Ethernet evaluation in data distribution traffic for the LHCb filtering farm at CERN

Rafał Dominik Krawczyk[1,*], *Flavio* Pisani[1], *Tommaso* Colombo[1], *Markus* Frank[1], and *Niko* Neufeld[1]

[1]CERN 1211 Geneva 23, Switzerland

Abstract. This paper evaluates the real-time distribution of data over Ethernet for the upgraded LHCb data acquisition cluster at CERN. The system commissioning ends in 2021 and its total estimated input throughput is 32 Terabits per second. After the events are assembled, they must be distributed for further data selection to the filtering farm of the online trigger. High-throughput and very low overhead transmissions will be an essential feature of such a system. In this work RoCE (Remote Direct Memory Access over Converged Ethernet) high-throughput Ethernet protocol and Ethernet flow control algorithms have been used to implement lossless event distribution. To generate LHCb-like traffic, a custom benchmark has been implemented. It was used to stress-test the selected Ethernet networks and to check resilience to uneven workload distribution. Performance tests were made with selected evaluation clusters. 100 Gb/s and 25 Gb/s links were used. Performance results and overall evaluation of this Ethernet-based approach are discussed.

1 Introduction

The present CERN LHCb upgrade [1, 2] involves a significant increase in the throughput in the filtering farm [3–5]. The new version of the LHCb data acquisition cluster will implement a fully software-defined selection. After its commissioning in 2021, it will process uncompressed data streams of the FPGA front-ends at a total input throughput of 32 Terabits per second.

The first stage of the processing consists of all-to-all transmissions between servers. This process is called *the event building*. Its purpose is to assemble *events*, which are the detector responses to particle bunch collisions. Events' fragments are scattered across all of the input FPGA streams. This assembly stage has been already evaluated and is now under commissioning.

The following step after the event building is to distribute the assembled structures to the data selection nodes. These servers reduce the output throughput that is kept in the persistent storage. One of the tested scenarios was to use a network to dispatch workloads between the *data producers* (that is, the event building nodes doing assembly) and the *data consumers* (that is, the filtering nodes doing selection). This many-to-many high-throughput event distribution traffic must be lossless. Otherwise, relevant information from the experiment will be lost during the acquisition.

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

The network must also be capable of handling different numbers of producers and consumers. Their ratio will be ultimately defined by the amount of event building versus filtering farm workloads. To optimize costs, links must use close-to-maximal speed. The network with the dedicated software must distribute workloads to the filtering farm as evenly as possible. In the event of the temporary busyness of some of the data selection servers, the workloads should be redirected to non-busy nodes.

For this specific use-case, Ethernet was considered as a potential alternative to InfiniBand. It allows for combining different link speeds in a single network. Another advantage is cost reduction when using a cheap shallow-buffered switch. This assumption is possible if the Ethernet flow control protocols are used to avoid congestion. Two such mechanisms were possible at the time of the evaluation: Priority Flow Control (PFC) and Explicit Congestion Notification (ECN).

This paper specifically focuses on the network tests with Ethernet and traffic control. A custom C++ benchmark has been implemented that generates the LHCb-like distribution traffic. Two Ethernet network setups have been evaluated in order to test the various scenarios, with different ratios of data producers versus data consumers. One variant combined 25 and 100 Gb/s links and the other consisted of 100 Gb/s links.

The remainder of the article is structured as follows. Section 2 discusses the details of the LHCb event distribution. Section 3 presents the custom benchmark that was developed to generate the LHCb-like traffic. Performance evaluation, network test benches, and results are discussed in section 4. Conclusions are covered in section 5.

2 Distribution of events to the filtering farm

The evaluated LHCb data acquisition cluster architecture is presented in figure 1.

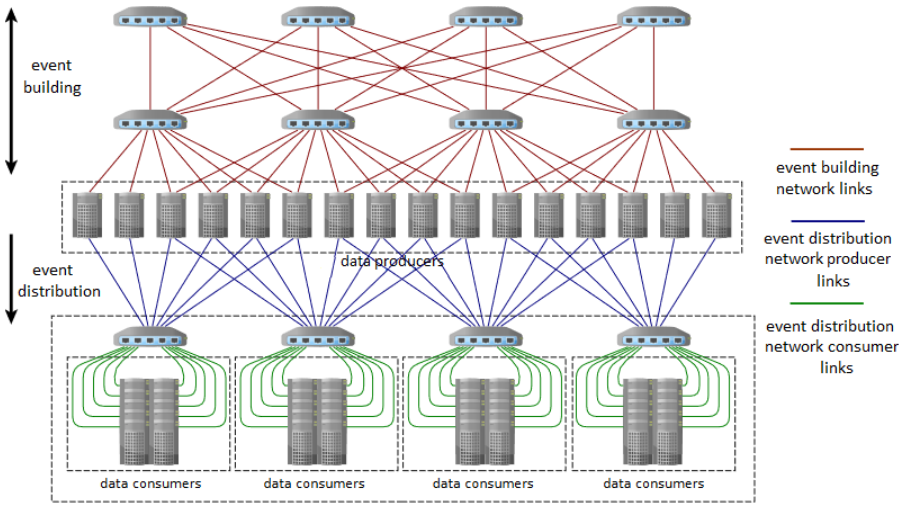

Figure 1. The evaluated LHCb data acquisition cluster architecture. InfiniBand network for the *event building* stage has been successfully tested and is currently commissioned. The *event distribution* over Ethernet in the *filtering farm* has been subject to evaluation.

Two networks can be distinguished - the *(*event building network) and the *filtering farm network*. It has been already decided to implement *event building network* over InfiniBand.

InfiniBand has been proven to be the most performant and reliable for this use case, where the same link speeds could be used. The related extensive evaluation supporting this statement has been documented in [4–10]. This stage is currently deployed. However, the *filtering farm* and its network evaluation have been subject to further study.

Different variants of processing in the *filtering farm* have been considered. One of the scenarios was to use a separate network to distribute assembled events between *data producers*(event builder nodes) and *data consumers* (filtering farm servers). This *event distribution* is depicted in figure 1. The numbers and link speeds of data producers and consumers can differ. The ratio depends on the data processing capacity of filtering nodes versus the output throughput of the event building nodes. Recent development in that scope has been documented in [1, 11, 12]. The architecture presented in figure 1 assumes using approximately 500 *event building nodes* and approximately 4000 *filtering farm nodes*. This is assuming the 100 Gb/s links for the *event building nodes* (data producers) and the 25 Gb/s links for the *filtering farm* (data consumers) nodes. The total input throughput of 32 Terabits per second.

The *filtering farm network* could potentially profit from Ethernet because of the possibility of using different link speeds and because of the cost reduction with cheap shallow-buffered switches.

In the event distribution network, groups of producers and consumers are connected with shallow-buffered Ethernet switches. This division is presented in figure 1. The workload distribution is handled by the switch connected to the group.

Data processing time can differ across filtering nodes. It depends on the type and content of processed events. To evenly distribute workloads, a mechanism had to be implemented to stop transmissions to busy nodes and to redirect traffic. Some data consumer nodes must be able to temporarily take over workloads and receive data at higher throughput.

To evaluate LHCb event distribution traffic, an application was developed to generate LHCb event-like structures with producers and to transmit them to consumers.

3 The benchmark for evaluating the LHCb-like distribution traffic

The custom *dedicated_eb_stresstest* benchmark was implemented in C++ [13]. It uses MPI for data exchange. The application's most important features are as follows:

- The LHCb-like event structures are generated in the *data producer processes*. The throughput and the generation intervals are adjustable. This simulates the input streams from the event building nodes (see figure 1).
- These data are then transmitted to the *data consumer processes*. This corresponds to feeding the filtering farm with assembled events.
- One management process schedules transmissions between the producers and the consumers. Its scheduler is notified through MPI messages about the readiness of the consumers and the producers. It also keeps track of the total amount of transmitted data for each of the producers and consumers. This allows to balance workloads across the processes.
- It is monitored if the data have been transmitted before the next generation is in place. If not, then this corresponds to a situation when the throughput of event distribution is too low to handle the input streams. In that case, the buffers of event builder servers saturate, which then leads to the loss of data. In the assumed LHCb operating conditions, this critical situation should never occur.
- Poisson distribution of processing time is optionally simulated. This forces consumer processes to be temporarily busy.

- Redirection of traffic is handled by the management process. As long as the scheduler is not notified about the readiness of the consumer process, it will not grant him new transmissions.
- The throughputs of data producers and data consumers are probed every second. It is also possible to monitor how many consumer nodes are busy.

The modified version of dedicated_eb_stresstest benchmark has been merged into the LHCb production software [14]. Its event transmission mechanism has been fully incorporated. The code is compliant with the production event data structures and compatible with the WinCC environment of the LHCb Experiment Control system [15].

The benchmark was used for the performance tests. The purpose was to check if stable and efficient and lossless LHCb-like event distribution over Ethernet could be sustained.

4 Performance tests

The *event distribution* stage in figure 1 compartmentalizes consumers and producers into groups of nodes connected with the same switch. The workload is very well scalable by simply increasing the number of such groups and the switches.

Therefore, it was sufficient to measure the Ethernet performance for only a single switch. This would prove the applicability for the entire *LHCb filtering farm network*, given the inherently good network scalability. This also applies to traffic management software. The entire event distribution can be implemented as separate and independent instances of programs, each assigned to service one Ethernet switch. This is possible assuming the sufficiently fair per-server distribution of output during the LHCb *event building*.

An evaluation setup was prepared to evaluate a single-switch fragment of the Ethernet network from figure 1. It is presented in figure 2. If sufficient performance is reached for the test bench, then a full-scale operation can be achieved by multiplying this configuration, each running an independent instance of software to manage its own part of the total traffic.

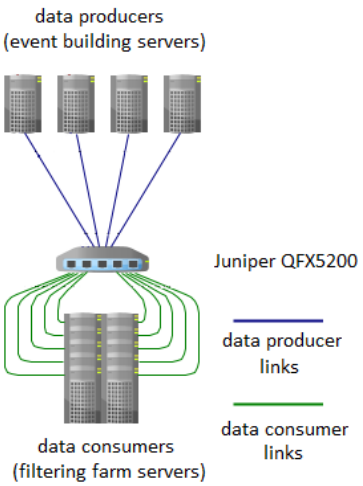

Figure 2. Evaluation test bench for testing event distribution over Ethernet. The setup represents a fragment the full architecture from figure 1. It services an independent and decomposable part of the total 32 Tb/s input data stream.

Two different variants of this test bench were used. They reflected the possible LHCb traffic scenarios and were based on the possible workloads of the event building and filtering stages:

- For the first variant, consumers used 25 Gb/s links, and producers were connected with 100 Gb/s links. The purpose was to evaluate the many-to-many traffic scenario with more consumers than producers.
- For the second variant, both the producers and the consumers used 100 Gb/s links. This setup was used to evaluate the many-to-many traffic scenario where there are more producers than consumers.

For 100 Gb/s links, Mellanox ConnectX-5 network cards were used in the servers. The 25 Gb/s links were connected to servers with Mellanox ConnectX-4 devices. Mellanox OpenFabrics Enterprise Distribution (OFED) was installed on Linux as a software stack. Its MPI distribution was configured to utilize Remote Direct Memory Access (RDMA), Unified Communication X (UCX), and RDMA over Converged Ethernet (RoCE v2). Servers with 100 Gb/s used either AMD Naples EPYC 7301 CPU or Intel Skylake Xeon Silver 4114 CPU. The servers with 25 Gb/s links had the following CPU model: Intel Haswell Xeon E5-2630 v3.

A Juniper QFX5200 shallow-buffered switch was connected with data producers and data consumers. In a scenario with combined 100 Gb/s and 25 Gb/s links, 100-to-25 Gb/s breakout cables were used.

In the tests, either there were more senders than receivers, or the senders had faster links than the receivers. For such traffic, Ethernet flow control protocols had to prevent network congestion. Explicit Congestion Notification (ECN) and Priority Flow Control (PFC) were enabled in the Juniper QFX5200 switch. The UCX framework allowed the configuration of flow control algorithms in the servers. The best performance was reached with PFC enabled. For the setups with ECN, or with a combination of PFC and ECN, performance was poorer. This can be explained by the fact that ECN is a point-to-point, TCP-like protocol that requires the sender to be notified by a receiver. It also needed fine-tuning and was very prone to topology changes. Conversely, the PFC is a link-level protocol that allowed for faster prevention of congestion.

An additional warm-up had to be implemented in the benchmark to mitigate the notable impact of the allocation of the MPI RDMA resources on performance.

For all of the tests, large data sets were generated each second for a period of 300 seconds. It was monitored if these sets were transmitted before the next generations. This allowed assessing whether or not the lossless LHCb-like traffic can be serviced in real-time.

4.1 Performance results for 25 and 100 Gb/s scenario with more consumers than producers

In this setup, 14 data producer servers were using 100 Gb/s. 72 data consumer nodes were connected with 25 Gb/s links. For this LHCb-like traffic scenario, the data producers were expected to generate data at a rate of 85 Gb/s.

The first network stress test was made without simulating the consumers' busyness. Consumers serviced data immediately after receiving them. The throughput of each of the data producers was 85.8 Gb/s. The expected average consumer receiving ratio was 85.8 * 14 / 72 = 16.8 Gb/s. The results are presented in figure 3. As shown in the two plots, the data producers were sending data at a very stable throughput of 85.8 Gb/s. Data consumers received at rates ranging from 11.8 to 20.4 Gb/s. The average receiving rate per consumer was 16.8 Gb/s. Stable and well-balanced data distribution was maintained.

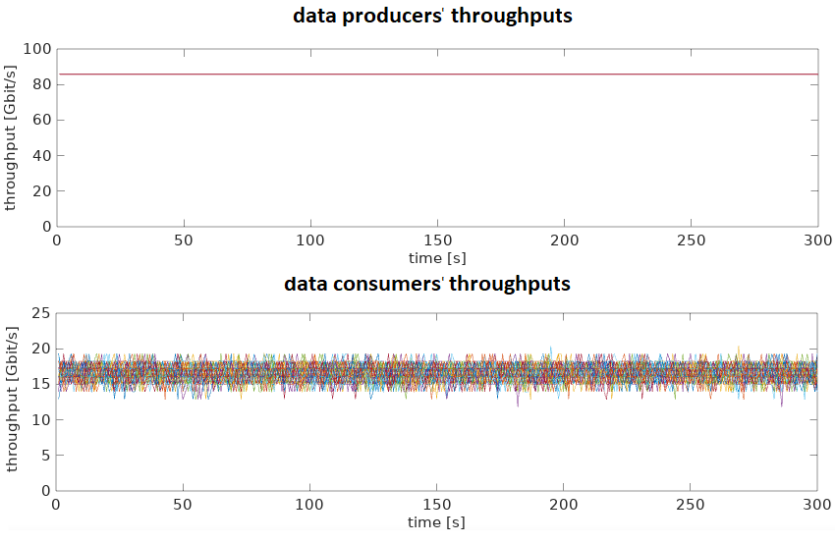

Figure 3. The performance stress test of the network with 100 Gb/s and 25 Gb/s links.

The second test was made to evaluate the network traffic when some data consumer nodes are busy. The Poissonian distribution was used to simulate the processing time. The producers-ports-versus-consumers-ports-throughput ratio was 18/14 = 1.28. Therefore, the average processing time per data consumer should not exceed 1.28 seconds. Above this value, filling the producers' buffers is inevitable. For the test, a close-to boundary average processing time of 1.2 seconds was adjusted. The results are presented in figure 4.

As shown in the top plot, up to 12 nodes were temporarily busy. The busyness of the data consumers can also be seen at the bottom plot, where some nodes temporarily had zero throughputs. The transmissions were correctly redirected to other nodes in these situations. As shown in the middle plot of figure 4, data producers managed to send out all the data before the next sets were generated. Some consumer nodes were receiving data at up to 20.4 Gb/s. A stable LHCb-like operation was sustained for the full test period of 300 seconds. The fatal scenario of saturating the buffers and losing data was avoided. The run was very stable despite a very aggressive context switch and traffic redirections across data consumers.

The two presented tests have proven that stable operation is possible for the combined link speeds and the LHCb-like traffic with more consumers than producers.

4.2 Performance results for 100 Gb/s links scenario with more producers than consumers

In the subsequent setup, 18 nodes were used as data producers and 14 as data consumers. The purpose was to test the LHCb event distribution traffic scenario where producers send data at a rate of 60 to 70 Gb/s and the receivers throughput is approximately 80 Gb/s.

In the *third* network stress-test, the data generation rate on the producers increased to the maximal point where a stable operation was still sustained. The processing time on the consumers was set to zero so that only the network was tested. The best results are presented in figure 5. As such, the stable operation was reached for an average throughput per data consumer of 95.2 Gb/s. As shown in figure 5, data producers were feeding data at a very

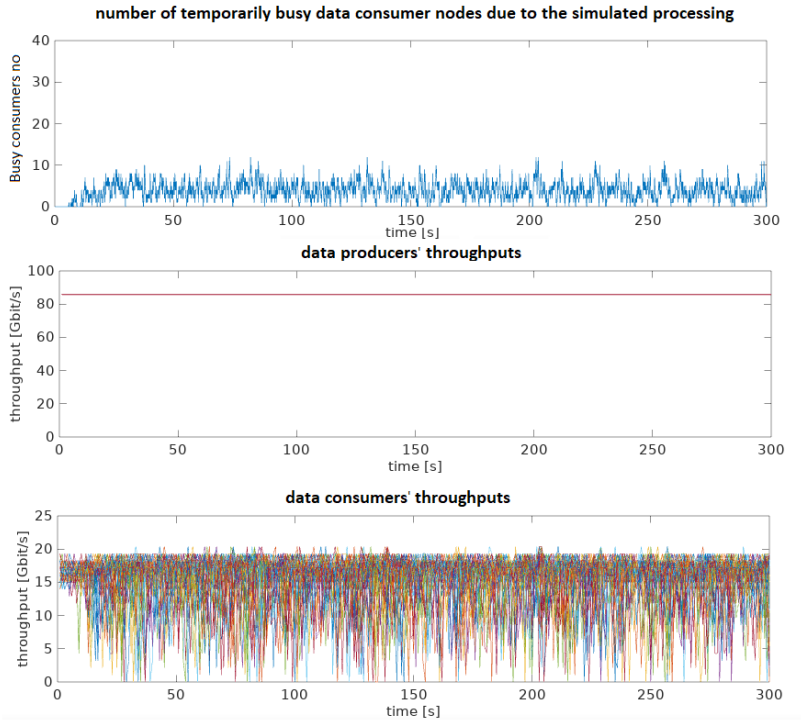

Figure 4. Performance stress test of the network with 100 Gb/s and 25 Gb/s links with simulated temporary busy state of nodes.

stable 74 Gb/s, and in every second generated data sets were sent out by producers before the subsequent generation. This test has proven that a very stable execution of a benchmark is possible at close-to-the-link speed. The following step was investigating the behavior of the network with temporarily busy data consumer nodes.

The *fourth* test was made with the simulated Poissonian processing time on the consumers' site. It was adjusted to the average value of 0.9 seconds. Producers were generating data at a rate of 63.3 Gb/s per producer. Consumers were receiving at an average rate of 81.4 Gb/s. In this situation, a sporadic busy state of some nodes was to be expected. Additionally, because of lower throughput in comparison with figure 5, the consumers had some throughput slack to take over from busy nodes. The results can be seen in figure 6. Up to 6 consumers were temporarily busy. In that case, the other consumer nodes were fed with the workloads. Their throughput increased, as shown at the bottom plot of figure 6. Stable operation of the data producers was sustained and the LHCb-like real-time operation caveats were met.

The last, *fifth* test evaluated the performance of the network when the nodes were overloaded. For this purpose, the average processing time on the data consumer nodes was increased from 0.9 seconds to 1.2 seconds. This inevitably led to filling the data producer buffers and a loss of data. Under these conditions, the stress test allowed getting the maximal throughput of consumer nodes exposed to intensive traffic switching. The results are presented in figure 7.

Slower processing than generation inevitably led to the filling of producers buffers, as shown at the bottom plot of figure 7. In such a case, buffer occupancy was above one. In the

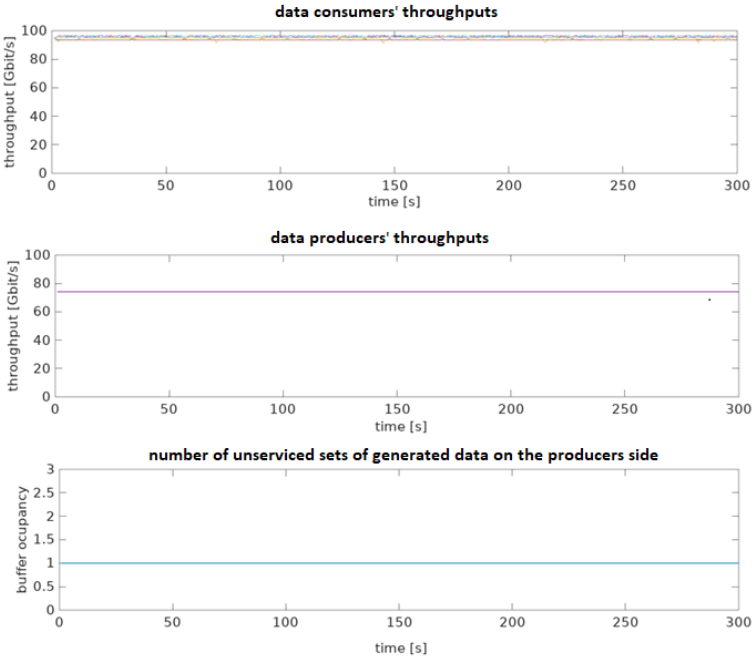

Figure 5. The performance stress test of the network with 100 Gb/s links.

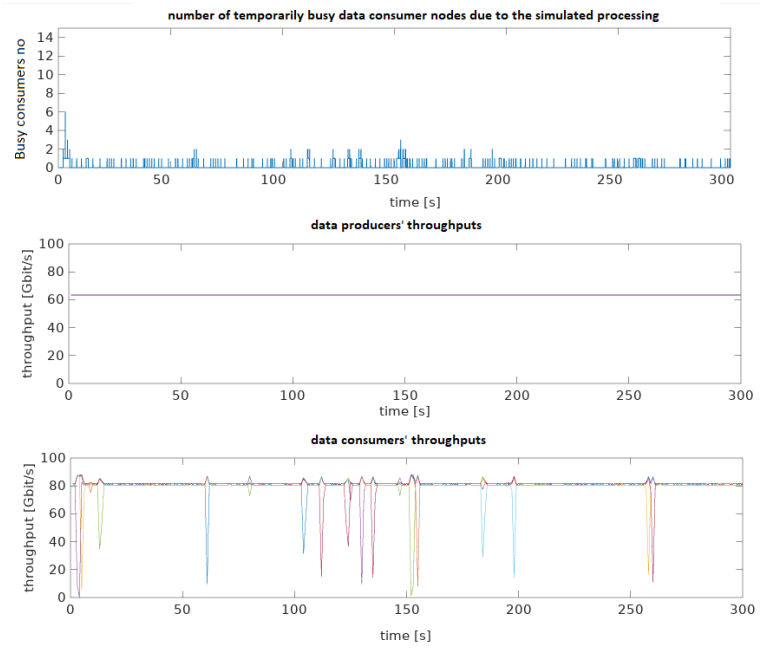

Figure 6. The 100 Gb/s test with higher data processing throughput versus data generation throughput.

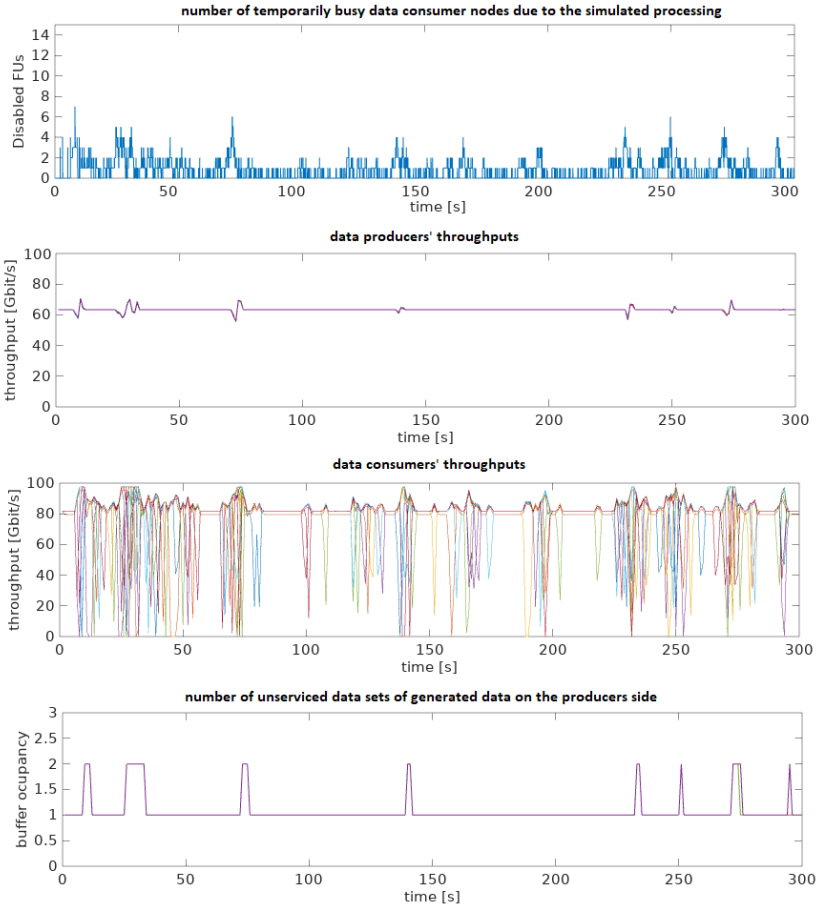

Figure 7. The 100 Gb/s test with lower data processing throughput versus data generation throughput.

test, it was assumed that the amount of generated data per second fills the producers' buffers. This means that the buffer occupancy above 1 in the benchmark reflects the situation where producers' buffers are saturated and data are lost.

However, the network exposed to this situation was still very stable and performant. During the first 200 seconds of the test, the producers' buffers were mostly not saturated. As shown at the top plot of figure 7 for most of the time, data consumers had to handle the temporary busy state of at least one node. Data consumer nodes that took over the workloads were reaching 95 Gb/s of input throughput. The tested network could handle the temporary busyness of up to 4 nodes and permanent busyness of up to 2 nodes. This can be observed when comparing the top and the bottom plots of figure 7 for the 40th and the 200th second of the test.

The tests have proven that this LHCb-like traffic for a scenario with fewer consumers than producers.

4.3 Results summary

The results for each of the scenarios from subsections 4.1 and 4.2 are summarized in table 1. The most important setup details and relevant results are included. In the assumed typical operation of LHCb for tested scenarios 1-4, the real-time stability of the network has been sustained (by not saturating the producers' buffers). For the test of scenario 5, the network saturation was forced with simulated computation time inevitably leading to buffers saturation. In such a case, the ability to redirect the traffic by the network was proven as well as the capability of consumer nodes to temporarily handle large workloads. Either of the tests proves sufficient network performance and reliability, even at the close-to-the-maximal link utilization.

Table 1. Summary of results from the five tested traffic scenarios

Test	Fig	Producers	Consumers	Per-producer input stream	Relevant results
1	3	14 x 100 Gb/s	72 x 25 Gb/s	85.8 Gb/s	No simulated busyness, sustained real-time
2	4	14 x 100 Gb/s	72 x 25 Gb/s	85.8 Gb/s	Simulated busyness, sustained real-time for up to 12 busy consumers, consumers reaching 20.4 Gb/s
3	5	18 x 100 Gb/s	14 x 100 Gb/s	74 Gb/s	No simulated busyness, sustained real-time, consumers reaching 95.2 Gb/s
4	6	18 x 100 Gb/s	14 x 100 Gb/s	63.3 Gb/s	Simulated busyness, sustained real-time for up to 6 busy consumers
5	7	18 x 100 Gb/s	14 x 100 Gb/s	63.3 Gb/s	Long busyness to force network saturation, consumers reaching 95 Gb/s

5 Conclusions and future studies

In this work, we presented a feasibility study, whether or not the Ethernet networks can handle the LHCb workload dispatch in the LHCb filtering farm. Our purpose was to check if Ethernet with flow control protocols is performant enough for the many-to-many traffic, with different link speeds with different numbers of data receivers and consumers.

We conducted the tests with the custom benchmark that produced the LHCb-like data, generated the traffic, and simulated the temporary busy state of data consumers. Two network setups were prepared that reflected the traffic of the presented potential LHCb data distribution. The first one consisted of the same 100 Gb/s links (with more producers than consumers). The second one combined 100 Gb/s and 25 Gb/s links (with more consumers than producers).

The workbench was installed with a single, fully-populated shallow-buffered switch that mimicked a fragment of the LHCb architecture presented in figure 1. The purpose was to measure the performance of the network traffic occurring for a single group of consumers and producers connected by a single switch. In such a case, satisfactory results for a single switch transform to satisfactory at a full scale. This because the *LHCb filtering farm network*

traffic is inherently very well-scalable, with very well decomposable workloads. A proof of concept for a single switch for our use case meant good performance at full scale.

In the network tests close-to-the maximal links capacity was reached either on the producer or the consumer side. A very stable operation was reached both for the same-speed links and the mixed-speed links. For the same-link-speed scenario, a stable real-time operation was reached with an average throughput of above 95 Gb/s for the 100 Gb/s links. This was achieved with Priority Flow Control (PFC) enabled. The tests of both of the setups proved that the networks could handle the temporary busyness of several consumers and efficiently redirect traffic to other devices in run time, provided that the remaining consumers can still handle the throughput. The feasibility studies have proven that stable lossless real-time LHCb-like event distribution traffic is applicable over Ethernet.

This solution can also profit in the future from 200 Gb/s links. Satisfactory results were reached with combined 25 and 100 Gb/s link speeds. Tests with both 100 and 200 Gb/s links will be possible in the future, as soon as 200 Gb/s Ethernet network cards are available for evaluation.

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] J. M. Jimenez, et.al, *Summary of session 8: Long Shutdown 2 strategy and preparation*, Proceedings of Chamonix 2014 Workshop on LHC Performance (2016), https://e-publishing.cern.ch/index.php/CYR/article/view/111/55

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

[4] 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

[5] T. Colombo, et al., *The LHCb Online system in 2020: trigger-free read-out with (almost exclusively) off-the-shelf hardware*, Journal of Physics: Conference Series vol. 1085 (3), (2018), https://inspirehep.net/literature/1699857

[6] 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

[7] F. Pisani, et al., *Network simulation of a 40 MHz event building system for the LHCb experiment*, EPJ Web of Conferences vol. 245 (2020).

[8] T. Colombo, et al., *Flit-level InfiniBand network simulations of the DAQ system of the LHCb experiment for Run-3*, IEEE Transactions on Nuclear Science, vol. 66 (2019).

[9] R. D. Krawczyk, et al. *Feasibility tests of RoCE v2 for LHCb event building*, EPJ Web of Conferences vol. 245 (2020).

[10] R. Krawczyk et al.: 32 Tb/s DAQ for the LHCb experiment at CERN, DAQFEET-21 Worksop slides, https://indico.cern.ch/event/974424/contributions/4217589/attachments/2186332/3694141/DAQFEET_9_02_21_FINAL.pdf

[11] C. Bozzi, et al., *Towards a computing model for the LHCb Upgrade* EPJ Web of Conferences vol. 214 (2019), https://www.epj-conferences.org/articles/epjconf/pdf/2019/19/epjconf_chep2018_03045.pdf

[12] R. Aaij, et al., *Design and performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC*, Journal of instrumentation vol. 4 (2019), https://arxiv.org/abs/1812.10790

[13] R. D. Krawczyk, *dedicated_eb_stresstest* benchmark, CERN GitLab Repository, https://gitlab.cern.ch/lhcb-online-eb/dedicated-eb-stresstest

[14] M. Frank, et al. *Online project*, CERN GitLab Repository, https://gitlab.cern.ch/lhcb/Online

[15] L. Granado Cardoso, et al., *Integration of custom DAQ Electronics in a SCADA Framework*, PJ Web of Conferences vol. 245 (2020), https://www.epj-conferences.org/articles/epjconf/abs/2020/21/epjconf_chep2020_01016/epjconf_chep2020_01016.html