ 1.3 Gb/s dedicated to slow control configuration and monitoring and 1 Gb/s dedicated to fast control. The leftover 1 Gb/s out of the effective data field is reserved in case more fast information should be transmitted in the future. Finally, 1.5 Gb/s is dedicated to error correction over the GBT link for a total of 4.8 Gb/s, resulting in a very robust and efficient link exploitation. A dedicated board, called SOL40, will be entirely devoted to control a slice of FE electronics and the details of how this is envisioned can be found in [13]. Such board is based on the same hardware as the Readout Board, but simply with a different firmware version.

The clock distribution to the FE will be provided via the same Master GBT link. The Master GBT will provide clocks with various frequencies and ensure constant latency and fine phase alignment. Moreover, robust resynchronization mechanisms are envisaged in order to have a robust and controlled timing distribution across the readout system as well as a quick and efficient recovery of the synchronization of the system.

3.3 The new Readout Board TELL40

The minimum requirements for both the detector and the readout architecture upgrades are to maintain the current excellent performance in harsher conditions, while keeping costs low. In this sense, common solutions and resources are being employed in the development of the new upgraded LHCb Readout Board, commonly referred to as TELL40.

This board is a compact, high density, FPGA-based ATCA [11] board. Thanks to this density, each module will be able to handle a throughput of $\sim$ 0.5 Tb/. The board can accept a total of 96 links at 5 Gb/s (4.8 Gb/s if considering the GBT protocol) and the connection to the processing farm will be via 48 x 10 Gigabit Ethernet (GbE) links. The board is a full ATCA card, composed of four Advanced Mezzanine Cards (AMC), each of them equipped with a powerful FPGA. For the first prototype, an ALTERA Stratix V FPGA was chosen, with 230,000 logical elements.

The board is so generic that by means of a firmware reconfiguration it can be used also for the fast control, clock distribution and slow control distribution and for the future Low Level Trigger. The firmware of the board will contain an interface layer code, common to all the boards, whose aim is to interface the hardware with the user code firmware using common blocks (like a GBT
decoder/encoder and a 10 GbE IP core). The actual user code firmware will define the flavor of the board and its functionality within the upgraded readout architecture. This reduces considerable the number of components to be developed within the upgrade of the readout architecture, optimizes manpower and optimises the effort on firmware design.

The same approach has also been put in place for a global simulation framework. The environment to develop the firmware for each flavor of the boards will be common across the entire LHCb experiment, with only the user code being exclusive. This collaborative method has been proven to be very effective in reducing unconformities and to enforce compatibility with specifications.

The user code dedicated to the functionality of the readout of events from the trigger-less FE faces considerable challenges. Events will arrive from the FE to the Readout Board asynchronously across all input links due to the variable latency in compression/zero-suppression mechanisms so the code has to be able to handle a big spread in latency between fragments of the same event. The readout code of the board must be able to decode the frames from the FE and realign them according to their BXID to then build an event packet and send it to the DAQ network. Common effort is ongoing to find the best technological solutions to this challenge.

TELL40 boards interface to the rest of the DAQ system by means of big core switches organized in a fat-tree topology, through which they route the fragments of each event to the corresponding processing node for event-building and processing. This is a classical approach of readout at the LHC.

Another more innovative approach takes ideas from the current and modern data-centres being used in many commercial environments. In this solution, a more compact, uniform and streamlined network could be envisaged in which each processing node is also a host for a PCIe Gen3 NIC readout board able to handle up to $\sim 100$ Gb/s. This board will actually act as a readout board and the event building is performed through a series of cheaper and commercially available switches in a more compact way. Disks and monitoring computers will also be within the same network, composing what could be seen as a readout box. An R&D phase is ongoing in order to evaluate the feasibility of this approach.

4 Conclusion

The LHCb experiment has proposed an upgrade of its detector and readout electronics in order to increase its physics yields by a factor 10 in muonic channels, a factor 20 in hadronic channels and a factor 10 in integrated luminosity by removing the first-level hardware trigger. The plan to upgrade the detector in 2018 consists of replacing some sub-detectors to newer technologies and to equip all of them with trigger-less Front-End electronics. Moreover, in order to cope with a multi-Tb/s readout network, the entire readout architecture has been redesigned and redeveloped as presented in this paper.

All LHC events will be recorded at the LHC bunch crossing frequency and made available to the processing farm in order to maximize signal efficiencies while increasing the readout rate. The installation and commissioning of the upgraded detector and readout system is planned for the LHC Long Shutdown II (2018–2019) over a period of about 18 months. After that, the LHCb experiment is expected to collect at least 50 fb$^{-1}$ at a leveled luminosity of $1–2 \times 10^{33}$ cm$^{-2}$s$^{-1}$. This is expected to increase dramatically the LHCb sensitivities beyond Flavor Physics and in the search for New Physics.
References

[1] LHCb collaboration, *The LHCb Detector at the LHC*, 2008 *JINST* 3 S08005.

[2] R. Jacobsson, *Performance of the LHCb Detector during the LHCb Proton Runs 2010–2012*, Proceedings of 2012 IEEE Nuclear Science Symposium and Medical Imaging Conference, Anaheim U.S.A. (2012), pg. 1479.

[3] LHCb collaboration, *Letter of Intent for the LHCb Upgrade*, CERN-LHCC-2011-001 (2011).

[4] LHCb collaboration, *Framework TDR for the LHCb Upgrade*, CERN-LHCC-2012-007 (2012).

[5] K. Wyllie et al., *Electronics Architecture for the LHCb Upgrade*, LHCb-PUB-2011-001 (2011) [astro-ph/0504449].

[6] F. Alessio and R. Jacobsson, *Timing and Fast Control for the Upgraded Readout Architecture of the LHCb experiment at CERN*, *IEEE Trans. Nucl. Sci.* 60 (2013) 3438.

[7] P. Moreira et al., *The GBT Ser-Des ASIC prototype*, 2010 *JINST* 5 C11022.

[8] S. Baron et al., TTC website, http://ttc.web.cern.ch/ttc.

[9] A. Pellegrino and W. Vink, *FPGA-based, radiation-tolerant on-detector electronics for the upgrade of the LHCb Outer Tracker Detector*, in proceedings of TWEPP 2013, Perugia Italy (2013).

[10] C. Faerber et al., *Radiation tolerance tests of SRAM-based FPGAs for the possible usage in the readout architecture for the LHCb experiment*, in proceedings of TWEPP 2013, Perugia Italy (2013).

[11] Advanced Telecommunications Computing Architecture, ATCA website, http://www.picmg.org/v2internal/resourcepage2.cfm?id=2.

[12] J.P. Cachemiche et al., *Study for the LHCb Upgrade readout board*, 2010 *JINST* 5 C12036.

[13] F. Alessio and R. Jacobsson, *A New Readout Control system for the LHCb upgrade at CERN*, 2012 *JINST* 7 C11010.