High throughput data acquisition with InfiniBand on low power architectures

Matteo Manzali
INFN CNAF - Università degli Studi di Ferrara

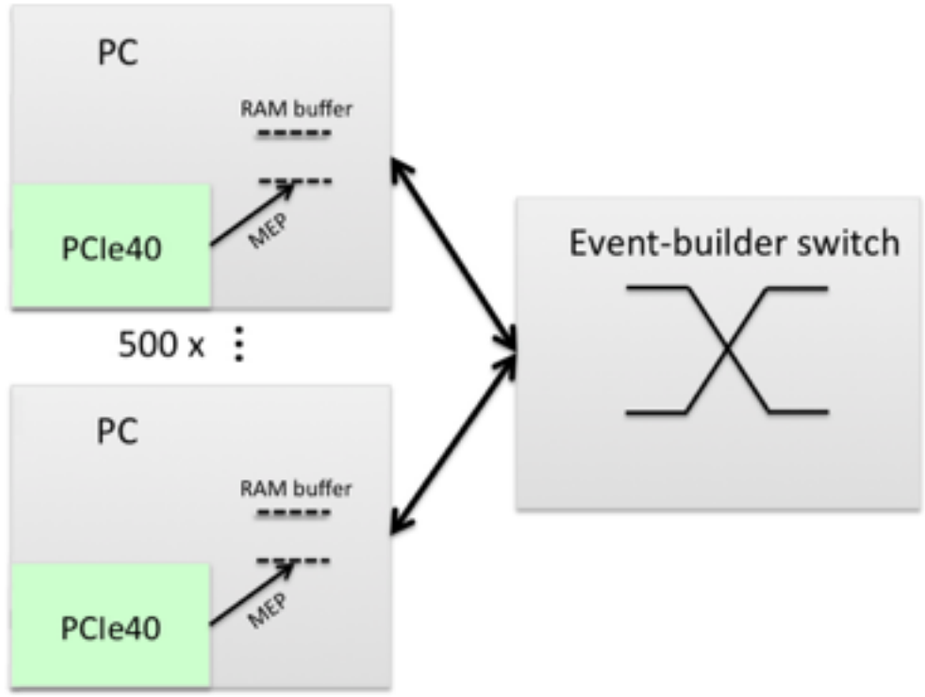
The LHCb experiment

- The LHC beauty (LHCb) experiment is one of four large experiments based at the CERN laboratory near Geneva (Switzerland).

- It will undergo an upgrade during the second long shutdown of the Large Hadron Collider (2018 - 2019)

- Trigger-less readout with expected rate increased from 1 MHz to 40 MHz

- ~500 DAQ nodes send and receive data in one-to-many pattern

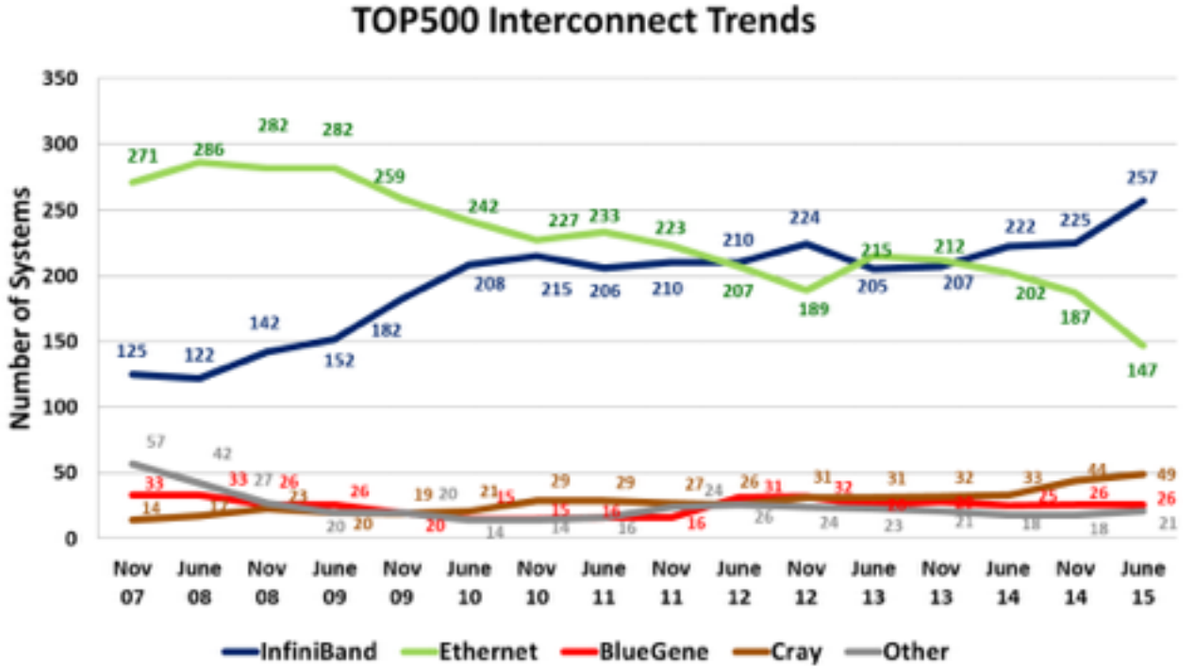


- Foreseen aggregated bandwidth of ~ 32 Tb/s

- In order to reach this performances a high-speed network is required

The InfiniBand standard

- A computer-networking communications standard that feature very high bandwidth and low latency (widely used in High Performance Computing)
- Low CPU utilization with RDMA (Remote Direct Memory Access)
- Constant speed evolution and cost-effective technology

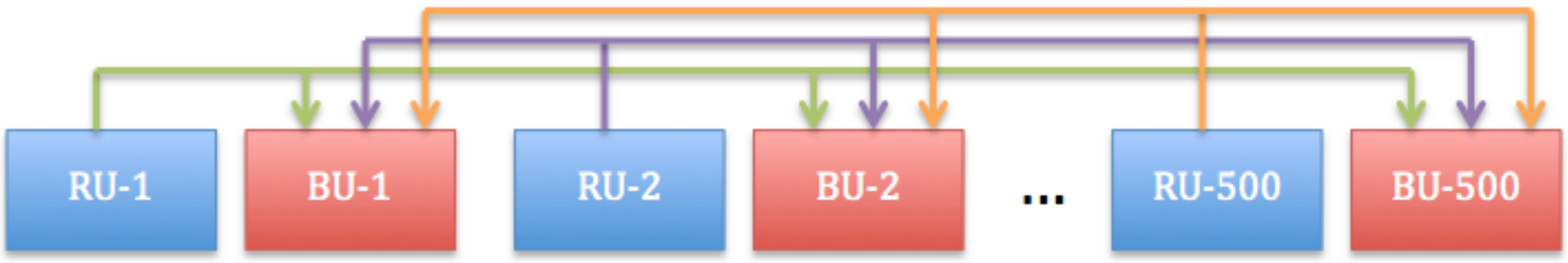


What are the “verbs”?

- Verbs are a low level description for RDMA programming and provide best performance
- Any other level of abstraction over verbs may harm the performance
- Same API for all RDMA-enabled transport protocols:
    - InfiniBand
    - RDMA Over Converged Ethernet (RoCE)
    - Internet Wide Area RDMA Protocol (iWARP)

The Event Builder Software

- It is designed to simulate the event building on InfiniBand based networks
- It relies on the verbs library to perform RDMA operations
- It is composed of two distinct logical components, the Readout Unit (RU) and the Builder Unit (BU):
    - Each RU receives data from a generator, creates the event fragments and ship them to receiving BU in a many-to-one pattern.
    - Each BU gathers event fragments together to generate full events.



Testbed setup

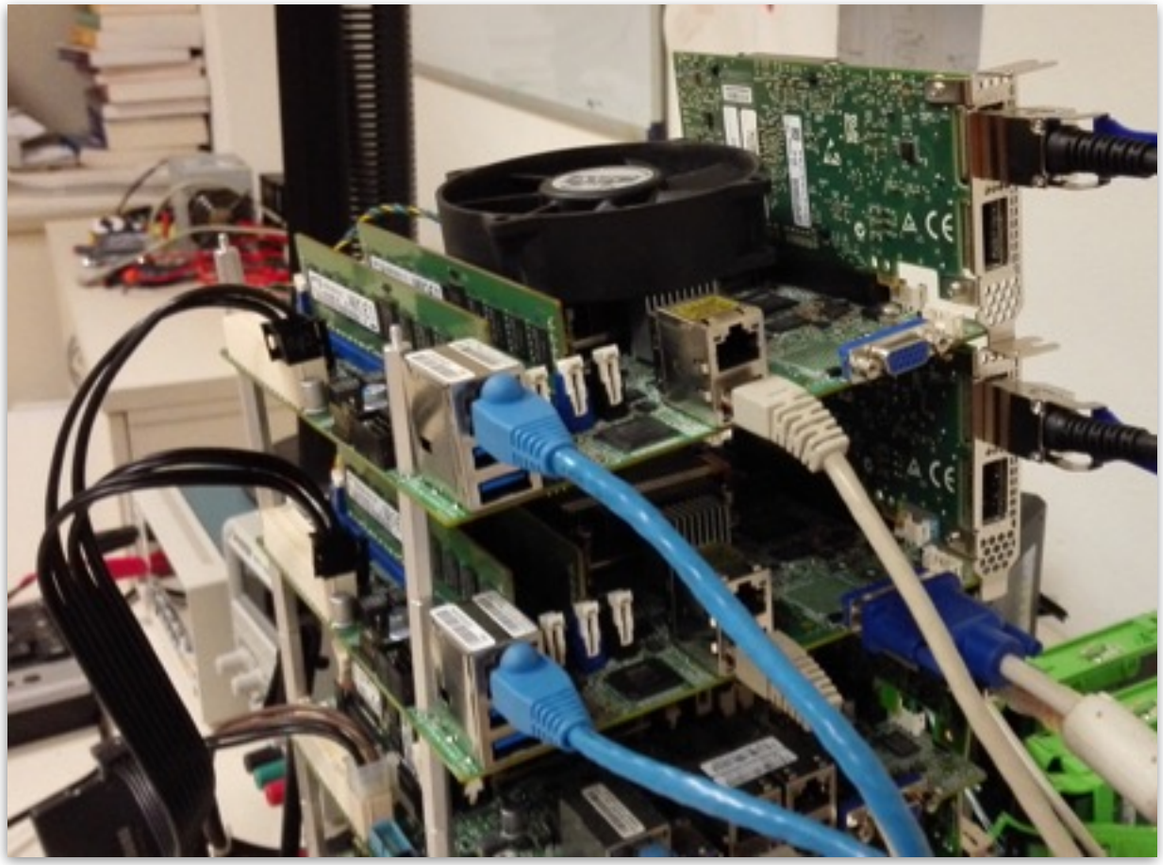
- Tests are performed in collaboration with the COSA project (see previous presentation) that provided the testbed. The architectures tested are:
    - Intel Xeon D-1540 (http://goo.gl/hqE6ai)
    - Intel Atom C2750 (http://goo.gl/4KN19A)
- First test: benchmark with the ib_write_bw tool (provided by the OFED package)
- Second test: execution of the Event Builder (comparing bandwidth and power consumption)
- Results of each tests are compared with those obtained on a dual socket Intel Xeon Processor E5-2683v3 (http://goo.gl/MRwCUL)

CPUFreq and c-states

- The CPUFreq governor allows the clock speed of the processor to be adjusted on the fly:
    - ondemand governor: it dynamically sets the maximum frequency when system load is high and minimum frequency when the system is idle
    - performance governor: forces the CPU to use the highest possible clock frequency
- c-states allows systems to save power by partially deactivating CPU parts that are not in use:
    - more CPU units are stopped, reducing the voltage or even completely shutting down, and more energy is saved
    - but more time required for the CPU to "wake up" and be again 100% operational

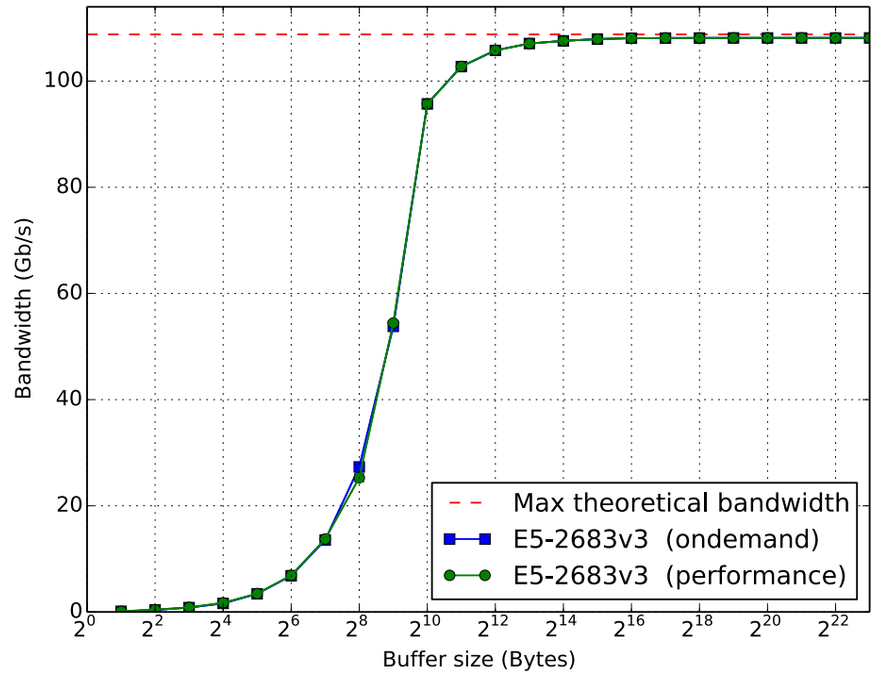
Intel Xeon D-1540

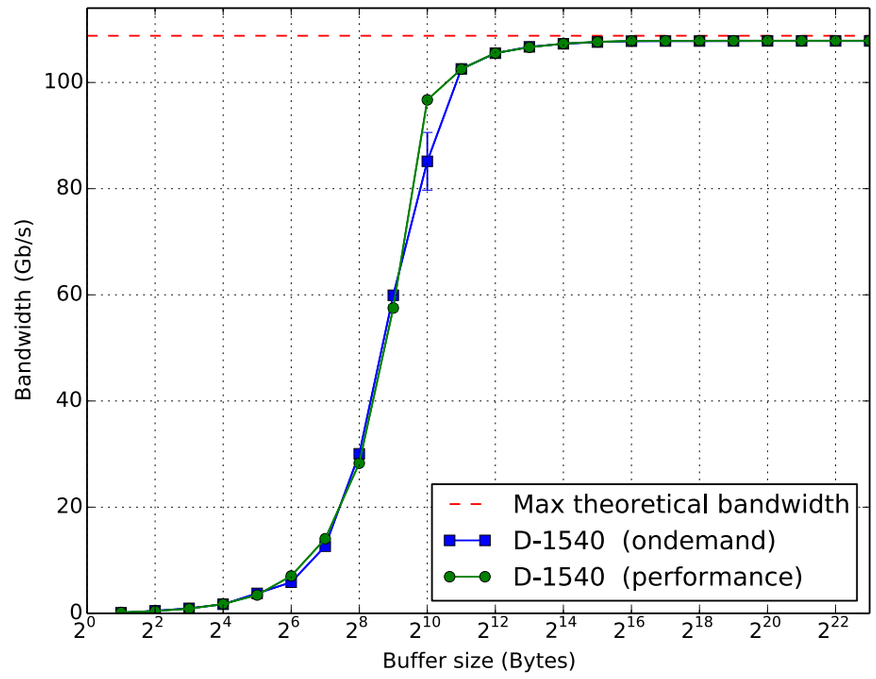
- Processor: 8 core Intel x86 SoC with 2 threads per core
- Development board: Supermicro X10SDV-F
- Two nodes connected back to back with Mellanox InfiniBand FDR cards:
  - 56 Gb raw bandwidth
  - 54,3 Gb theoretical bandwidth (64b/66b encoding)
  - Requires 1 slot PCIe 3.0 16x



ib_write_bw benchmark

- The ondemand governor doesn't affect performances
- Plots of both the architectures are comparable:
  - The D-1540 reaches the 99.35 % of the theoretical bandwidth
  - The E5-2683v3 reaches the 99.57 % of the theoretical bandwidth
- Plateau reached with 16KB of buffer dimension



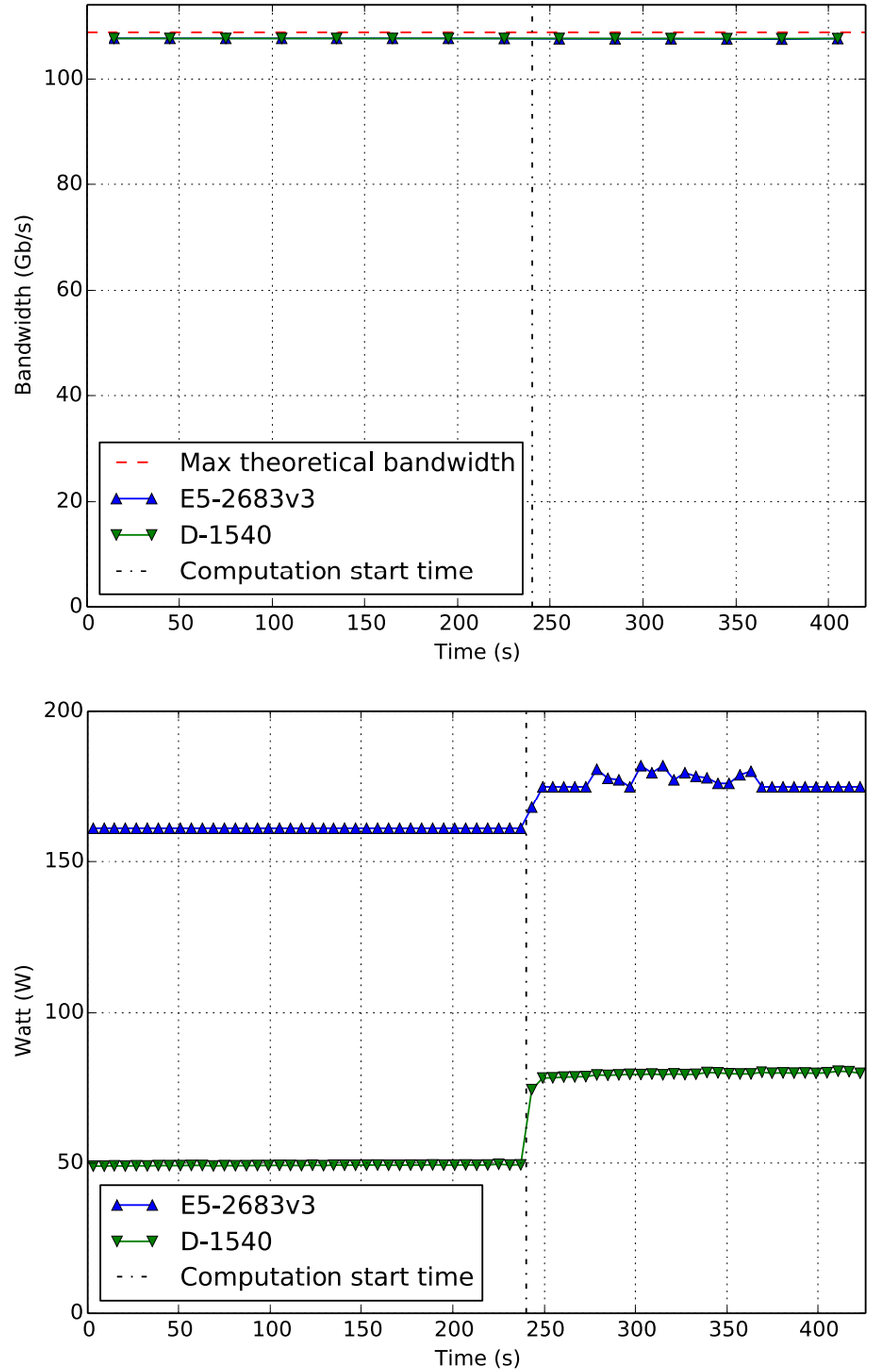


Event Builder test

- In the second part of the test a pure computation process is started on four cores (in order to simulate a software trigger)

- The performances of the Event Builder are comparable, but the D-1540 requires a third of the power consumption of the E5-2683v3

	E5-2683v3	D-1540
Idle power consumption	80.78 W	28.23 W
EB power consumption	161.00 W	49.02 W
EB power consumption with computation	176.54 W	79.12 W
Max temperature	56.0 C	59.0 C
Average bandwidth	107.63 Gb/s	107.65 Gb/s



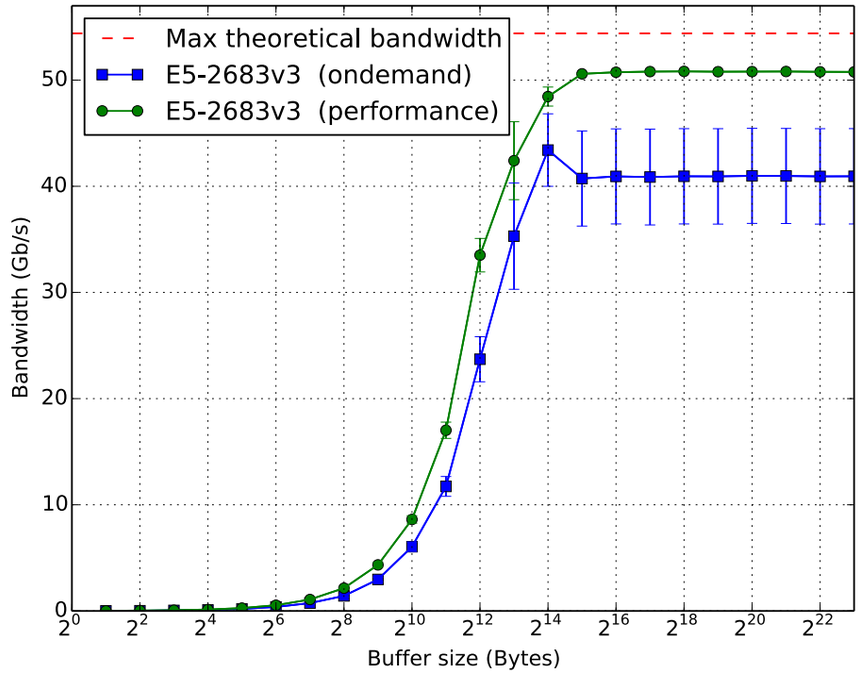
Intel Atom C2750

- Processor: 8 core Intel x86 SoC
- Development board: Supermicro A1SAi-2750F
- Two nodes connected back to back with QLogic InfiniBand QDR cards:
  - 40 Gb raw bandwidth
  - 27,2 Gb bandwidth declared by QLogic
  - Requires 1 slot PCIe 3.0 8x (with PCIe 2.0 it is foreseen a 20% of bandwidth degradation)

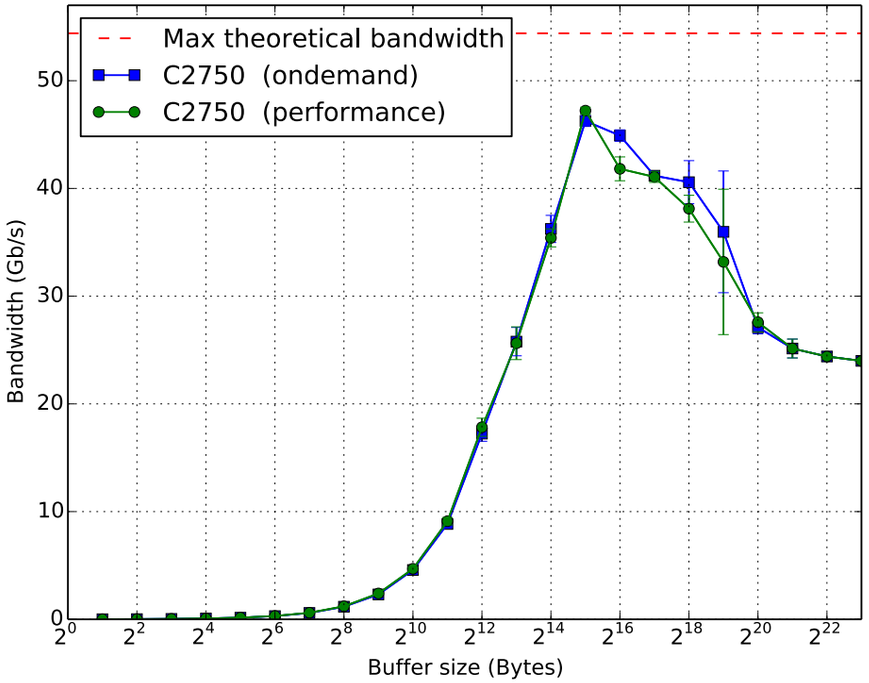
ib_write_bw benchmark

- The ondemand governor affects the performances of the E5-2683v3 with QDR cards

- The performances of the C2750 are worst in both cases (relevant degradation with sizes greater than 32KB)



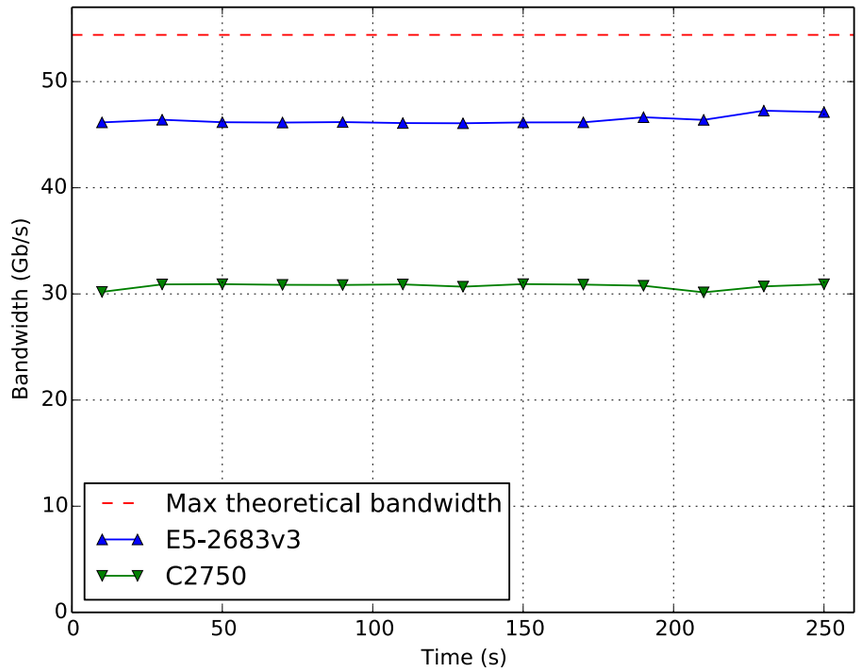
- This can be caused by several factors:
  - PCIe 2.0
  - CPU
  - More investigation is needed…



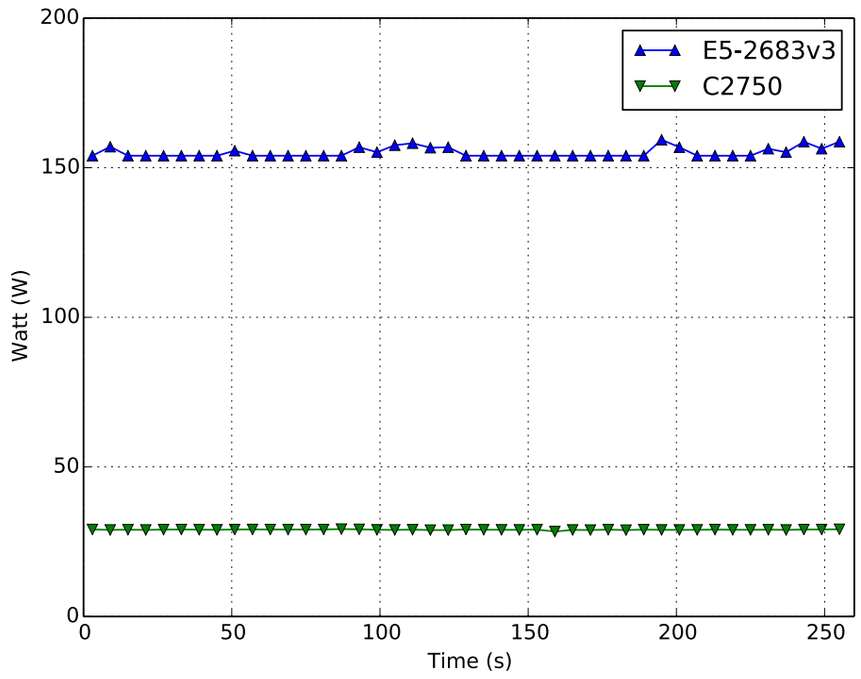
Event Builder test

- The Event Builder on the C2750 reaches the 65.71% of the performances of the E5-2683v3, but it requires a 18,73 % of the power consumption

- The bandwidth becomes really unstable running a computation process on the C2750



	E5-2683v3	C2750
Idle power consumption	77.46 W	18.20 W
EB power consumption	154.44 W	28.93 W
Max temperature	52.0 C	37.0 C
Average bandwidth	46.38 Gb/s	30.74 Gb/s



Conclusions

- The Intel Xeon D-1540 seems a really interesting processor for high-throughput data acquisition purposes.

- It brings all the functionalities of the XEON family processors but reducing costs and power consumption.

- The Intel Atom C2750 can't compete with the high-performance XEON family processors.

- It is still interesting for data acquisition purposes in case the requirements are not so high, due to its extremely low power consumption.