Evaluation of 100 Gb/s LAN networks for the LHCb DAQ upgrade

Balázs Vőneki, Sébastien Valat, Rainer Schwemmer, Niko Neufeld, Jon Machen, Daniel Hugo Cámpora Pérez

***Abstract*–The LHCb experiment[1] is preparing a major upgrade resulting in a need for a high-end network for a data acquisition system. Its capacity will grow up to a target speed of 40 Tb/s, aggregated by 500 nodes. This can only be achieved reasonably by using links capable of coping with 100 Gb/s line rates.**

The constantly increasing need for more and more bandwidth has initiated the development of several 100 Gigabit/s networks. There are 3 candidates on the horizon, which need to be considered: Intel® Omni-Path, 100G Ethernet and EDR InfiniBand.

We present test results with such links both using standard benchmarks (e.g. iperf) and using a custom built benchmark called LHCB-DAQPIPE. The key benefit of these measurements is that we can get to know better the behavior of the system in the early development stage, thus we can find out the limitations of the different network components. This can give an idea as to whether there is a need for more focus on some elements, which can be optimized in the future.

I. Introduction

The Large Hadron Collider (LHC) has 4 big experiments, one of them is LHCb. It has an underground detector which gathers information from particle collisions at high energies. The designed rate of the collisions is 40 MHz. In order to be able to deal with the large quantities of data generated at the collider, one needs to reject the irrelevant events and keep only those events, which are interesting to us for later analysis. This procedure is called triggering.

Currently the LHCb operates by applying two levels of triggers: a low-level trigger (LLT) and a high-level trigger (HLT), where the first is realized by FPGA-based custom hardware, reducing the 40 MHz input rate to 1 MHz. So the HLT has to further process only this 1 MHz rate of incoming events.

The LHCb experiment will be upgraded during the LHC Long Shutdown 2 starting from 2018 until 2019[2]. One of the key features of this upgrade is the removal of the low-level trigger. Hence, it will end with a trigger-free readout and fully software-driven trigger.

In this paper we will shortly describe the network communication which will be needed on those 500 nodes, then we will discuss the benchmarking results we had going from really simple available bandwidth benchmark to more realistic custom made benchmarks. This analysis will end up with a full running test on 16 nodes equipped with 100 Gb/s InfiniBand EDR boards.

II. 100 Gbit/s Technologies Overview

The LHCb experiment is considering three different 100 Gbit/s technologies for the upgrade[3]:

- 100 Gb/s Ethernet
- Intel® Omni-Path
- EDR InfiniBand

III. Event Building And Data Flow

For the next LHCb upgrade, we will setup an event building cluster of 500 nodes to read and aggregate the 40 Tb/s of data going out of the detector. As shown by Fig. 1, the data will arrive from the sub-detectors to the surface. Using standard servers to host those readout units now permits to manage easily the buffering and to handle the event building onto a 100 Gb/s standard fabric from the HPC field. Once the data have been aggregated by the Builder Units (BU) they have to be sent to one of the 3500 Filter Units (FU) onto a second network. Those filter units will apply the software triggering rules.

We decided to host the Readout and Builder units on the same host for cost efficiency, it will require an inner memory throughput up to 400 Gb/s for each node. This requirement seems to be acceptable as shown in Fig. 2 by using the stream (see Fig. 3) on a dual socket Intel® Xeon E5-2690 server. This event building process will also be required to handle up to 100 Gb/s bi-directionally on the event building network (which requires some experimentation). This is what we want to benchmark.

Manuscript received June 10, 2016.

B. Vőneki, S. Valat, R. Schwemmer and N. Neufeld are with CERN, Switzerland, 978-1-5090-2014-0/16/$31.00 2016 European Union.

J. Machen is with DCG IPAG, EU Exascale Labs, Intel® Semiconductor AG, INTEL Switzerland.

D. H. Cámpora Pérez is with Universidad de Sevilla, Spain.

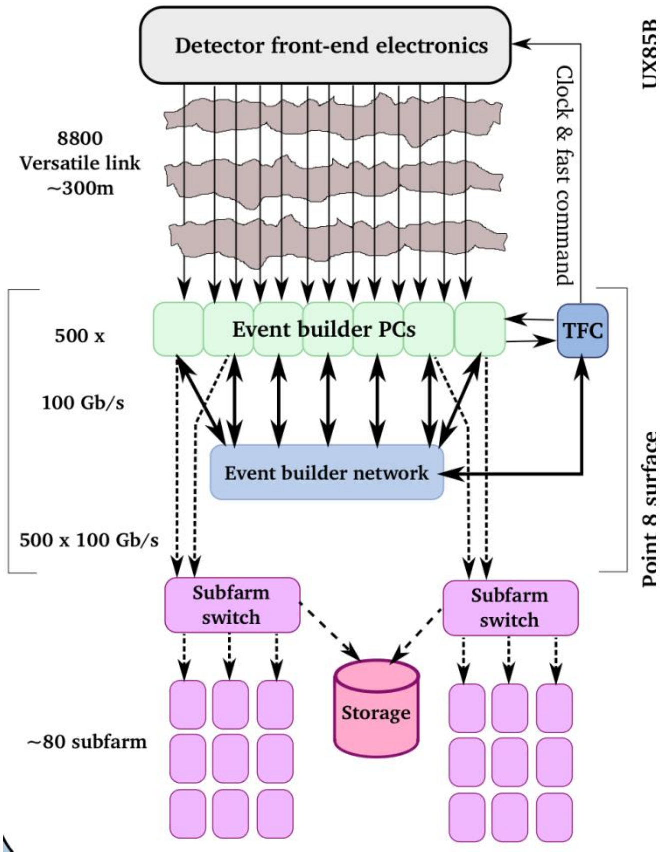

Fig. 1. The architecture of the upgraded LHCb readout-system.

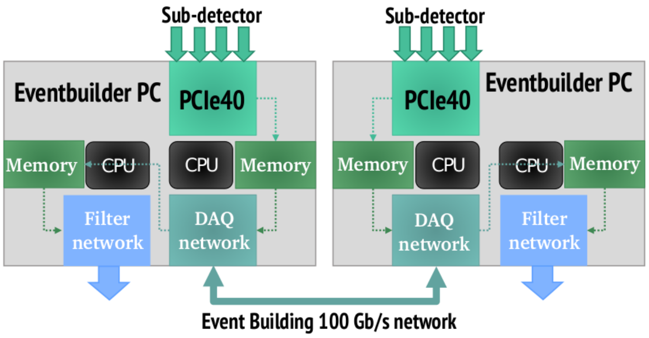

Fig. 2. Dataflow per node.

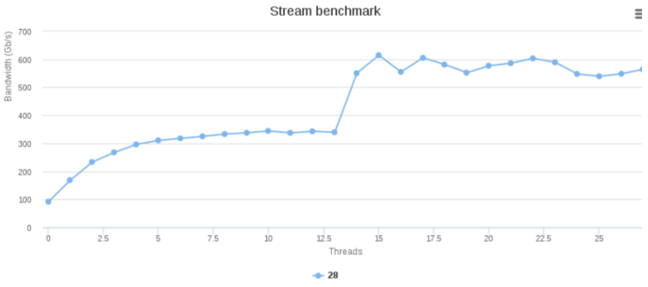

Fig. 3. Stream benchmark.

IV. Simple Benchmarks

As a first step we wanted to evaluate the available network technologies with simple available benchmarks. On HPC networks we can use the MPI OSU benchmark which simply applies a ping-pong communication pattern between the two nodes to evaluate the bandwidth. This ping-pong can be in one way (OSU_BW) or two way (OSU_BIBW). The result for Intel® Omni-Path are provided in Fig. 4.

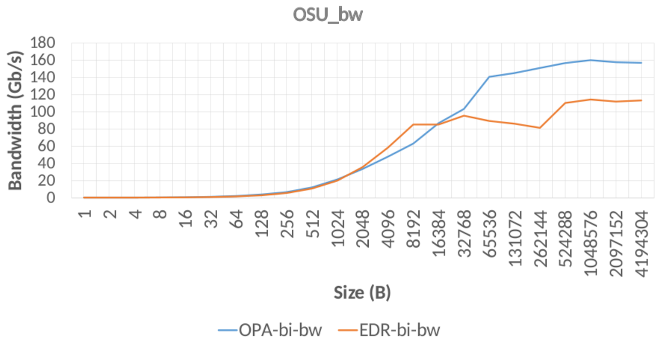

Fig. 4. Simple MPI OSU benchmark between two nodes.

In the same idea we can also evaluate the 100 Gb/s Ethernet solution with the iperf3 benchmark. We have recently received Intel® Boulder Rapids cards, which are very likely to gain transfer speeds.

V. Using Synthetic Benchmark

Before moving to the more realistic benchmark we wanted to explore some synthetic ones to evaluate the potential issues which can arise. We implemented 4 benchmarks to simulate various communication patterns: One-to-one, Many-to-one and Gather. The benchmarks also permit to handle a controlled number of on-the-fly messages. The best results obtained show that we achieve bandwidth larger than 86 Gb/s in any case by selecting the good configuration.

VI. DAQPIPE Benchmark

In order to evaluate the real event building solution we built DAQPIPE (DAQ Protocol Independent Performance Evaluator). This benchmark can run in various running modes. The results are given by Fig. 5.

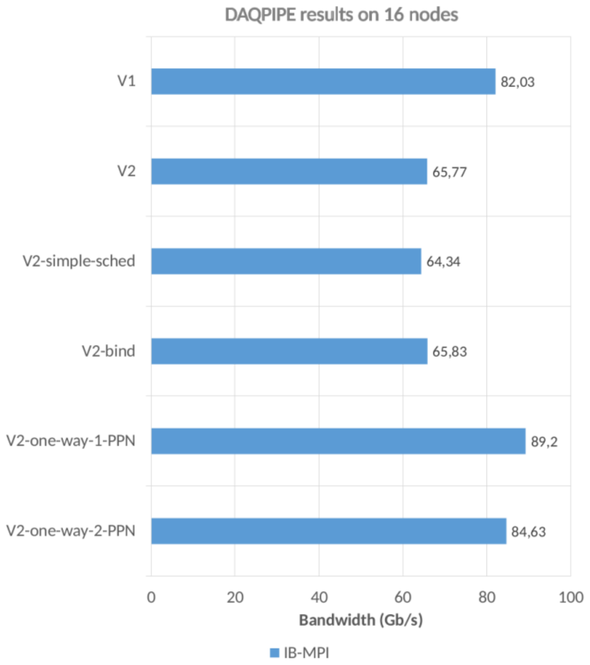

Fig. 5. DAQPIPE results on 16 nodes

VII. Conclusion

We have proved that we can achieve good bandwidth usage up to 16 nodes with 86 Gb/s per process which is what we expected to see. The next step will be to evaluate our benchmark at larger scale. We are also looking on failure recovery support to handle failure of nodes during data taking.

References

[1] Daniel Hugo Cámpora Pérez, Rainer Schwemmer, Niko Neufeld, *Protocol-Independent Event Building Evaluator for the LHCb DAQ System*, IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 62, NO. 3, JUNE 2015

[2] The LHCb Collaboration *et al.*, "LHCb Trigger and Online Upgrade Tech. Design Rep.," CERN-LHCC-2014-016; LHCB-TDR-016, 2014.

[3] Adam Otto, Daniel Hugo Cámpora Pérez, Niko Neufeld, Rainer Schwemmer, Flavio Pisani, *A first look at 100 Gbps LAN technologies, with an emphasis on future DAQ applications*, in 21st International Conference on Computing in High Energy and Nuclear Physics, 2015