IL NUOVO CIMENTO **40 C** (2017) 71
DOI 10.1393/ncc/i2017-17071-0

COLLOQUIA: IFAE 2016

The 40 MHz trigger-less DAQ for the LHCb upgrade

A. FALABELLA([1]), M. MANZALI([2]) and U. MARCONI([3])

([1]) *INFN CNAF - Bologna, Italy*
([2]) *Università degli studi di Ferrara - Ferrara, Italy*
([3]) *INFN, Sezione di Bologna - Bologna, Italy*

received 17 October 2016

Summary. — The LHCb experiment focuses on flavour physics, aiming to enhance the current knowledge of CP violation parameters and exploit new physics signatures studying rare decays of b and c hadrons. A major upgrade of the detector is foreseen during the second long shutdown (2018–2019) to allow to collect an order of magnitude more data with respect to Run 1 and Run 2. The current maximum readout rate of 1 MHz is a limitation for the hadronic trigger. The upgraded detector will implement a full read-out running at the LHC bunch crossing frequency, using a software trigger. A high-throughput interface board has been designed to read-out the detector at 40 MHz. The read-out boards allow a cost-effective implementation of the DAQ by means of a high-speed PC network. The redesigned DAQ system collects data fragments from the subdetector, performs the event building, and transports data to the High-Level software trigger at an estimated aggregate rate of ~ 32 Tbit/s. Possible technologies candidates for the high-speed network under study are InfiniBand and Gigabit Ethernet. In order the explore and find the best implementation we performed several tests using an Event Builder evaluator on small size test beds and HPC scale facilities. Up to date performance results are presented.

1. – Introduction

LHCb [1] is one of the four main experiments at the Large Hadron Collider at CERN, Switzerland. It aims at the precise measurement of CP violation parameters and the study of rare decays of b- and c-quark hadrons. Figure 1 shows a schematic side view of the detector.

LHCb has already performed successful physics measurements during Run 1. After two years of stop, it has just started the Run 2 data-taking period (2015–2017). A major upgrade is foreseen during Long Shutdown 2 (LS2 2018–2019) in order allow record data at the design LHC energy of 14 TeV with an instantaneous luminosity of $2\times10^{33}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. The new read-out system will allow for a *full software trigger* [2].

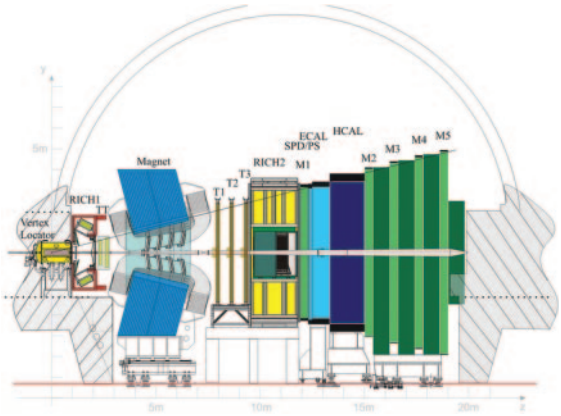

Fig. 1. – Schematic side view of the LHCb detector.

2. – Trigger evolution

The Run 1 trigger system consists of a Level-0 (L0) fixed latency near-detector trigger and a software Higher-Level trigger (HLT). The L0 trigger purpose is to reduce the visible bunch crossing rate from 30 MHz to $\sim$ 1 MHz at which frequency the detector could be read-out. The HLT then applies more advanced selections in order to further reduce the rate to $\sim$ 5 kHz for offline storage and reprocessing.

In Run 3 the upgraded LHCb detector will operate at luminosity five times higher with respect to Run 1. The limited information used by the L0 trigger would causes a loss of efficiency especially for hadronic channels as shown in fig. 2.

3. – DAQ implementation for the upgrade

The LHCb upgrade will then remove the L0 trigger implementing a trigger-less readout system [3]. The readout system will be composed by the Event Builder (EB), the

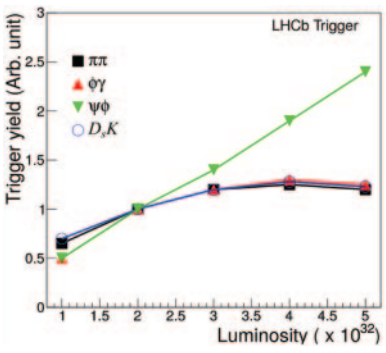

Fig. 2. – L0 trigger efficiency normalised to that of Run 1 as a function of luminosity for selected hadronic decays. At the nominal Run 3 luminosity of $2 \times 10^{33}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ several hadronic modes saturate.

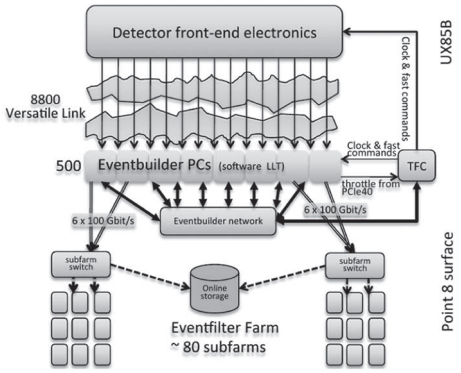

Fig. 3. – The architecture of the upgraded LHCb readout system.

TABLE I. – *Constraints for the online system.*

Event rate	40 MHz
Mean nominal event size	100 KBytes
Readout board bandwidth	up to 100 Gbit/s

Timing and Fast Control (TFC) distribution, the Experiment Control System (ECS) and the Event Filter Farm (EFF) and it can be seen from the schema of fig. 3.

Table I summarizes the constraints for the trigger-less readout system. It can be seen that the EB network must than handle an aggregated bandwidth of 32 Tbit/s. Such a network can be implemented at a reasonable cost using commercial local area network technologies such as Ethernet or InfiniBand. Here we present up to date scalability tests for InfiniBand-based EB network. We considered the InfiniBand solution because of its constant performance improvement and cost effectiveness.

4. – EB Performance Evaluator

In order to test the DAQ implementation we developed a performance evaluator software for the EB. With respect to our previous work [4] we decided to simplify the implementation of the software keeping the Builder Units (BU) and the Readout Units (RU) while discarding the Event Manager.

5. – EB nodes tuning

In order to obtain the best performances from the InfiniBand network adapters we follow the prescriptions in [5]. In particular a key aspects that must be considered are the CPU power management that can be controlled in Linux in several ways. Another aspect concerns the NUMA (*Non Uniform Memory Access*) architecture in multiprocessors systems. In such systems the I/O is better for processes running on CPU local to the network interface.

We measured the network traffic through the BU during a simulation of a few minutes using built-in test provided by the Open Fabrics Alliance (OFED) driver suite. The results are shown in fig. 4. The blue curve represents the bandwidth with the power

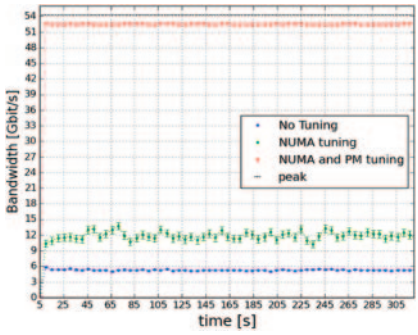

Fig. 4. – EB software test on point-to-point two machines test bed equipped with Mellanox MCB194A-FCAT 56 Gbit/s (FDR). The blue curve represents the bandwidth as seen by the BUs without any tuning. The green curve represent the same test when binding to the proper NUMA node, the red curve has been obtained disabling also the power management.

management active and without binding the running processes to a particular NUMA node. Considering that Mellanox FDR allows for 54.3 Gbit/s (taking into account of encoding) data transfer the performance is around $\sim 10\%$ of the maximum value. Binding to the proper NUMA node gives an average value of ~ 11.8 Gbit/s (green curve), while disabling the power management gives and average value of ~ 52.5 Gbit/s (red curve) that is the $\sim 98\%$ of the maximum. This means that the EB software is able to saturate the network prior to a proper setup of the system. The bandwidth is also stable over time.

6. – Scalability tests

A key test for the EB software is the scalability of its performances on large systems. We performed such test on the Galileo cluster of the Cineca consortium.

Table II summarizes the Galileo system architecture. We run the EB software on a increasing number of nodes running an a RU and a BU on each node. In each test we measure the full-duplex bandwidth. In fig. 5 the results for a 128 node test are shown (maximum number of nodes allowed by the batch scheduler). As can be seen the bandwidth is stable over time and $\sim 60\%$ of maximum full-duplex. This shows that the EB prototype performs good in term of scalability and stability.

7. – Conclusion

The LHCb DAQ system has been redesigned in order to cope with LHC Run 3 higher luminosity. One of the elements of the new implementation is the EB that requires a

TABLE II. – *Galileo cluster architecture.*

Nodes	516
Processors	8-cores Intel Haswell 2.40 GHz (2 per node)
Network	Infiniband with 4x QDR switches

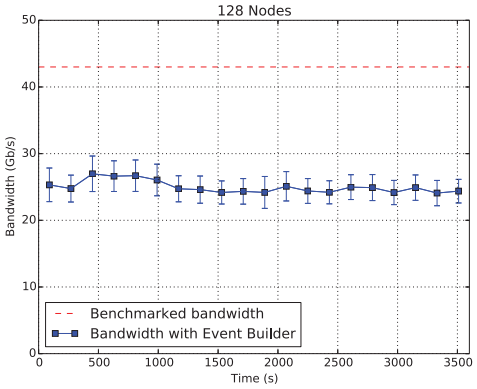

Fig. 5. – Data transfer full-duplex for 128 nodes test of the EB software.

high-throughput network that must handle an aggregated traffic of 32 Gbit/s. The InfiniBand standard is a good candidate for it, and to prove it we implemented a prototype software. We proved the scalability of the EB up to 128 nodes at the Galileo CINECA cluster.

* * *

The authors thank the HPC User Support team at CINECA for their prompt support during the tests.

REFERENCES

[1] LHCb Collaboration, *JINST*, **3** (2008) S08005.
[2] LHCb Collaboration, Trigger and Online Upgrade Technical Design Report *CERN*, **LHCC** (2014) C10026 LHCb.
[3] LHCb Collaboration, *JINST*, **9** (2014) C10026.
[4] Antonio Falabella, *Nucl. Instrum. Methods Phys. Res. A*, (2016) 280.
[5] Mellanox Technologies, Performance Tuning Guidelines for Mellanox Network Adapters, `www.mellanox.com` (2014).