Feasibility tests of RoCE v2 for LHCb event building

Rafał Dominik Krawczyk[1,*], *Tommaso* Colombo[1], *Niko* Neufeld[1], *Flavio* Pisani[1,2], and *Sébastien* Valat[3]

[1]CERN 1211 Geneva 23, Switzerland
[2]University of Bologna, Via Zamboni, 33, 40126 Bologna BO, Italy
[3]Atos SE, 1 Rue de Provence, 38130 Échirolles, France

Abstract. This paper evaluates the utilization of Remote Direct Memory Access (RDMA) over Converged Ethernet (RoCE) for the Run 3 LHCb event building at CERN. The acquisition system of the detector will collect partial data from approximately 1000 separate detector streams. The total estimated throughput equals 32 Terabits per second. Full events will be assembled for subsequent processing and data selection in the filtering farm of the online trigger. High-throughput transmissions with up to 90% links utilization will be an essential feature of the system. The data exchange mechanism must support zero-copy transmissions. In this work, the RoCE high-throughput kernel bypass Ethernet protocol is benchmarked as a potential alternative to InfiniBand. A RoCE-based event building network is presented and two implementations are considered. The former variant combined shallow-buffered and deep-buffered switches with enabled flow control. In the latter setup, only deep-buffered devices are used, where operation relied on their memory throughput and capacity. Feasibility tests were conducted with selected Ethernet switches. Memory bandwidth utilization was investigated, in comparison with InfiniBand. Relevant utilization and interoperability issues of RoCE flow control are detailed with lessons learned along the road.

1 Introduction

CERN LHCb detector upgrade for upcoming Run 3 [1] imposed significant changes in the data acquisition system [2]. In its new revision, the hardware trigger will be dropped and the data selection will be fully software-defined. The currently developed Software High-Level Trigger will have to handle 32 Terabits per second of input. Before being fed to the trigger filter farm, the data stream must be first reorganized in the event building process. Fragments of events from all detector parts must be assembled into one structure and then dispatched across the nodes handling the trigger data selection. A network must, therefore, handle lossless, many-to-one traffic. For the sake of cost optimization, close-to-maximal link utilization is expected, with of average of 80 Gbit/s per second and reaching 90 Gbit/s. With the currently considered 100 Gbit/s links, efficient transmissions with a small memory footprint can be reached with a Remote Direct Memory Access. This zero-copy protocol, as opposed to TCP and UDP, does not make a copy of data in the kernel invocation and

[*] e-mail: rafal.dominik.krawczyk@cern.ch

allows direct transfer between memory regions between remote nodes. This drastically lowers necessary memory bandwidth and CPU usage.

So far, InfiniBand has been the most promising and most widely tested RDMA-supporting technology for the LHCb Event building [3][4]. However, a protocol called “RDMA over Converged Ethernet version 2” (RoCE v2) [5] has recently become a promising and potentially applicable alternative.

2 RoCE for the LHCb Event Building

RoCE v2 is a zero-copy technology that supports sending InfiniBand frames encapsulated within the Ethernet frames, using the Ethernet infrastructure. It has several relevant features from the standpoint of implementing the LHCb Event Building.

2.1 Advantages and limitations of RoCE v2

RoCE v2 is an Ethernet-based protocol. Unlike InfiniBand, it is supported by multiple vendors, an important consideration given the long life-time of the Event Builder.

Another advantage of RoCE v2 over InfiniBand is the possibility of combining different link speeds in a single subnet. A cost reduction is possible because events assembly and dispatching can be fused within a single network.

However, RoCE is a less mature and tested protocol. The most important limitation of the protocol is that it is not running over a lossless lower layer (Ethernet is not loss-less). Whereas InfiniBand ascertains that data is sent only when the receiver has enough resources, drops can happen in the Ethernet networks during the transmission. If such losses occur, RoCE v2 performance drops sharply. These disadvantages enforce using additional mechanisms to prevent dropping packets, such as relying on the switches with large buffers with high throughput or the Ethernet flow control protocols.

2.2 Proposal of Ethernet RoCE v2 based Event Builder

Considering the advantages and the limitations of the RoCE v2 protocol, a new event building network has been designed, presented in figure 1. In such 25 Gbit/s and 100 Gbit/s links are combined in a single subnet. Event Building and dispatching of assembled structures are both conducted in this infrastructure. Two variants of switches can be distinguished. The core switches connect data source servers with Top-of-the-Rack (ToR) switches and use 100 Gbit/s links. For ToR switches 100 Gbit/s uplinks and 25 Gbit/s downlinks are used. ToRs are connected to both the core switches and the data consumer servers.

Two types of ToR switches are considered. In the cheaper variant, the shallow-buffer devices are used. This configuration must use additional Ethernet flow control. In the more expensive variant, deep-buffered devices can potentially ascertain lossless transmissions without using the Ethernet flow control. For this, the memory of the devices must have large-enough throughput to efficiently handle up to 90 Gigabits per second transmissions in all their ports at the same time.

3 Evaluating RoCE v2 setup

The tests were conducted to evaluate RoCE v2 stability and performance. The purpose was to assess whether the proposed, Ethernet-based event building design would be performant enough.

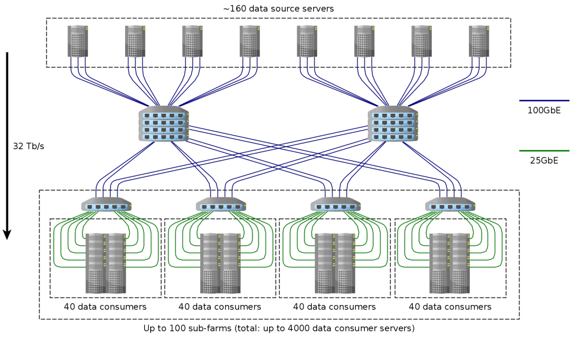

Figure 1. RoCE v2 Ethernet-based LHCb event builder.

3.1 Long-run RoCE v2 test and comparison with InfiniBand

Tests were made to compare Ethernet with RoCE v2 versus InfiniBand in terms of performance and stability. A custom event-building traffic simulator, DAQPIPE, was used [6]. 3 switches were selected:

- Mellanox MSB7700 - an InfiniBand reference device with 36 100 Gbit/s ports,
- Juniper QFX10000-30C - an Ethernet deep-buffered line card supporting RoCE v2. It contains 12 GB of memory and 32 100 Gbit/s ports, Hybrid Memory Cube (HMC) is used for the buffers [7].
- Arista DCS-7280CR2K-30-F a second, deep-buffered Ethernet switch with a RoCE v2 support. It has 12 GB of memory and 30 100 Gbit/s ports. DDR RAM is used for the buffers, which has lower band-width than HMC.

In the first test, it was assessed if the RoCE devices can constantly ascertain sufficient throughput over a longer period of time. All available 100Gbit/s ports were used on the selected devices. The DAQPIPE parameters were first tuned in 15-second runs. For each switch the optimal configuration was selected of message size, number of transmission grants per data consumer and number of parallel sends per data producer. Then the DAQPIPE was run for 3 hours and throughput of data producers was probed every two seconds. Results are depicted in figure 2. For Mellanox MSB7700 Infiniband switch, average throughput was 95.1 Gbit/s and the minimal value was 94.7 Gbit/s. For Juniper QFX10000-30C, the average throughput was 94.8 Gbit/s, the minimal probe was 87.1 Gbit/s. Fluctuations were relatively small (although higher than in the case of Mellanox MSB7700 InfiniBand device). However, for Arista DCS-7280CR2K-30-F, both the performance and the stability were worse. Average throughput was 87.4 Gbit/s and it dropped to 66.7 Gbit/s. Additionally, these results were achieved only when the message size was limited to 128 KiB.

Stable performance was reached for a single, fully populated Juniper QFX10000-30C line card. Conversely, transmissions with Arista switch were not performant enough. Further tests were conducted to check if stable RoCE v2 based event building performance can be achieved while relying on buffers' throughput with the selected deep-buffered devices.

3.2 Evaluating switches throughput for the deep-buffered only variant

Following the long-run test, a parameter scan was made with the DAQPIPE to compare the behaviour of the fully populated Arista DCS-7280CR2K-30-F switch and the Juniper QFX10000-30C line card. Message size, number of parallel sends and data consumer send requests were adjusted for 15-second benchmark runs. As depicted in figure 3, above 80 Gbit/s

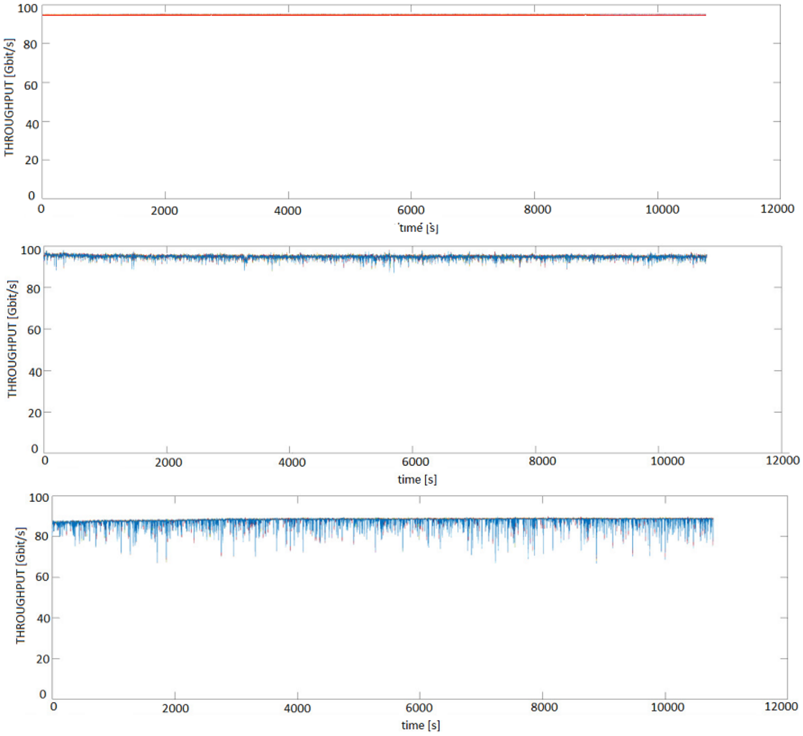

Figure 2. Mellanox InfiniBand (top), Juniper QFX10000-30C RoCE (middle) and Arista DCS-7280CR2K-30-F RoCE results from Event Builder simulator 2-hour runs.

were reached only for the message size of 128 KiB. Another DAQPIPE test was made to investigate this limitation. The message size was set to 2 MB and the number of ports was adjusted. Data consumers and data producers were separated and assigned to ports serviced by different on-switch ASICs. As indicated in figure 4, a notable performance drop occurred when more than 6 ports were used. Knowing that both the switches had 12 GB of memory, it was clear that the bandwidth of DDR RAM was the limitation of the Arista DCS-7280CR2K-30-F switch. The on-ASIC memory throughput was too low to buffer all the upcoming data in parallel from all ports at a rate of above 90 Gbit/s. The device was not performant enough for the deep-buffer only event builder and discarded due to its limitations.

Further tests were continued with Juniper QFX10000-30C devices. DAQPIPE was run with 88 ports used on 3 line cards installed in Juniper QFX100008 switch. In such a configuration, data was partially transmitted between fabrics connecting the line cards. As depicted in figure 5, the performance was slower in comparison with figure 3, where fabrics connecting the line cards were unused. Data producers throughput was below the bottom limit of 80 Gbit/s. Although performant-enough for a single line card, event building was underperforming when several line cards had to partially tranmit data between with one another. Performance was likely limited by the scheduling of traffic in the fabrics connecting the cards.

In conclusion, the throughput of the two selected switches was insufficient for the deep-buffer-only configuration for the event builder proposed in figure 1. The following step was, therefore, testing Ethernet Flow Control for the alternative configuration combining deep-buffered and shallow-buffered switches.

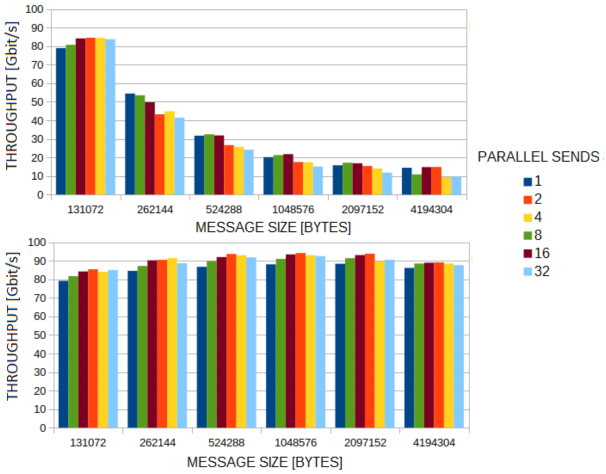

Figure 3. Arista DCS-7280CR2K-30-F switch (top) versus Juniper QFX10000-30C (bottom) results with average data producers throughput of the DAQPIPE parameter scan for 8 parallel transmission grants per data consumer. Number of parallel sends per data producer and message size were adjusted.

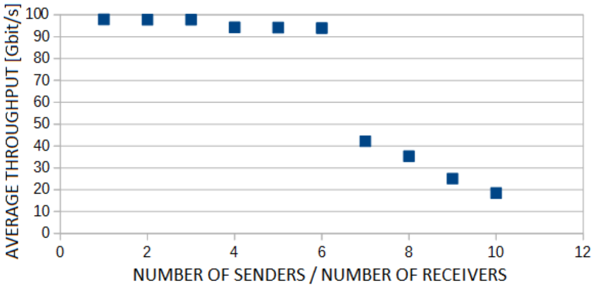

Figure 4. Data producers throughput for Arista switch with DAQPIPE versus the number of used ports.

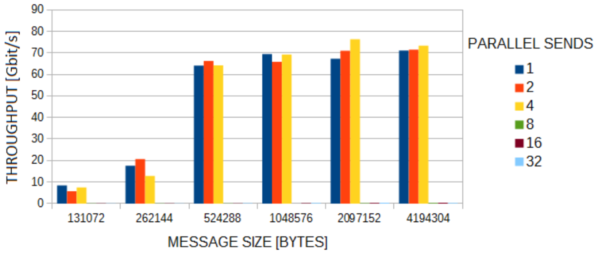

Figure 5. Average data producers throughput for DAQPIPE with 88 ports used on three Juniper QFX10000-30C line cards

3.3 Flow Control tests for the shallow-buffered and deep-buffered variant

The second variant of Ethernet-only event builder assumes using shallow-buffered switches as ToR (see figure 1). In this configuration packet losses in many many-to-one transmissions cannot be avoided by buffering and Ethernet flow control mechanisms must be used.

At the time of writing this paper, 2 such mechanisms were supporting RoCE v2. The link-level Priority Flow Control (PFC) [8] and point-to-point Explicit Congestion Notification (ECN) [9]. Although Packet Pacing [10] would allow explicit throughput control between data producer and data consumer in the event building, this mechanism was not yet supported.

When configuring PFC for combined Juniper QFX10000-30C deep-buffer line card and QFX5200 shallow buffer switch, interoperability issues were witnessed. The following test bench was evaluated under network congestion conditions. A sender node was connected to the switch with 100 Gbit/s link. It was was transmitting to a receiver node connected to the switch with 25 Gbit/s link. When a single Juniper QFX1000-30C and QFX5200 devices were used, PFC worked correctly. However, for a combination of QFX1000-30C and QFX5200 switches connected with 100 Gbit/s, the PFC was not working and the performance dropped. As further investigated, the Juniper QFX1000-30C line card, was receiving Ethernet Pause frames from the Juniper QFX5200 device but it did not correctly react to them and did not reduce throughput in the link between the two switches. As further confirmed by the Juniper support, PFC was not supported in such configuration.

The only mechanism working correctly for the setup with the two switches was the ECN. A test was made to evaluate its impact on performance with Juniper QFX5200 shallow-buffered switch. 8 data producers were connected to the switch with 100 Gbit/s links. 30 data consumer nodes were connected via 25 Gbit/s links using breakout cables. A custom benchmark was developed [11] to test performance of 100-to-25 Gbit/s transmissions. The results are depicted in figure 6, presenting throughput of data consumers with and without ECN. Only a small performance improvement was observed, which needed a lot of ECN parameters tuning in the switch. Moreover, the throughput was above 15 Gbit/s only for small messages sizes, up to 256 KiB. At this stage of evaluation, it was, therefore, concluded that flow control mechanisms were insufficient to ascertain stable and performant-enough event building.

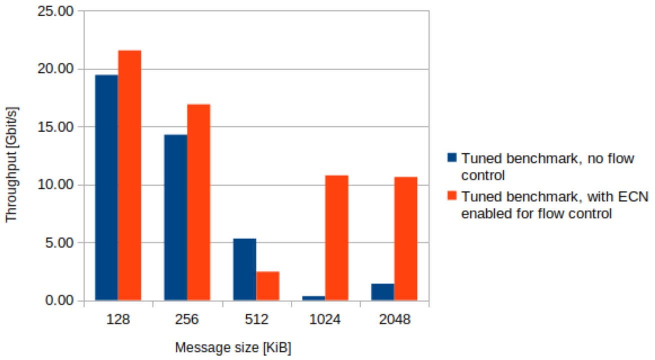

Figure 6. Performance improvement for the the data consumers for the benchmark generating 100-to-25 Gbit/s traffic and Juniper QFX5200 switch

4 Conclusions

The tests of the Ethernet RoCE v2 devices have indicated that, at this stage of development, the standard is not as performant as InfiniBand. The throughput and stability are insufficient to ascertain the relevant operation of the LHCb DAQ upgrade. Therefore, it has been decided to implement the event building with InfiniBand. The Ethernet RoCE is still a considerable option for the dispatch of assembled events. However the tests also showed that if these challenges can be overcome Ethernet is a very realistic alternative for the future evolution of the LHCb Data Acquisition.

References

[1] A. Piucci, *The LHCb Upgrade*, Journal of Physics: Conference Series vol. 878 (2017), https://iopscience.iop.org/article/10.1088/1742-6596/878/1/012012/pdf

[2] LHCb collaboration, *LHCb Trigger and Online Technical Design Report*, CERN-LHCC-2014-016; https://cds.cern.ch/record/1701361/files/LHCB-TDR-016.pdf

[3] T. Colombo, et al., *The LHCb DAQ Upgrade for LHC Run3*, IEEE Transactions on Nuclear Science vol. 66 (7), (2019), https://ieeexplore.ieee.org/document/8727952

[4] S. Valat, et al., *An Evaluation of 100-Gb/s LAN Networks for the LHCb DAQ grade*, IEEE Transactions on Nuclear Science vol. 64 (6), (2017), https://ieeexplore.ieee.org/document/7886309

[5] Mellanox official web page, *RoCE v2 Considerations*, https://community.mellanox.com/s/article/roce-v2-considerations

[6] S. Valat, et al., DAQPIPE benchmark repository page, https://gitlab.cern.ch/lhcb-daqpipe/lhcb-daqpipe-v2

[7] Juniper networks Documentation, *QFX10000 Switches System Architecture*, https://www.juniper.net/assets/us/en/local/pdf/whitepapers/2000599-en.pdf

[8] Juniper networks TechLibrary, *Configuring CoS PFC (Congestion Notification Profiles)*, https://www.juniper.net/documentation/en_US/junos/topics/task/configuration/cos-congestion-notification-qfx-series-cli.html

[9] Juniper networks TechLibrary, *Understanding CoS Explicit Congestion Notification* https://www.juniper.net/documentation/en_US/junos/topics/concept/cos-qfx-series-explicit-congestion-notification-understanding.html

[10] Mellanox Technologies documentation, *HowTo Configure Packet Pacing on ConnectX-4* https://community.mellanox.com/s/article/howto-configure-packet-pacing-on-connectx-4

[11] R. Krawczyk, et al., Distributed Event Building Benchmark repository page, https://gitlab.cern.ch/lhcb-daqpipe/distributed_network_benchmark