A first look at 100 Gbps LAN technologies, with an emphasis on future DAQ applications.

Journal of Physics: Conference Series **664** (2015) 052030 doi:10.1088/1742-6596/664/5/052030

A first look at 100 Gbps LAN technologies, with an emphasis on future DAQ applications.

Adam Otto, Daniel Hugo Cámpora Pérez, Niko Neufeld, Rainer Schwemmer, Flavio Pisani
CERN, Geneva, Switzerland

E-mail: adam.otto@cern.ch

Abstract. The LHCb experiment is preparing a major upgrade, during long shutdown 2 in 2018, of both the detector and the data acquisition system. A system capable of transporting up to 50 Tbps of data will be required. This can only be achieved in a manageable way using 100 Gbps links. Such links recently became available also in the servers, while they have been available between switches already for a while. We present first measurements with such links using standard benchmarks and using a prototype event-building application. We analyse the CPU load effects by using Remote DMA technologies, and we also show comparison with previous tests on 40G equipment.

1. Introduction

In 2018, during Long Shutdown 2 (LS2) of the Large Hadron Collider, the LHCb experiment will undergo a major upgrade. This upgrade has two principle goals: to improve detectors and electronics such that the experiment can run at an instantaneous luminosity of $2 \times 10^{33}\ cm^{-2}s^{-1}$ and to read out every detector element for every bunch-crossing, thus essentially creating a trigger-less experiment [1]. This will increase the event-size from $50 - 60$ kB (LHC runs 1 and 2) to 100 kB. It will increase the event-rate from 1 MHz to 40 MHz. These two improvements will result in an aggregated bandwidth of 32 Tbit/s. The input data will be distributed in about 500 data-sources, connected to the detector front-end electronics [2]. The key parameters are compared with the situation in Runs 1 and 2 in Table 1. In order to achieve such a high

Table 1. Key paramaters of the LHCb DAQ

	Runs 1&2	Run3
event-size[kB]	50 - 60	100
event-rate[MHz]	1	40
# data sources	313	500
# data sinks	up to 2000	up to 5000

bandwidth rate, the LHCb experiment will need to deploy 100 Gbit/s technology. The future architecture of data aquisition network is presented in figure 1, 100 Gbit/s technology will be deployed in 500 Eventbuilders PCs[1].

[1] to achieve an average link-load of about 80 Gbit/s

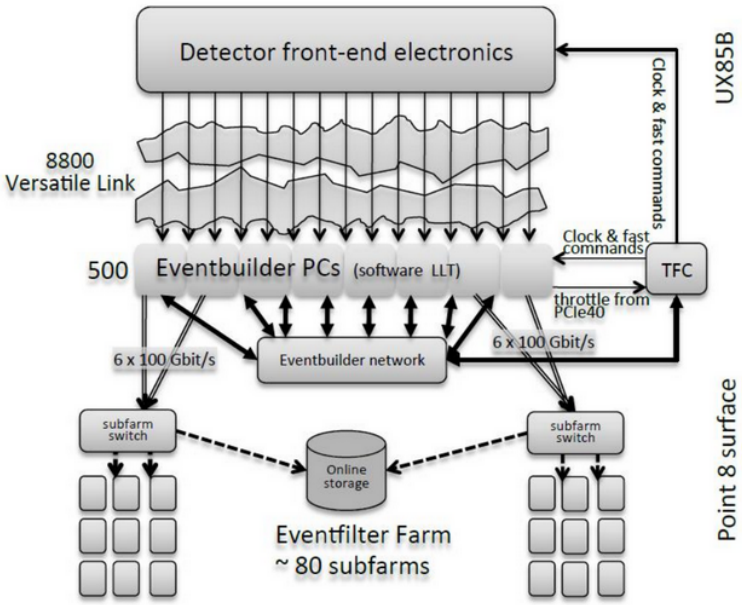

Figure 1. LHCb DAQ Architecture 2018.

2. 100 Gbit/s technologies overview

For this upgrade, the LHCb experiment is considering three different 100 Gbit/s technologies: 100G Ethernet, Intel Omni-Path, EDR InfiniBand.

2.1. 100G Ethernet

100 Gigabit ethernet is a technology defined in the IEEE 802.3ba-2010 standard. At the time of writing this paper, no end-devices were available with this technology implemented.

2.2. Intel Omni-Path

The Intel Omni-Path is new high-speed interconnect technology optimized for HPC deployments. It is supposed to offer 100 Gbit/s bandwidth, and a 56 % gain in lower latency over InfiniBand fabrics [3]. The release date is yet unknown.

2.3. EDR InfiniBand

Enhanced Data Rate (EDR) is the next stage on the InfiniBand roadmap after Fourteen Data Rate (FDR). It is able to handle 100 Gbit/s bandwith with latency lower than 0.5 μs. An EDR link is composed of four lanes, each able to handle 25 Gbit/s. The latest generation is PCIe 3.0 Gen compliant.

3. Study of 100G InfiniBand

The LHCb experiment was able to test a very first release of the ConnectX-4 network card (NIC), which supports both 100 Gbit/s ethernet and InfiniBand. Unfortunately due to preproduction version of drivers, only InfiniBand mode was supported at the time of testing. We tested the network card in two different configurations: two directly connected NICs and NICs connected through a switch (figure 2).

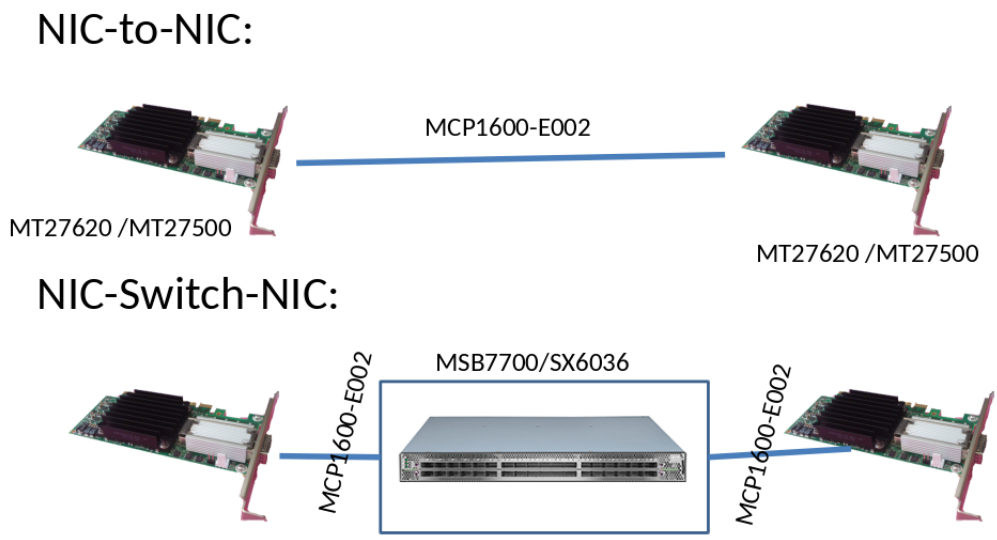

Figure 2. Two setups:1)NIC-NIC 2)NIC-SWITCH-NIC

3.1. Testbed details
For testing purposes, we used four servers with the following configuration *2 x Intel(R) Xeon(R) E5-2650 v3 CPU* running at 2.30 GHz, each of the CPU have 10 cores and 20 threads and 64 GB of RAM. For comparison, we tested two generation of Host Channel Adapters (HCA): *Mellanox Technologies MT27620 Family- ConnectX-4* and *Mellanox Technologies MT27500 Family- ConnectX-3*. Switches: EDR capable*MSB7700*, FDR capable *SX6036*. We ran *Scientific Linux CERN SLC release 6.6*, which is very similar to CentOS of the same version. Preproduction OFED[2] version of drivers used was *2.4-1.0.1* and the EDR switch was running on beta version of MLNX OS *3.4.1102*. The used OFED benchmarks were version *5.33* and MPI[3] was *1.8.4*.

3.2. Measurement methodology.
Most of the results were obtained using standard OFED benchmarks, which were run for 40 seconds per specific message size. The obtained network bandwidth, latency was averaged over a period of 40 s. CPU usage was measured using sar[4] from the sysstat package[4]. It was run in parallel to transmission for specific message sizes, the result was an average usage for time of this transmission. Memory bandwidth was measured with the Intel PCM (Performance Counter Monitor) during transmission, and the result was average usage over this transmission. The network and memory bandwith were represented as Gbit/s, latency as microseconds and CPU usage as used percentage of total CPU power.

4. Results
We performed our tests with different OFED and MPI benchmarks and with our proprietary EventBuilder simulation software. The results are represented on the plots in order to facilitate read. Concerning the representation of bandwidth results, on the left side of the plot, there is a shared scale for network and memory bandwidth in Gbit/s, on the right side, we have a scale

[2] OpenFabrics Enterprise Distribution, https://www.openfabrics.org/index.php/openfabrics-software.html
[3] Message Passing Interface, http://www.open-mpi.org/
[4] system activity reporter

for CPU usage in %, the x-axis is logarithmic and represents message size in bytes. For latency results, the left side of the plot represents latency in microseconds, the right side is for CPU usage, and the logarithmic x-axis is in bytes for specific message size.

4.1. OFED benchmark before tuning

From the OFED benchmark suite, we used the following benchmarks: *ib_read_bw, ib_read_lat, ib_send_bw, ib_send_lat, ib_write_bw, ib_write_lat.* Each of these benchmarks test different aspects of the infiniband fabric. Ib_read_bw/ib_read_lat tests the bandwidth or latency of RDMA[5] read transactions; ib_send_bw/ib_send_lat tests parameters of RDMA send transactions; ib_write_bw/ib_write_lat tests RDMA write transactions. Tests were performed for both cases. The plots only represents results for read benchmarks, and the rest of the results are regrouped in a table (table 2). The maximum attained bandwidth with ib_read_bw was around 70 Gbit/s, CPU usage was stable at 2.5 % of total utilization. Memory bandwidth only started to increase after messages larger than 1 Mbyte were used. The results were identical for both test cases (figure 3). Latency was linear and only started to increase significantly after 100 KB messages were transmitted (figure 4).

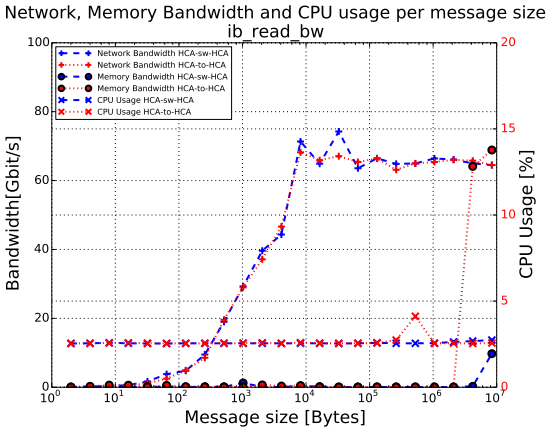

Figure 3. ib_read_bw bandwith comparison between two testcases.

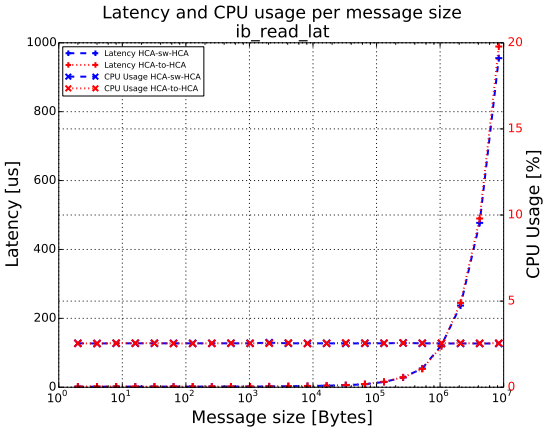

Figure 4. ib_read_lat latency comparison between two testcases.

The results for the rest of benchmarks are presented in table 2.

Table 2. The results for the rest of benchmarks.

Benchmark	Maximal Bandwidth(Gbit/s)	Average CPU usage(%)
ib_send	78	3
ib_write	80	2.5

4.2. Tuning

In order to improve performance, the machines were tuned at BIOS and kernel level. The tuning was done according to the Mellanox documentation [5]. The following options were set in the BIOS: power profile was set to maximum performance, c-states were disabled, turbo mode was enabled and finally memory speed was increased to maximum performance. On the kernel

[5] Remote direct memory access

level, we ran an affinity script from the Mellanox website (http://www.mellanox.com/related-docs/prod_software/mlnx_irq_affinity.tgz) in order to set interrupts accordingly to system architecture.

4.3. OFED benchmark after tuning

After tuning, a 10 Gbit/s gain in bandwidth was observed for ib_read_bw benchmark tests resulting in 80 Gbit/s total bandwidth. CPU usage stayed at the same level 2.5 % and memory bandwidth also started to increase for messages larger than 1 MB. The latency did not change. For comparison, the figures 5 and 6 present results from pre-tuned (red dotted curve) and post-tuned (blue dashed curve) runs. The results for the rest of benchmarks are represented in table 3.

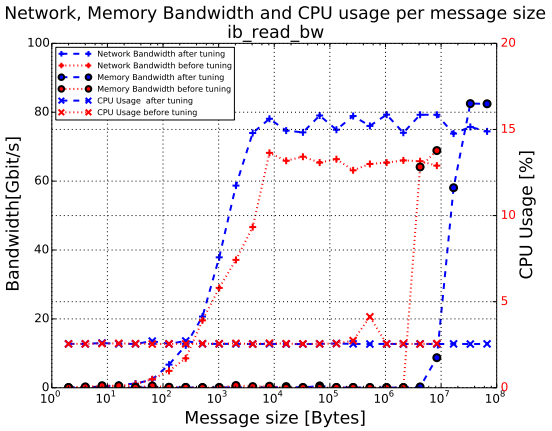

Figure 5. ib_read_bw bandwith comparison between tuned and untuned runs.

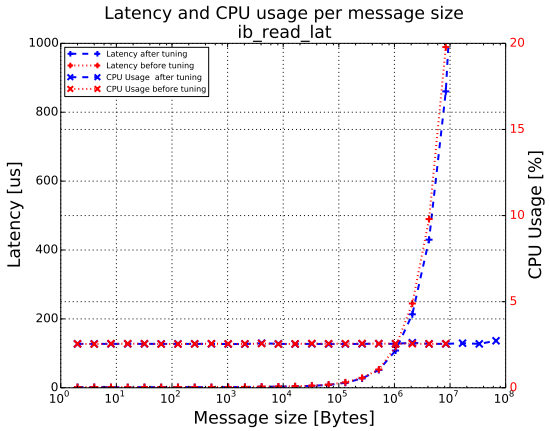

Figure 6. ib_read_lat latency comparison between tuned and untuned runs.

Table 3. The results for the rest of benchmarks.

Benchmark	Maximal Bandwidth(Gbit/s)	Average CPU usage(%)
ib_send	85	3
ib_write	84	2.5

4.4. OSU Benchmark

To test full duplex connectivity,the OSU benchmark[6] was run in both directions. As can be seen in figure 7, the maximum attained bidirectional bandwidth rate was of 150 Gbit/s and the maximum unidirectional bandwidth rate was of 75 Gbit/s. The cpu usage was around 5 %. The fluctuation that can be seen on the left side of the plot, are related to very small message sizes, and this result in very short transmission times, so these CPU usage readouts weren't accurate. The CPU usage values for messages larger than 10KB were only considered for this test.

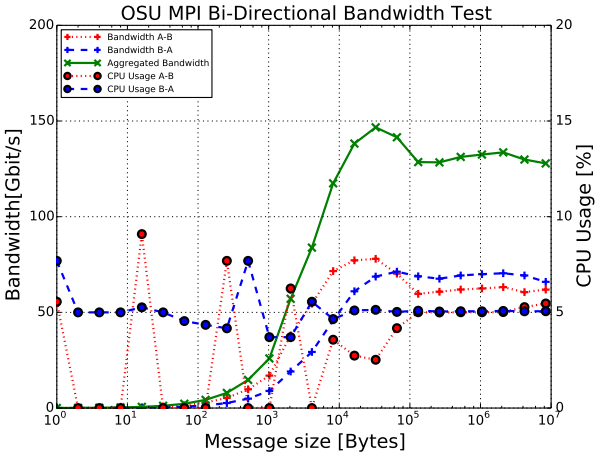

Figure 7. OSU Benchmark resulting plot

4.5. LBDAQPIPE[6]

Next tests were performed with our Event Builder simulation software LBDAQPIPE [7] which can be run using MPI messages or standard verbs sockets. For the RDMA run, the overall bandwith performance was very stable regardless of message size (figure 8). The best bandwidth rate achieved was 80.6 Gbit/s. For the MPI implementation the highest bandwidth rate achieved was 80.5 Gbit/s. Both results are very promising and a 97 % of standard benchmark performance was achieved.

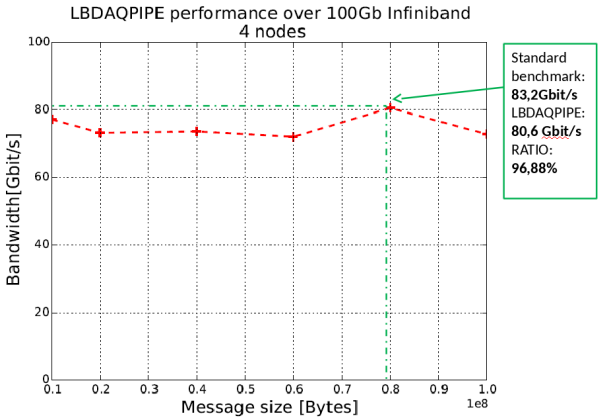

Figure 8. LBDAQPIPE RDMA performance

4.6. Generation comparison. ConnectX-3 vs. ConnectX-4

The last test consisted of comparing two generation of connectX and checking if there are any differences in their performance, except for bandwidth. As we can see on figures 9 and 10 the CPU usage is slightly higher for previous generation (red dotted curves). Memory bandwidth

[6] LBDAQPIPE is a evaluation tool designed specifically to measures the performance of the LHCb Data Acquisition system including network, hardware components, network protocol and DAQ schemes. It provides protocol independent services for network-based DAQ system.

behaves the same for both generations and only starts to increas for messages larger than 1 MB. Regarding latency, ConnectX-4 (blue dashed curve) performs better, with about 200 μs difference for 5 Mbyte messages. Latency only starts to increase significantly after the size of messages becomes larger than 100 KB.

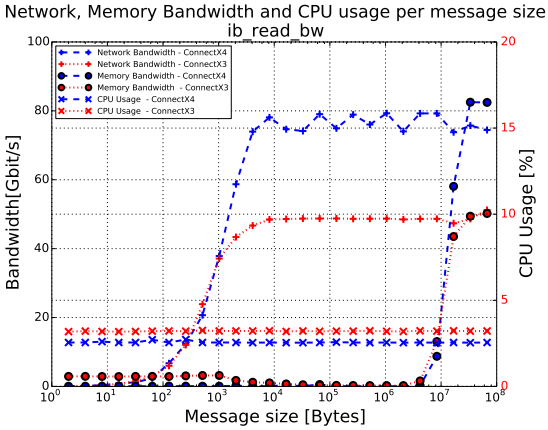

Figure 9. ib_read_bw bandwith comparison between generations.

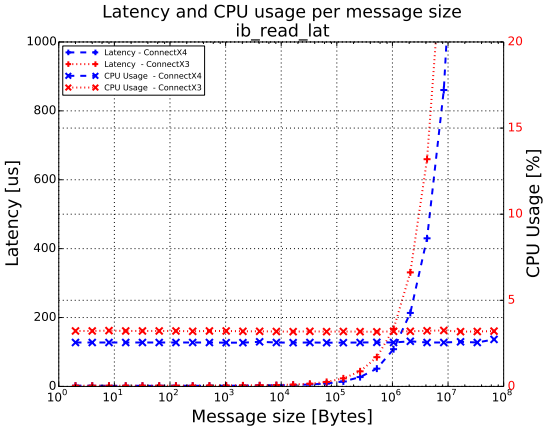

Figure 10. ib_read_lat latency comparison between generations.

5. Conclusion

In spite of a very early versions of the drivers used for the newest generation of ConnectX, we could achieve 85 bit/s in half-duplex communication and 150 Gbit/s in full-duplex. Our own Event Builder simulation software ran at about 80 Gbit/s in both possible modes: MPI, VERBS. These results are very encouraging, as they can be obtained at modest CPU load in modern servers as required by the future LHCb event-builder. Further 100G tests will be performed as Ethernet mode become available.

References

[1] Bediaga I et al. (LHCb collaboration), *2012 Framework TDR for the LHCb Upgrade: Technical Design Report*, Tech. Rep. CERN-LHCC-2012-007. LHCb-TDR-12 CERN Geneva
[2] Guoming Liu and Niko Neufeld, *DAQ Architecture for the LHCb Upgrade*, CHEP 2013
[3] Intel Press Release, http://newsroom.intel.com/community/intel_newsroom /blog/2014/11/17/intel-reveals-details-for-future-high-performance-computing-system-building-blocks-as-momentum-builds-for-intel-xeon-phi-product
[4] Sysstat Documentation, http://www.linuxfromscratch.org/blfs/view/svn/general/sysstat.html
[5] Mellanox Documentation, http://www.mellanox.com/related-docs /prod_software/Performance_Tuning_Guide_for_Mellanox_Network_Adapters_v1.6.pdf
[6] OSU Benchmark Documentation, http://mvapich.cse.ohio-state.edu/benchmarks/
[7] Daniel Hugo Cámpora Pérez, Rainer Schwemmer, Niko Neufeld, *Protocol-Independent Event Building Evaluator for the LHCb DAQ System*, Transactions of Nuclear Science, 2014