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

by Daniel Hugo Campora Perez, Niko Neufeld, Adam Otto (adam.otto@cern.ch), Flavio Pisani, Rainer Schwemmer

LHCb DAQ Upgrade

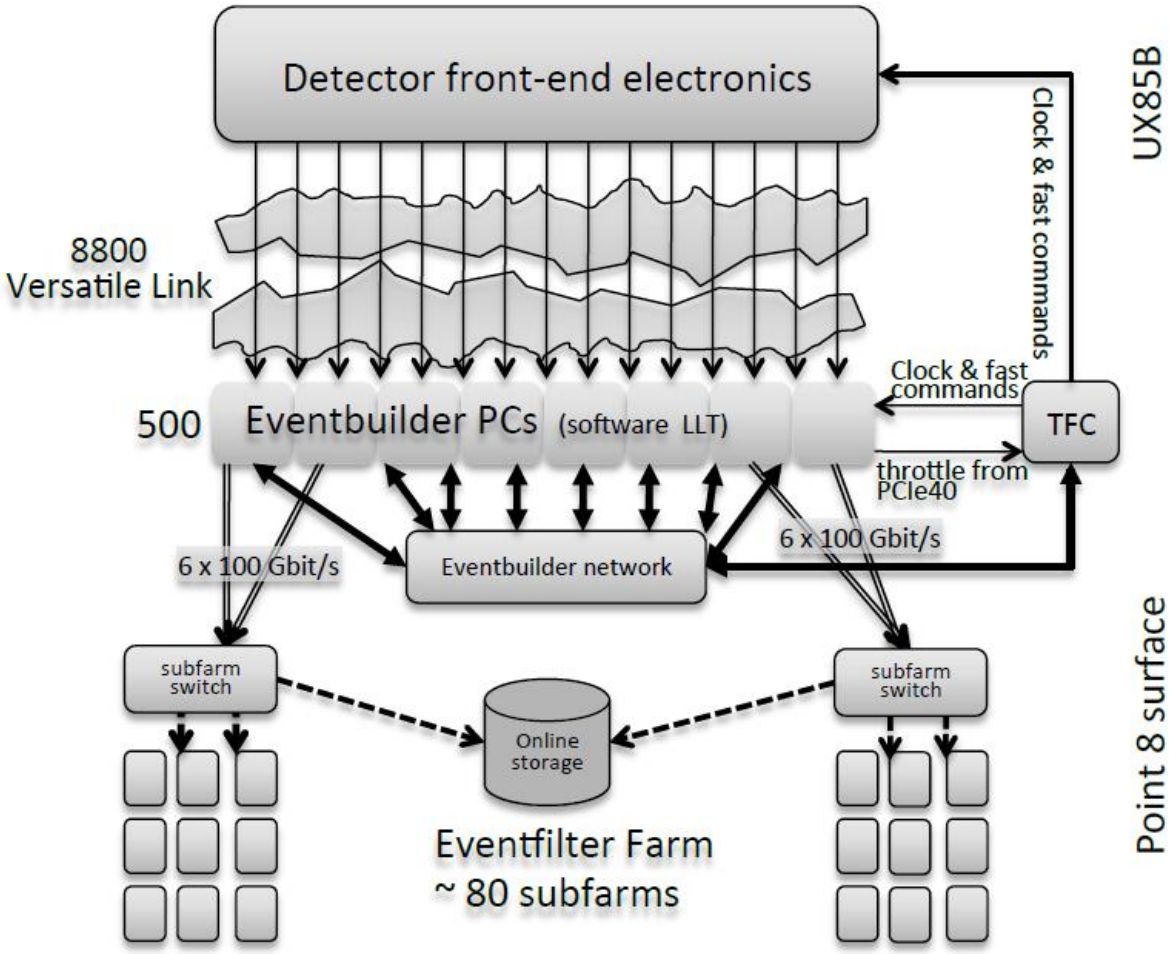
- When? Long Shutdown 2
- What?
  - improve detectors and electronics such that the experiment can run at an instantaneous luminosity of 2*10^33 cm^-2 s^-1
  - increase the event-rate from 1 MHz to 40 MHz
- Impact:
  - increased the event-size from between 50 and 60 kB to 100 kB
  - aggregated bandwidth of 32 Tbit/s

Table 1. Key paramaters of the LHCb DAQ

	Runs 1 & 2	Run 3
event-size [kB]	50 - 60	100
event-rate [MHz]	1	40
# data sources	313	500
# data sinks	up to 2000	up to 5000

Reference [1]

LHCb DAQ Upgrade - Architecture



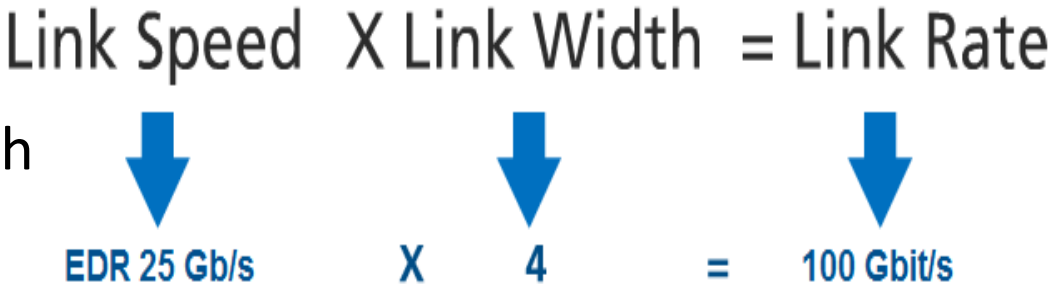
Possible 100Gbps solutions

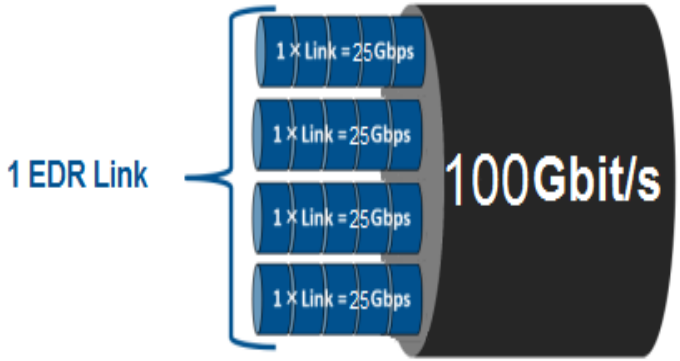
- Intel® Omni-Path
- 100G Ethernet
- EDR InfiniBand

InfiniBand 100Gbit/s

Mellanox ConnectX 4:

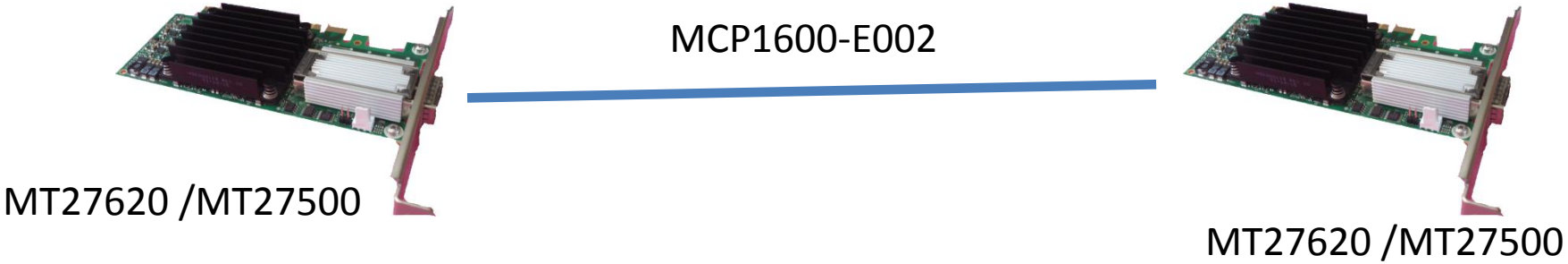
- VPI – supported both technologies:
  - -Ethernet
  - -InfiniBand
- PCIe Gen 3.0 compliant
- 256 to 4Kbyte MTU



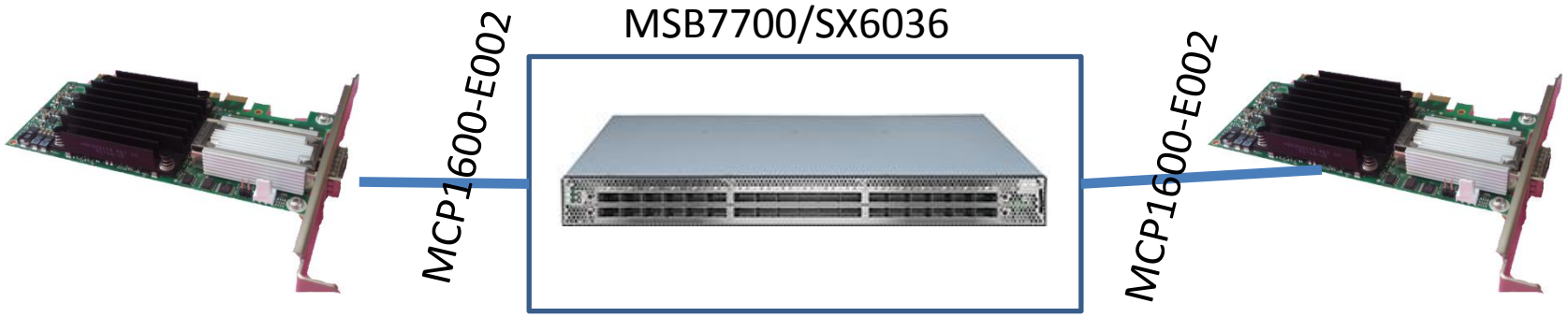


Study of 100G Infiniband

NIC-to-NIC:



NIC-Switch-NIC:



Testbed Details

- 2 x Intel(R) Xeon(R) CPU E5-2650 v3 @ 2.30GHz
  - 10 CORES
  - 20 THREADS
- 64GB RAM
- NIC :
  - Mellanox Technologies MT27620 Family [ConnectX-4]
  - Mellanox Technologies MT27500 Family [ConnectX-3]
- SWITCH:
  - MSB7700
  - SX6036

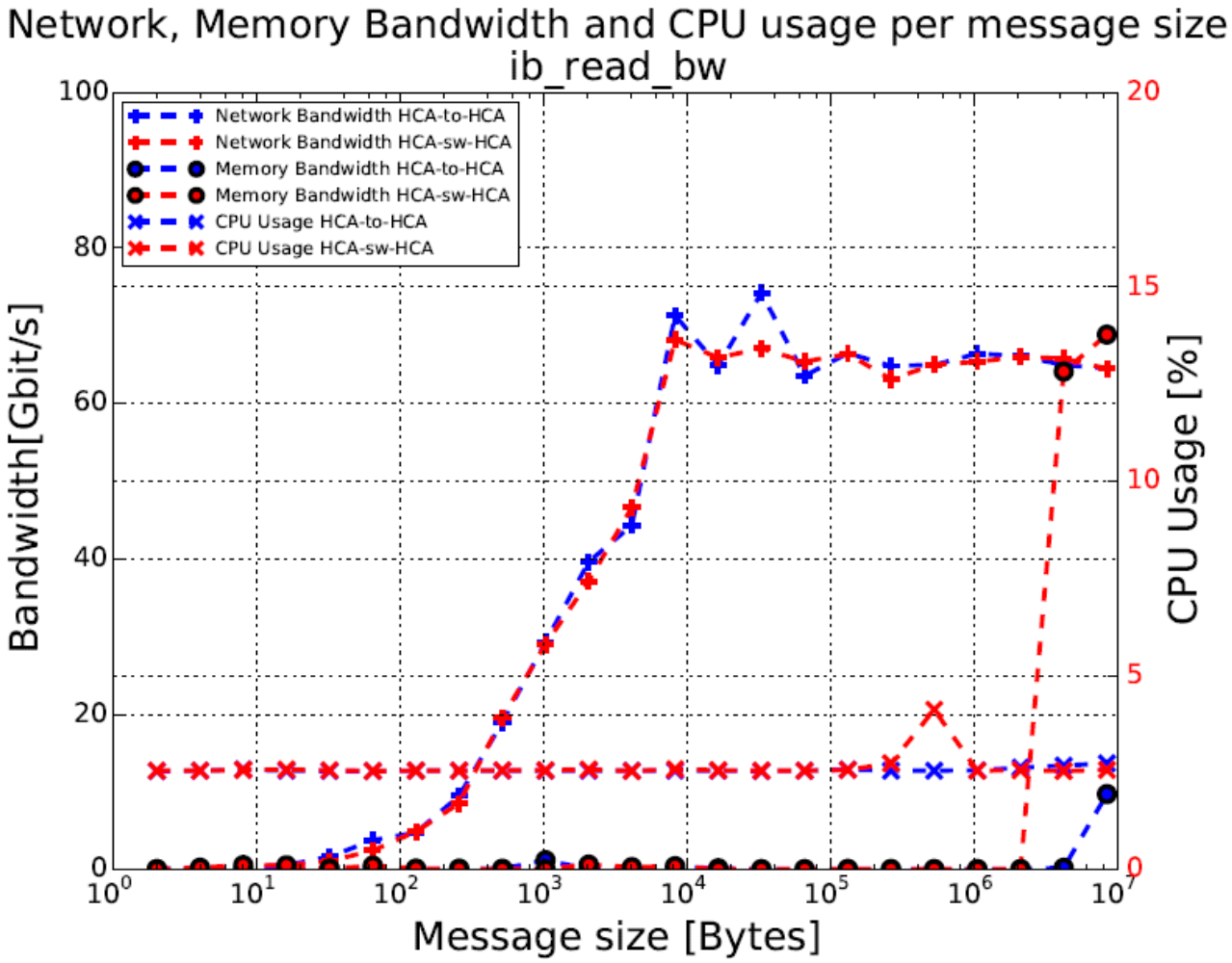
Testbed Details

- Scientific Linux CERN SLC release 6.6
  - kernel: 2.6.32-504.el6.x86_64
- HCA: Mellanox OFED 2.4-1.0.1
  - Alfa version of driver for ConnectX4
- Switch: MLNX OS 3.4.1102
  - Beta version
- OFED Benchmarks : Version: 5.33
- MPI : openmpi-1.8.4

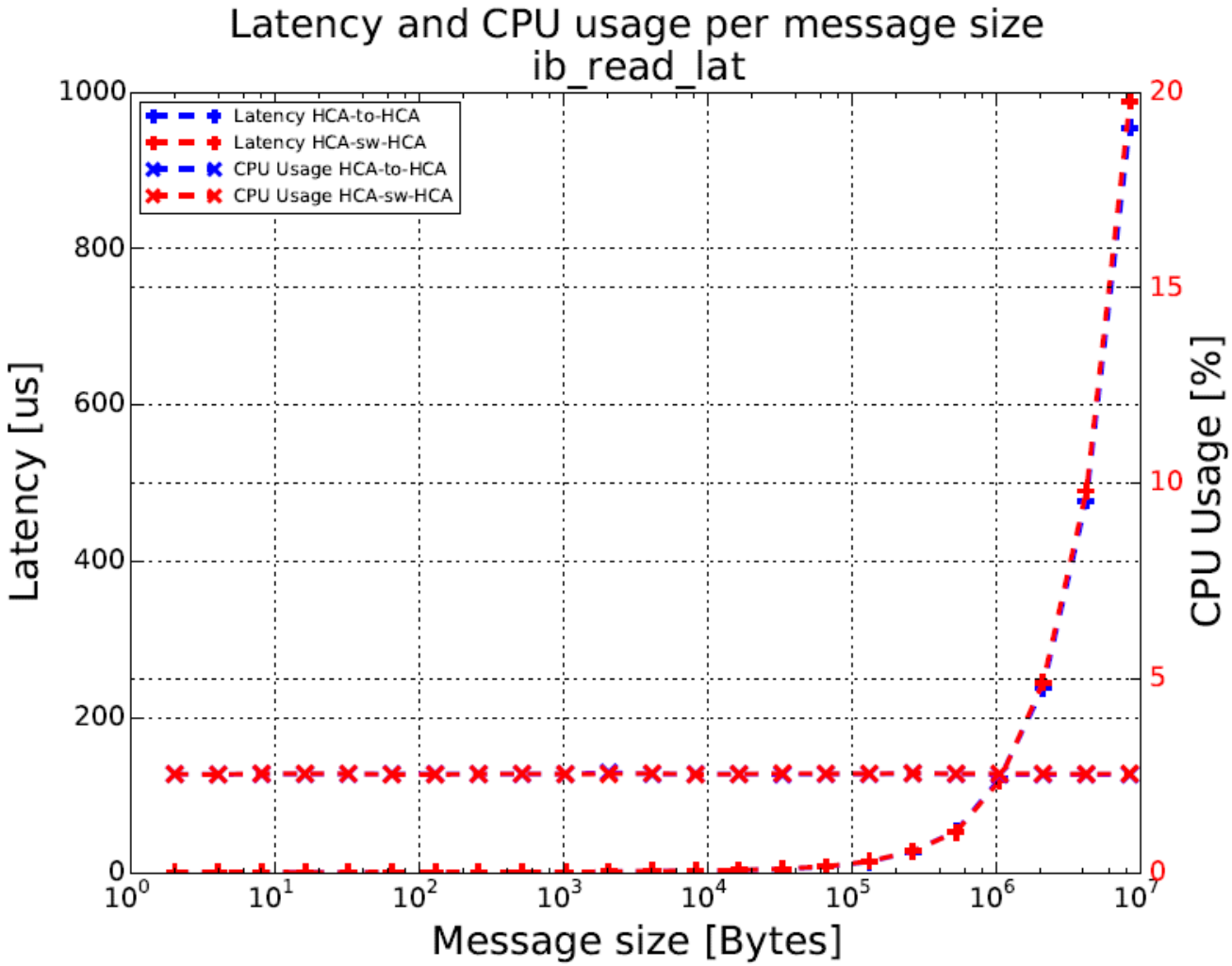
Measurement methodology

- Bandwidth/Latency:
  - standard infiniband OFED benchmarks ran for 40 seconds per message size
  - result = average bandwidth/latency for period of 40s
- CPU Usage:
  - measured with sar during transmission
  - result = average usage for time of transmission
- Memory Bandwidth
  - measured with Intel PCM during transmission
  - result = average memory bandwidth for time of transmission

Measurement of bandwidth, without tuning.



Measurement of latency, without tuning.



Tuning

BIOS Settings:

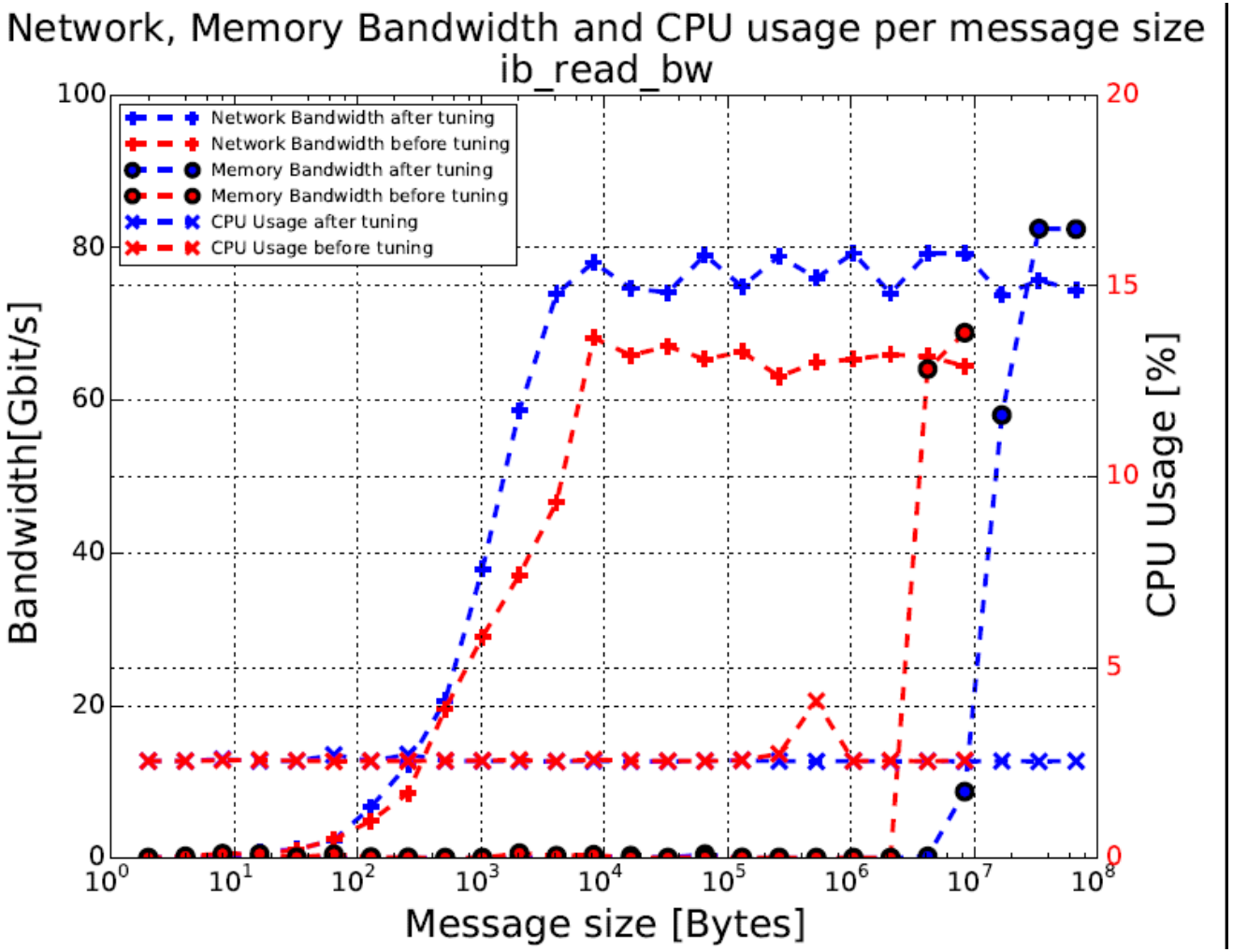
- Power profile: Maximum performance
- C-States disabled
- Turbo mode: enabled
- Memory speed: Max performance

Linux optimization:

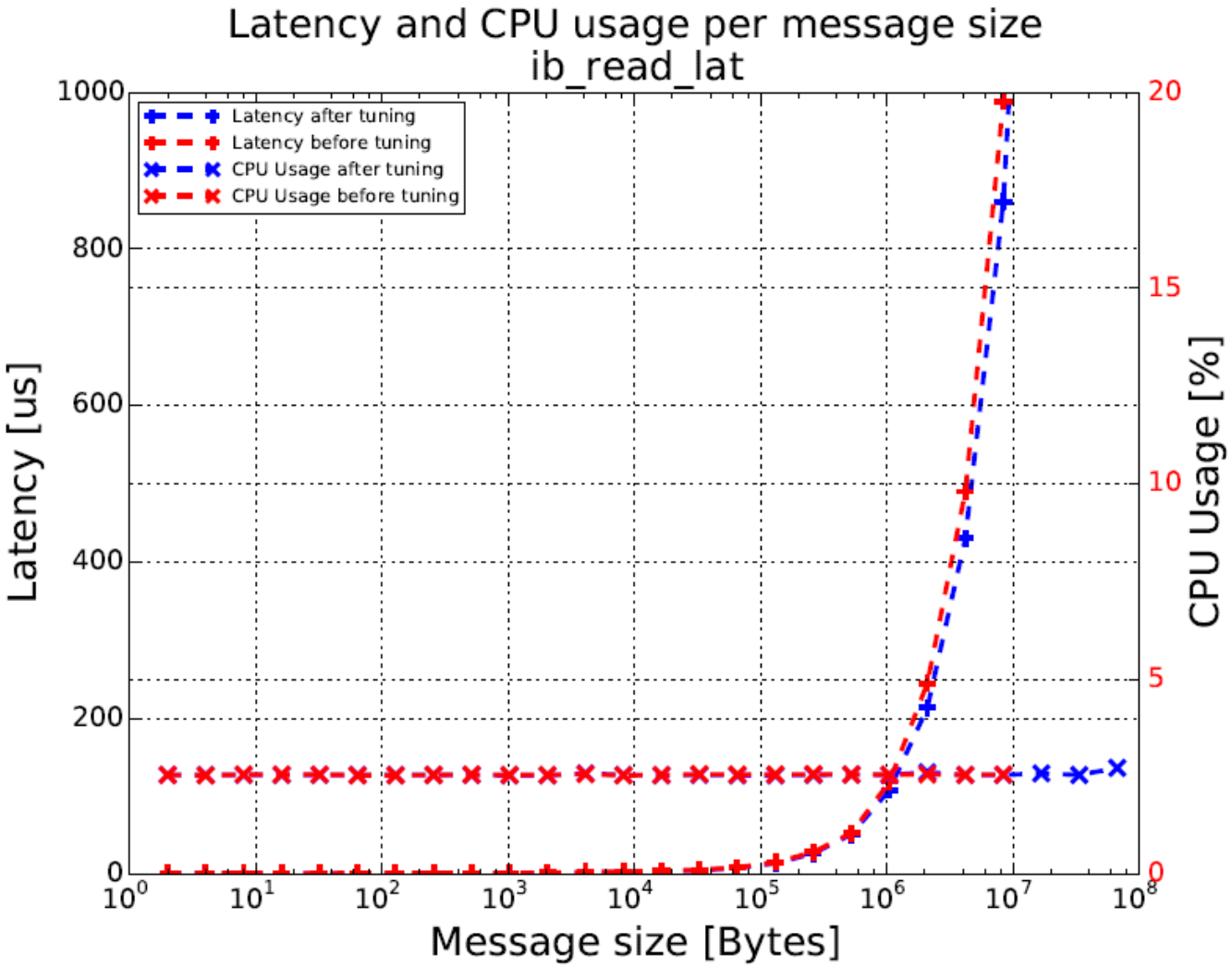
- MLNX_AFFINITY

http://www.mellanox.com/related-docs/prod_software/Performance_Tuning_Guide_for_Mellanox_Network_Adapters_v1.6.pdf

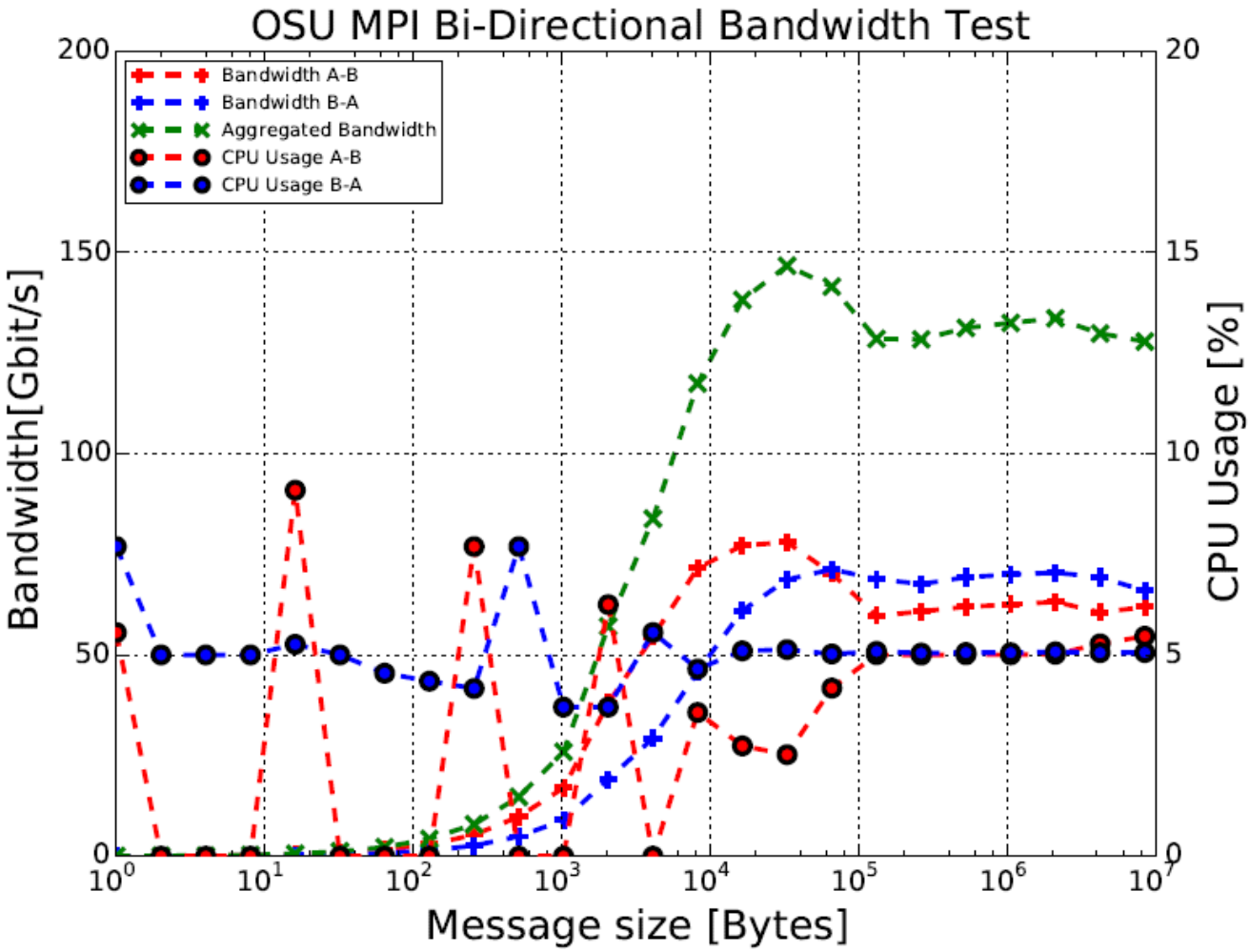
Measurement of bandwidth, after tuning.



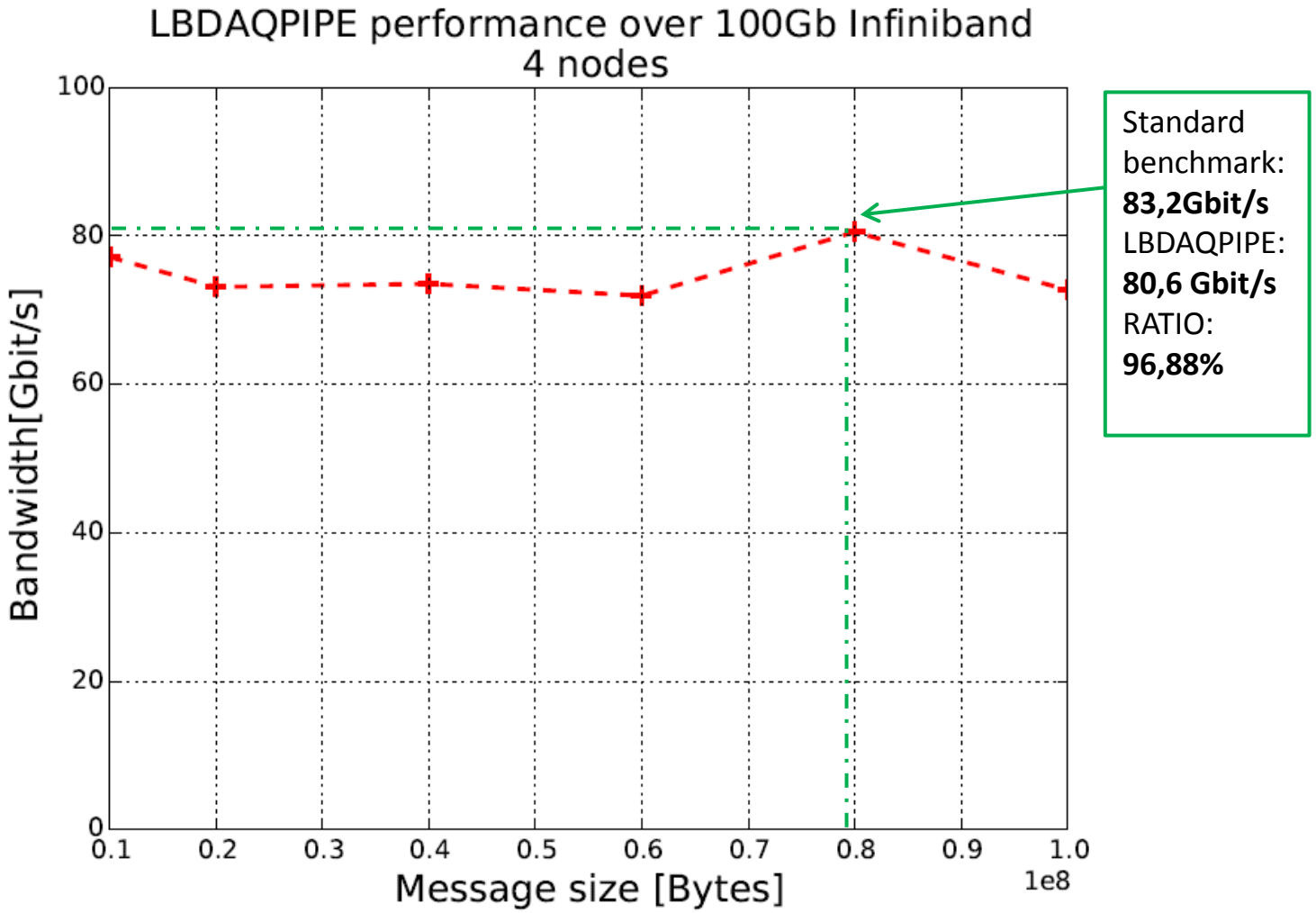
Measurement of latency, after tuning.



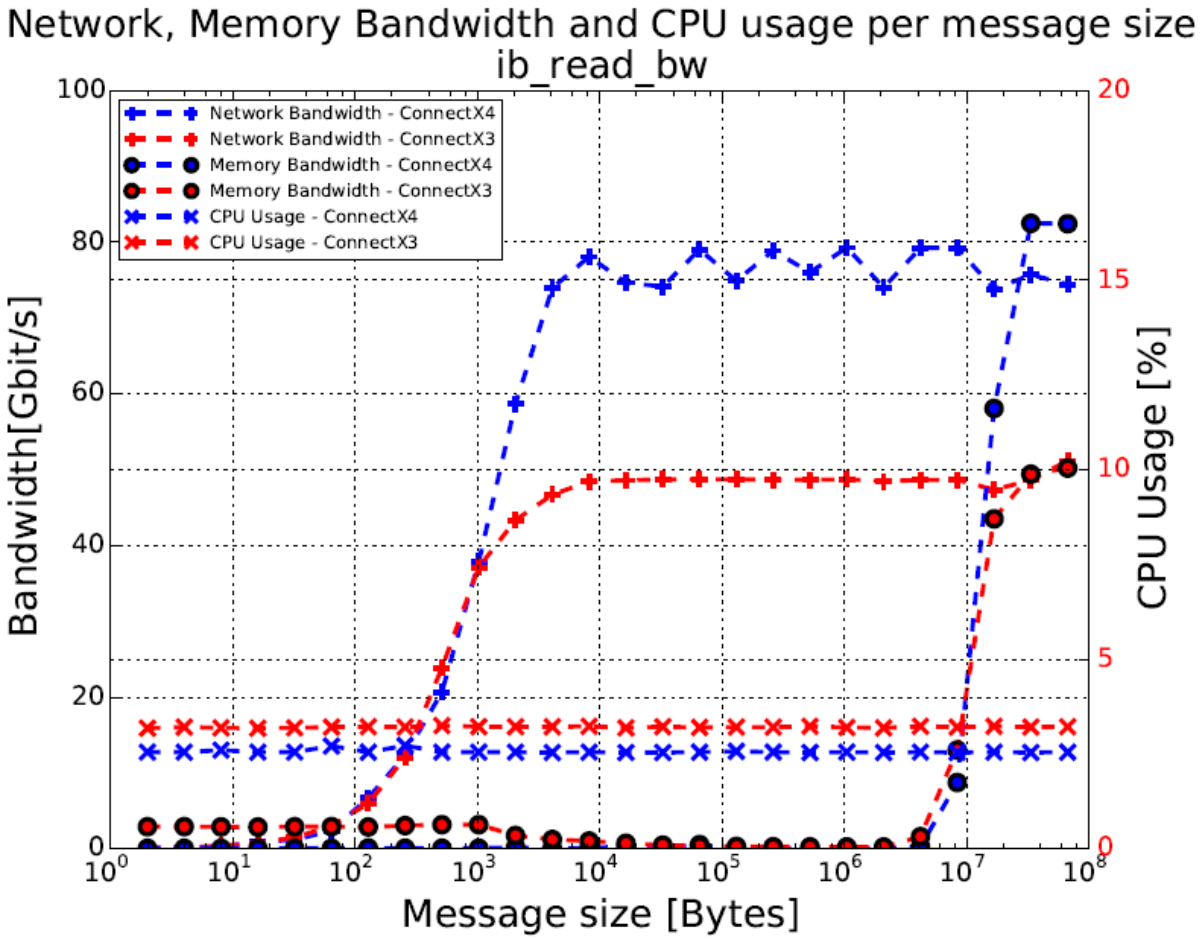
OSU Benchmark



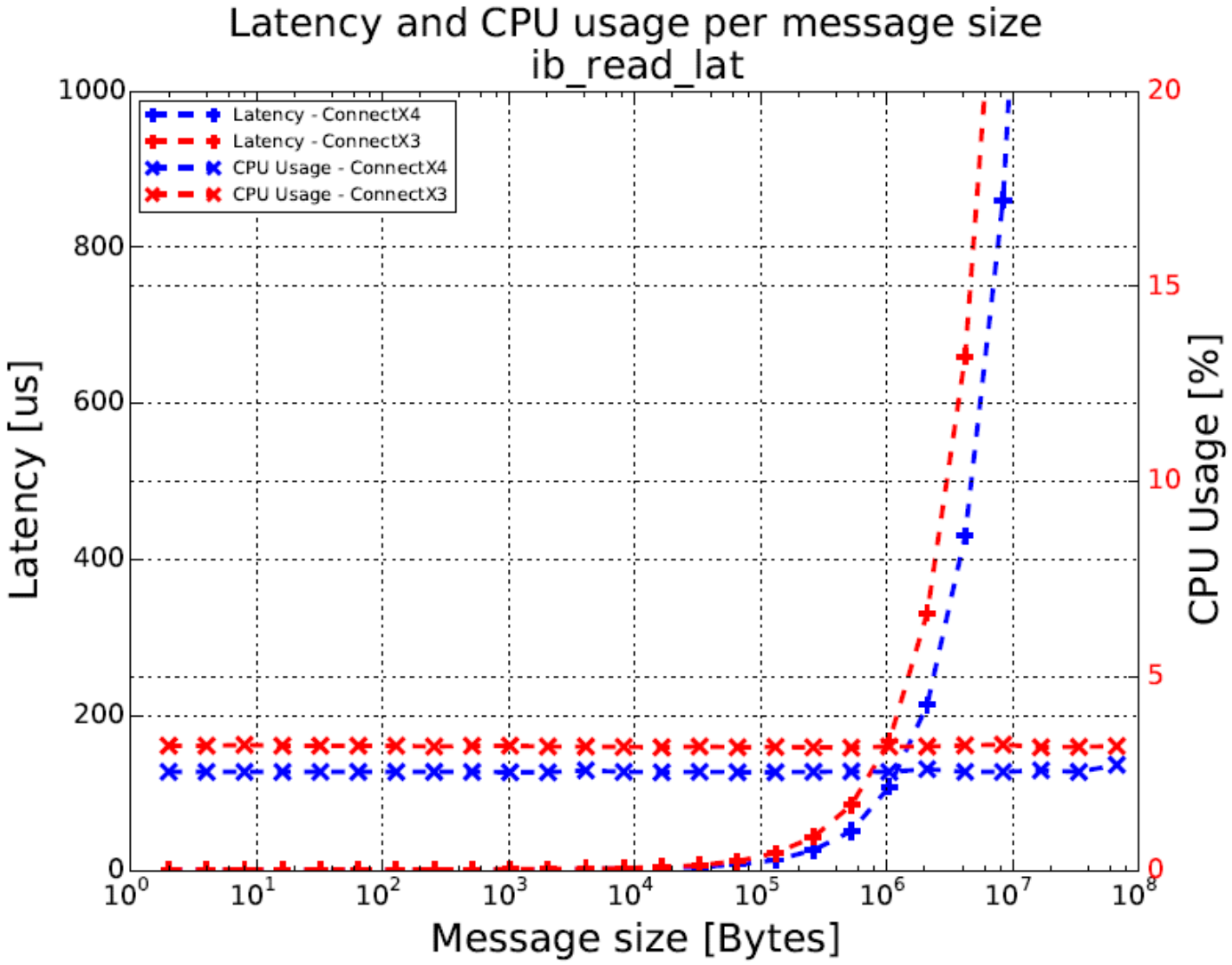
LBDAQPIPE [2] - RDMA



ConnectX 3 vs. ConnectX 4



ConnectX 3 vs. ConnectX 4



Conclusion:

- Achieved 80Gbit/s with standard OFED benchmarks.
- 150Gbit/s Bi- Bandwidth performance
- LBDAQPIPE runs at 80Gbit/s which is so far the best obtained result.
- General results are very promising and we should expect a bandwidth to get closer to declared 100Gbit/s, once stable version of driver is released .
- Unfortunatelly, alfa version of driver doesn’t support Ethernet mode.

Thank you

References

[1]DAQ Architecture for the LHCb Upgrade, Guoming Liu and Niko Neufeld, CHEP 2013

[2] Protocol-Independent Event Building Evaluator for the LHCb DAQ System, Daniel Hugo Campora Perez, Transactions of Nuclear Science 2014