Event building and reconstruction at 30 MHz using a CPU farm

This content has been downloaded from IOPscience. Please scroll down to see the full text.
2014 JINST 9 C10029
(http://iopscience.iop.org/1748-0221/9/10/C10029)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 137.138.139.20
This content was downloaded on 21/11/2014 at 12:49

Please note that terms and conditions apply.
WORKSHOP ON INTELLIGENT TRACKERS,
14–16 MAY 2014,
UNIVERSITY OF PENNSYLVANIA, U.S.A.

Event building and reconstruction at 30 MHz using a CPU farm

J. Albrecht, a D. Campora, b V. Gligorov, b B. Jost, b U. Marconi, c,1 N. Neufeld, b
G. Raven d and R. Schwemmer b

a Technische Universität Dortmund,
August-Schmidt-Straße 4, 44227 Dortmund, Germany
b CERN,
Route de Meyrin 385, 1217 Meyrin, Switzerland
 c INFN Sezione di Bologna,
Via Irnerio 42, 40126 Bologna, Italy
d Vrije Universiteit Amsterdam,
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

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

ABSTRACT: The LHCb experiment will be upgraded between 2018 and 2019 in order to reach unprecedented precision on the measurements of the main observables of the beauty and charm quarks. This paper describes the trigger-less readout system foreseen for the upgrade.

KEYWORDS: Data acquisition concepts; Online farms and online filtering

1 Corresponding author.
1 Introduction

One of the main limitations of the current LHCb detector [1] is that the collision rate must be reduced to match the maximum rate at which all the sub-detectors can be readout of 1.1 MHz. The rate reduction is achieved by the Level-0 hardware trigger, which uses the basic events signatures available (calorimeters and muon system objects), operating within a fixed latency of few microseconds. Owing to its implementation the Level-0 causes the largest inefficiencies in the entire trigger chain, especially for purely hadronic decays of beauty and charm hadrons. Therefore, one of the main objectives of the LHCb upgrade, planned during the Long Shutdown 2 by 2018-2019, is to remove the hardware trigger bottleneck.

In the foreseen upgraded detector the event yields useful for physics will be maximised by operating a synchronous readout of each bunch-crossing. No more trigger decision is sent to the front-end electronics, making the upgraded LHCb readout completely trigger-free. This requires a change of all the front-end electronics of all the detectors. Also, several detectors will be replaced or upgraded. The upgraded LHCb will operate with an instantaneous luminosity of $2 \times 10^{33}$ cm$^{-2}$s$^{-1}$, five times higher the current value. The luminosity will be kept constant during the fill using the levelling scheme that was successfully operated during Run 1. At this luminosity the expected inelastic collision rate of about 30 MHz, will be processed entirely by the software trigger, which will run on the dedicated Event Filter Farm (EFF). The software trigger selections will be as similar as possible to those applied in offline analyses to maximise trigger efficiencies and to minimise systematic uncertainties.

Although the low level trigger (LLT) has not been chosen as part of the upgrade baseline design, the CPU power available in the event building farm permits implementing the LLT in software. The LLT could be useful to regulate the rate at the input of the EFF at the start of data taking of Run 3, if the EFF is not fully in place at the start of data taking. This solution represents the best compromise between cost, flexibility and added security.

A dataset of at least 50 fb$^{-1}$ will be collected by the upgraded experiment in less than ten years. These data will allow LHCb to reach unprecedented precisions in the analysis of the beauty and charm quark flavour state transitions.
2 System design

The architecture of the upgraded LHCb readout-systems is shown in figure 1. All the hardware of the DAQ and ECS/TFC system (slow control and timing and fast control respectively) will be concentrated in the foreseen data-centre located on the surface. This requires to operate the 4.8 Gb/s radiation-hard Versatile Links [2] over the approximately 300 m long optical fibres, which will be necessary to cover the distance between the underground areas (UX85B) and the surface [3]. Before sending data to the DAQ all detectors must perform zero-suppression on the front-end. Data will be then pushed over simplex links. Control and timing information use bi-directional links instead. On the majority of the links the so-called wide mode of the GBT will be used, with effective bandwidth of 4.5 Gb/s. The nominal event-size of 100 kB is estimated from the total of 8800 data links required (the average link usage factor is 80%) and assuming 30 MHz of non-empty bunches. The total number of optical fibres for both DAQ and ECS/TFC has been estimated to be 17000 as an upper limit, including spares.

The readout board, called the PCIe40 board, is a generic hardware component, which has been designed for the data acquisition of all detectors, the distribution of the timing and fast commands and the slow control. The different functionalities of the boards will be selected by firmware. Initially, several prototypes have been developed to check the feasibility of mapping the readout system over an ATCA architecture. Detailed studies of the evolution of the network technologies and the global optimisation of the readout system have shown that a cost-effective implementation can be achieved when the readout board is embedded in a PC server. In fact, the cost of the event-builder, to connect the readout-boards to the filter-farm nodes, is minimised by using data-centre
technology in the network and ensuring short distances between components. Data-centre technologies in the network require the use of PCs as end-points. The collaboration decided therefore to adopt the PCI Express (PCIe) standard to connect the readout board to the PC’s motherboard as a peripheral. The scheme of the PCIe based readout system is shown in figure 2.

The PCIe40 board will present 48 bi-directional optical links for interfacing the FE electronics and one bidirectional optical link for interfacing the TFC. All the boards will be equipped with a large-size Arria10 ALTERA FPGA. The latter is interfaced to the CPU through two 8-lanes PCIe Generation 3 busses connected to a PLX PCIe switch to form a single 16-lanes bus. The maximum data transfer rate is fixed by the PCIe Generation 3 output to about 110 Gb/s, which corresponds to 24, fully loaded, input links, running the wide mode GBT protocol.

Since the event-building requires to bring the data from all readout-boards into a single CPU node, a local area network (LAN) is used for this. Several LAN technologies are or will be likely available, however, at the moment there are only two, which have a certain market-share and are known outside very specialised contexts: Ethernet (IEEE 802.3) and InfiniBand [4]. Ethernet exists today in 10 Gbit/s and 40 Gbit/s versions (10G and 40G) and FDR InfiniBand offers effectively about 50 Gbit/s. In both cases a variant with 100 Gbit/s speed will be available at the time of the upgrade, which will be cheaper and simply reduce the number of necessary links by a factor two.

The main steps of the event building in this context are the following. Event-fragments are pushed by the FPGA of the PCIe40 into the main-memory of the hosting PC. Data from several bunch-crossings are coalesced into a multi-event fragment packet (MEP) to reduce the message rate and ensure efficient link-usage. Event-building is done then grouping in a single PC all MEPs containing data from the same bunch-crossings. For each MEP one PC is elected to be the event-builder PC, while the other PCs will send it their MEP. The PCs will use the same links also to receive the MEPs, when they are themselves the elected event-builder. In this way the link is used in both directions and the number of ports in the high-speed event-building network is as large as the number of event-builder PCs.

Figure 2. The PCIe based readout system. The PCIe40 readout boards are directly connected to the event-builder PCs through 16-lanes PCIe edge-connector.
Table 1. Summary of the readout-system.

<table>
<thead>
<tr>
<th>Versatile Links for DAQ</th>
<th>8800</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean nominal total event-size</td>
<td>100 kB</td>
</tr>
<tr>
<td>PCIe40 boards for DAQ</td>
<td>PCIe40 boards for DAQ</td>
</tr>
<tr>
<td>Versatile Links / readout board (DAQ)</td>
<td>up to 48</td>
</tr>
<tr>
<td>Event builder-PCs</td>
<td>500</td>
</tr>
<tr>
<td>PCIe40 boards for ECS and TFC</td>
<td>80</td>
</tr>
<tr>
<td>Core switch ports (100 Gbit/s)</td>
<td>500</td>
</tr>
<tr>
<td>Event-filter nodes</td>
<td>up to 4000</td>
</tr>
<tr>
<td>Output rate</td>
<td>20 kHz</td>
</tr>
<tr>
<td>Nominal instantaneous output rate</td>
<td>2 GB/s</td>
</tr>
</tbody>
</table>

The TFC, together with the distribution of the critical timing signals and fast commands to the front-end, implements a central, robust flow-control mechanism, the so-called throttle. Back pressure in the readout-system, from the PCIe40 boards onwards, will eventually make one of the PCIe40 trigger the throttle and ensure that synchronously the in flux of new events is stopped until the back-pressure has stopped.

The size of the whole system is given in the table 1. The limitation of 4000 event-filter nodes for the software trigger EFF comes from the power, cooling and space constraints of the data-centre, assuming a node to need 1U rack-space and about 400W of power.

3 Test results

Modern FPGAs, like for instance ALTERA Stratix V and newer models (Arria10), offer several embedded PCIe Generation 3 hard IP blocks to implement the PCIe protocol. The Stratix V allows to set two 8-lanes hard IP blocks, which can be connected to a PLX PCIe switch to form a 16-lanes PCIe Generation 3 bus such to reach the theoretical transmission capability of 128 Gbit/s. The PCIe hard IP blocks available are very efficient: one 8-lanes block uses less than 1% of the FPGA resources. We used evaluation boards equipped with Stratix V\textsuperscript{1} to test the performance of a DMA based data transfer from the FPGA to the RAM of the host Linux PC through PCIe. The measured bandwidth of about 55 Gb/s does not depend on the size of transmitted record, which has been varied in the range between 1 kB and 2 MB. The observed performance are in agreement with the literature \cite{5}: by specification the duty cycle of the DMA engine is 88% and the 8-lanes PCIe Generation 3 connection provides a theoretical bandwidth of 64 Gb/s.

The maximum bandwidth of the PCIe40 board is fixed by the 16-lanes PCIe Generation 3 protocol to about 110 Gb/s. The load on the single event-builder server is therefore quite high. The figure 3 shows the data-flow in one event-builder server. In total is at the level of 200 Gb/s full-duplex, when there is no data-reduction before events are sent to the EFF farm-nodes. Such a system has become possible since the advent of the Intel’s SandyBridge micro-architecture, which is the first CPU that handles the PCIe Generation 3 protocol.

\textsuperscript{1}ALTERA development board, Stratix V GX FPGA, model 5SGXEA7K2F40C2N and Bittware S5PE, Stratix V GX FPGA, model 5SGXEA7K2F40C3.
Figure 3. Data flow in the event-builder server

We have built a realistic test-system to measure performance, stability and resource-usage of a server.\(^1\) The other event-builder servers and the farm-nodes have been emulated by four different servers. The amount of transferred data from one server is the same as it will be in the final system. Since a PCIe40 was not available, it has been emulated using a GPU card from Nvidia. This generator produces the same data-pattern as the FPGA firmware, with all associated protocol overheads. It is 100% compatible with the FPGA version and it can send data over 16-lanes PCIe-3 at maximum speed. The prototype event-builder uses InfiniBand FDR dual-port cards with 16-lanes PCIe-3. This allows event-building and the sending of completed events at 100 Gb/s over two bundled ports of 54 Gb/s for each connection. The figure 4 shows a long-term test of the event-building performance. About 90% of the theoretical maximum link-speed (about 104 Gb/s) is achieved and the server is sustaining four times this I/O as required. The event-building is zero-copy, in the sense that the only copy operation is the one from the DMA engine (either of the GPU-FPGA or the network interface card) into the memory of the receiving PC. As can be expected from a purely zero-copy the CPU-load is rather modest: at about 400 Gb/s more than 80% of the CPU resources are free. In the current architecture the memory pressure is more important than the CPU. To evaluate this effect, we have run, in parallel to the event-building, the LHCb trigger application in offline mode (data come from a file in parallel to the event-building). On the test-machine we can launch 18 instances of the trigger application without negatively influencing the event-building application as seen in figure 4. The limitation does not come from the CPU needs of the event-building, which is rather small (about 15%), but from the total available memory-bandwidth in the server. The available memory bandwidth will increase in future server architectures,\(^3\) while the bandwidth-

\(^2\)The CPUs used in the test are Intel E5-2670 v2 with a C610 chipset. The servers are equipped with 1866 MHz DDR3 memory in optimal configuration. Hyper-threading has been enabled.

\(^3\)In fact it will already go up by almost 50% in the generation following the one on which the present tests have been performed.
needs of the event-builder remain constant at 200 Gb/s per PCIe40 card. Very conservatively we estimate that at least 80% of the event-building server will be available for opportunistic use by the high-level trigger or a software version of the low-level trigger. For an acquisition in 2019 we estimated a growth-factor with respect to the reference node of about 16. We expect to be able to run about 400 instances of the trigger application on a single EFF server. Therefore, the CPU time budget for each trigger application is 13 ms assuming an EFF farm of 1000 servers and an input rate of 30 MHz [6].

4 Conclusions

LHCb will undergo one major upgrade between 2018 and 2019 to allow operating the experiment at a luminosity of \(2 \times 10^{33} \text{cm}^{-2}\text{s}^{-1}\). The detector upgrade consists of a complete redesign of the data acquisition system to read out the full detector at the bunch crossing rate of 40 MHz. In order to maximise trigger efficiencies and to minimise systematic uncertainties in the selection of the interesting flavour decays, events will be processed at the full speed entirely by the versatile software trigger. The optimal implementation of the event-builder can be achieved interfacing the event-builder PC farm directly to the detector front-end electronics by means of PCIe Generation 3 based readout boards. The upgraded readout and trigger system, as well as the detector optimisation, will allow the LHCb physics program and running conditions to adapt to any signature which may come out. The plan is to collect a dataset of at least 50 fb\(^{-1}\) in less than ten years. It will allow LHCb to reach unprecedented precision in flavour physics.

References

