time structure analysis of the lhcb daq network

Gianni Antichi, Marc Bruyère, Daniel Campora Perez, Guoming Liu, Niko Neufeld, Philippe Owezarski, Andrew W Moore, Stefano Giordano

Time structure analysis of the LHCb DAQ network

G Antichi[1], M Bruyere[2,3,4], D H Cámpora Pérez[6], G Liu[6], N Neufeld[6], S Giordano[5], P Owezarski[2,3] and A W Moore[1]

[1] University of Cambridge
[2] CNRS
[3] Université de Toulouse
[4] DELL Inc.
[5] University of Pisa
[6] CERN

E-mail: gianni.antichi@cl.cam.ac.uk, mbruyere@laas.fr, dcampora@cern.ch

Abstract.

The LHCb DAQ Network is a real time high performance network, in which 350 data sources send data over a Gigabit Ethernet LAN to more than 1500 receiving nodes. The aggregated throughput of the application, called Event Building, is more than 60 Gbps. The protocol employed by LHCb makes the sending nodes transmit simultaneously portions of events to one receiving node at a time, which is selected using a credit-token scheme. The resulting traffic is very bursty and sensitive to irregularities in the temporal distribution of packet-bursts to the same destination or region of the network.

In order to study the relevant properties of such a dataflow, a non-disruptive monitoring setup based on a networking capable FPGA (Netfpga) has been deployed. The Netfpga allows order of hundred nano-second precise time-stamping of packets. We study in detail the timing structure of the Event Building communication, and we identify potential effects of micro-bursts like buffer packet drops or jitter.

1. Introduction

In the LHCb experiment [1], particle collisions produce data at a rate of 40 MHz. This data is gathered by FPGA-based readout unit boards, known as *TELL1s* [2], which apply a filter based on selection algorithms. This *hardware level trigger*, discards uninteresting events and reduces the data rate to 1.1 MHz. The TELL1 boards pack the data and inject it into the Data Acquisition (DAQ) network, where the data is subsequently processed in software, in stages known as *Event Building* and *High Level Trigger*.

The DAQ network [3], depicted in figure 1, is a routed network composed of 350 TELL1s acting as sources, and more than 1500 receiving nodes. The network is built following a fat-tree topology, with two core routers and 65 Top of the Rack (TOR) switches, connected to each other by Link Aggregation (LAG) uplinks for an aggregated bandwidth of 6 Gbps for each TOR switch. All sending nodes have four 1 GbE links, two of them connected to each core router to achieve traffic load-balancing, and each receiving node is connected to a TOR switch by a 1 GbE link.

The traffic flows unidirectionally through the network following a many-to-one pattern, from the TELL1s to the farm nodes, using a custom format named *Multi-Event Packets* (MEP). MEP

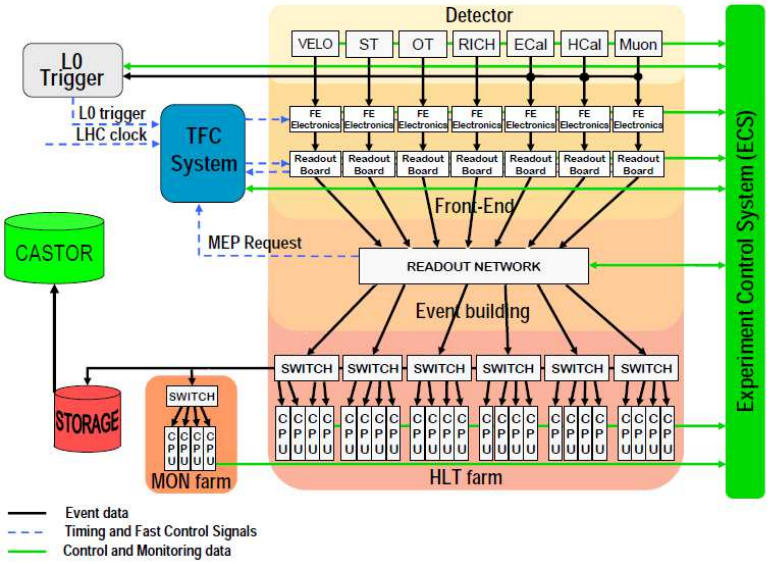


Figure 1: The LHCb readout-system.

datagrams are sent in a connectionless communication scheme, sensitive to losses. This format allows the transmission of several event fragments packed in each datagram, and in this fashion up to 15 event fragments can be sent within each packet. The typical aggregated packet rate of the network is 25 MHz.

All TELL1s are highly synchronized in time by the use of the *Timing and Fast Control* (TFC) signal, to a sub-microsecond level. All sources send MEP datagrams to receiving farm nodes, one at a time, which are selected using a credit-token scheme. As a consequence, the resulting traffic shape is bursty locally on the receiving nodes, and deep-buffers are required in the network.

In this paper, we present studies about the shape and characterization of the traffic in the LHCb DAQ network. We show the time distribution of packet bursts and latency irregularities in a section of the network, and we outline the effects of such a traffic.

2. Monitoring setup

We have performed our measurements during a data taking period of the LHCb experiment. In order not to be disruptive with the performance of the DAQ network, we have identified a portion of the network which would be of interest, and replicated the data using network TAPs. Figure 2 depicts the monitoring setup.

The presented setup has been created to account for the monitoring of data flows. The detailed monitoring setup for each port is as follows:

- *nf2c0* - Data incoming from TFCODIN-TELL1 (*TFCODIN*), with any destination.
- *nf2c1* - Data incoming from ECAL-TELL1 (*ECAL*), with any destination.
- *nf2c2* - Any source to a destination in the *A01* farm of PCs (*A01*).
- *nf2c3* - Credit requests (*MEP Requests*) coming from all subfarms.

After the flow has been replicated, the *Netfpga platform* is used on the monitoring side for high-performance timestamping.

Netfpga is a low-cost platform, developed by the High Performance Networking Group at Stanford University. The card contains a Xilinx Virtex-II Pro Field-Programmable Gate Array

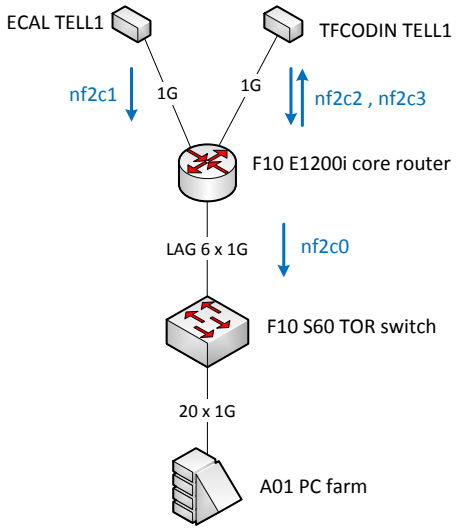


Figure 2: Monitoring setup.

(FPGA) which is programmed with user-defined logic and a clock of 125 MHz. The PCI interface connecting the host PC to the Netfpga is managed by a small Xilinx Spartan II FPGA. Four Gigabit Ethernet ports, 4.5 MB of static RAM (in 2 banks) and 64MB of DDR2 dynamic RAM are also on-board.

In our system, the Netfpga takes care of filtering and timestamping traffic of interest, while existing libpcap-based software provides a familiar user interface, and flow selection is accomplished by commands to a simple, shell-like Command Line Interface. Only traffic that matches a filter rule is passed to host software, with an associated timestamp. Packets are timestamped at the very earliest possible moment to minimise jitter caused by FIFOs. Since the Ethernet preamble can be of variable length, we sample the timestamp counter when the Ethernet start-of-frame delimiter arrives to minimise jitter.

3. Time analysis results

We have carried out an analysis of the latencies of the packets received, related to our traffic type.

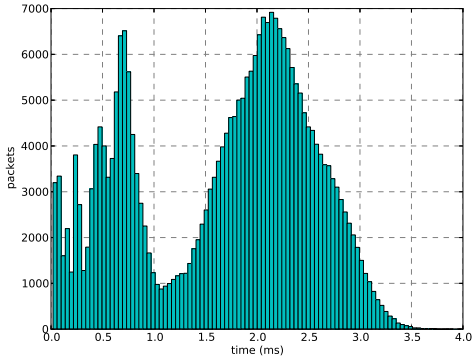
The router under analysis is a *Force10 E1200i* [5], a terascale 1/10 GbE core router with a raw 3.5 Tbps switching capacity. In order to perform background latency tests on the device, a Local Area Network composed of six nodes, synchronized with the peer-to-peer time protocol [6] has been deployed. At the moment of the test, the switch had no packet load.

We measure a consistent latency of 50 microseconds on packet switching, for varying packet sizes. The latency has been measured both on the same linecard and across linecards with similar results, verifying the non-locality feature of the switch.

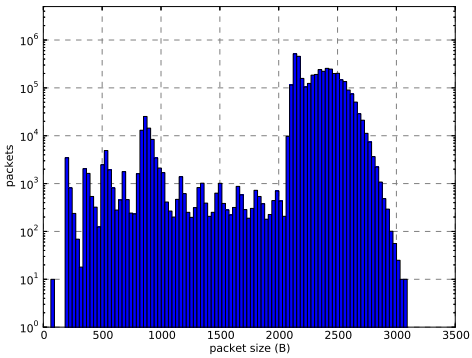
Using the Netfpga setup, we have monitored a data-taking period of the LHC to perform traffic shape analysis. Figure 3 depicts a latency distribution of the traffic flowing from *ECAL* to *A01*, over proton-proton data-taking runs in January and February, 2013. A distribution between 50 microseconds and 4.0 milliseconds is observed, with a non-uniform density.

The difference between the maximum and minimum latency are indicators of the utilization of the buffers. The maximum difference is of 4.0 milliseconds, or 512 KB for an egress link with 1 Gbps capacity.

This has implications as to what requirements the current LHCb DAQ network has, related



(a) Latency histogram.



(b) Histogram of packet sizes on logarithmic scale.

Figure 3: Characterization of *ECAL - A01* flow, in proton-proton collision runs.

to the architecture of the switch. In our current setup, our aggregated bandwidth to the aggregation layer is of 360 Gbps. With a shared load, a naive extrapolation implies a 155 MB egress buffer utilization, and since we have around 700 ingress ports in our setup, an average 250 KB port buffer requirement. These static buffer requirements have to be met by the switch architecture, a consideration which is valid and extrapolable for the upcoming network upgrade in Long Shutdown 2 [4].

We have identified the density distribution to be related to our packet size distribution, depicted in figure 3b. A similar shape to that of figure 3a is observed, suggesting there is a connection between the packet length and its latency.

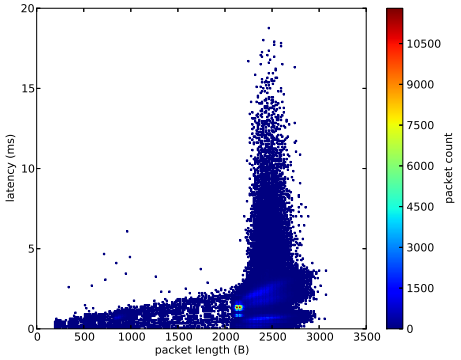
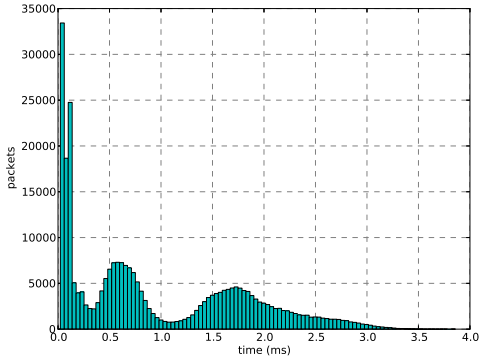


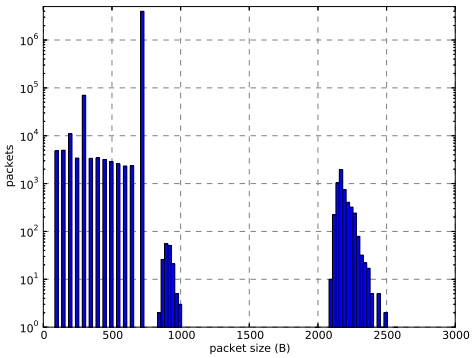
Figure 4: Characterization of length of packets in flow *ECAL - A01*, in proton-proton collision runs. Latency of packets versus length.

In our network, all packets are sent off sender cards in a highly synchronized way in bursts, over a 25 ns window-frame. All these packets do not have the same length, and so smaller packets are processed faster due to the store-and-forward switching policy of the E1200i. As a consequence, there is a direct relation between the frame size and the latency. Figure 4 shows an increase in the latency as a function of the packet size.

The flow from TFCODIN to A01 shows a different latency distribution, as shown in figure 5a. Three peaks are observed, and the maximum latency difference is similar to the previously observed.



(a) Latency histogram.



(b) Histogram of packet sizes on logarithmic scale.

Figure 5: Characterization of *TFCODIN - A01* flow, in proton-proton collision runs.

A comparison with the packet length distribution yields a similar shape to the observed in the latency distribution. As figure 5b shows, the packet length histogram keeps similarities with figure 5a. The length versus latency depicted in figure 6 shows an increasing function with a peak at the length value 708. Packets with this length value represent 97% of the total.

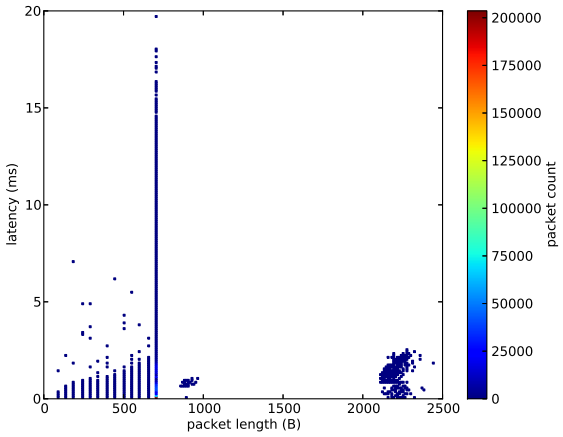


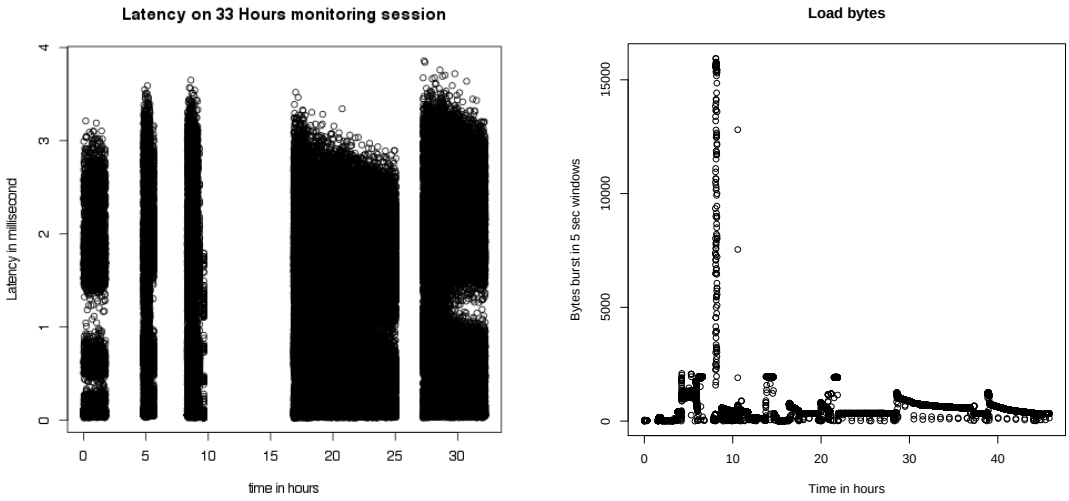
Figure 6: Characterization of length of packets in flow *TFCODIN - A01*, in proton-proton collision runs.

The load of MEP Requests is an accurate indirect indicator of the load of the router. The latency maxima observed in the latency distribution over time agree with the router load peaks. These effects can be observed in a side-to-side comparison between figure 7a and 7b.

The traffic shape inside a run, depicted in figure 7a, is related to the decreasing collision rate during the same time period, in figure 8. A decreasing peak latency over time is observed.

4. Conclusion

We have characterized the data traffic in the Online DAQ network of LHCb. Maximum latency values give a view of the requirements in buffer size in our network. We have found correlations in the traffic shape of the network with the collision rate of the LHC, and we characterize the



(a) Latency density versus time.

(b) Burst size versus time.

Figure 7: Burst and latency characterization over time.

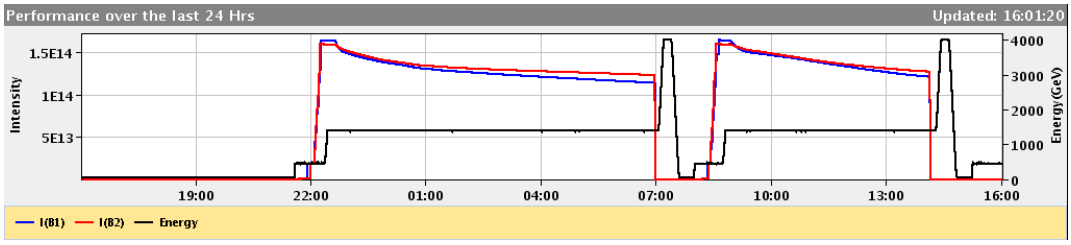


Figure 8: Intensity of beams in the LHC over time. Two proton-proton runs are shown, in a 24-hour period starting on the 12-02-2013 at 16:00.

latency as a function of the size of the packets, an effect seen with the store-and-forward policy of our routers.

Our Netfpga monitoring setup accounts for precise network shaping. Bursts effects are observed, which may be disruptive at higher data frequencies. Drops or backpressure effects originating from variable latencies can be observed with our monitoring system, which will play a decisive role at later stages of the LHC.

[1] The LHCb Detector at the LHC. Alves, A. Augusto et al. 10.1088/1748-0221/3/08/S08005

[2] The LHCb DAQ interface board TELL1. Haefeli, G. et al. 10.1016/j.nima.2005.12.212

[3] The LHCb Eventbuilder: Design, Implementation and Operational Experience. Neufeld, N. and Frank, M. and Garnier, J.-C. and Gaspar, C. and Jacobsson, R. and Jost, B. and Guoming Liu

[4] Letter of Intent for the LHCb Upgrade. LHCb Collaboration ; CERN, Geneva. CERN-LHCC-2011-001. LHCC-I-018

[5] Dell Networking E-Series. High-performance 1/10GbE chassis core systems. DELL Inc.

[6] IEEE 1588 Standard for A Precision Clock Synchronization Protocol for Networked Measurement and Control Systems. IEEE Instrumentation and Measurement Society.