High Throughput Data Acquisition with InfiniBand on x86 Low-power Architectures for the LHCb Upgrade

D. Cesini, A. Ferraro, A. Falabella, F. Giacomini, M. Manzali, U. Marconi, N. Neufeld, S. Valat, B. Voneki

***Abstract*—The LHCb Collaboration is preparing a major upgrade of the detector and the Data Acquisition (DAQ) to be installed during the LHC-LS2. The new Event Builder computing farm for the DAQ requires about 500 nodes, and have to be capable of transporting order of 32 Tbps. The requested performance can possibly be achieved using high-bandwidth data-centre switches and commodity hardware. Several studies are ongoing to evaluate and compare network and hardware technologies, with the aim of optimising the performance and also the purchase and maintenance costs of the system. We are investigating if x86 low-power architectures can achieve equivalent performance as traditional servers when used for multi gigabit DAQ. In this talk we introduce an Event Builder implementation based on InfiniBand network and show preliminary tests with this network technology on x86 low-power architectures, such as Intel Atom C2750 and Intel Xeon D-1540, comparing measured bandwidth and power consumption.**

I. Introduction

THE LHCb experiment is one of the four main experiments operating at the Large Hadron Collider at CERN. It will undergo a major upgrade during the second long shutdown (2018 - 2019), aiming at collecting an order of magnitude more data the possible with the present detector. The upgraded detector foresees a full software trigger, running at the LHC bunch crossing frequency of 40 MHz. A new high-throughput PCIe Generation 3 based read-out board, named PCIe40, has been designed on this purpose [1]. The read-out board will allow an efficient and cost-effective implementation of the Data Acquisition System by means of high-speed PC networks [2]. The final foreseen read-out system for the upgrade is shown in Figure 1, in which each of the 500 nodes of the Event Builder cluster will aggregate the data coming from the detector, building events with an expected bandwidth of 100 Gb/s per node per direction. Moreover, in case of no data filtering performed by the event-building phase, the completed events will be sent to an event filter farm with the same bandwidth of 100 Gb/s.

Even using commodity PCs and commercial network technologies, the Event Builder cluster will have a relevant cost in terms of hardware, power consumption and cooling. The aim of this work is to investigate a possible use of upcoming low-power architectures for the event-building purpose of LHCb and for high-throughput data acquisition in general. Several studies performed by the LHCb collaboration are ongoing in order to establish which will be the most suitable candidate as network technology for the event-building [3]. For this work the InfiniBand network technology has been chosen [4]. we designed and implemented an Event Builder (EB) software prototype based on the ”verbs” [5], a set of structures and functions that allow to access the RDMA capabilities of supported network devices.

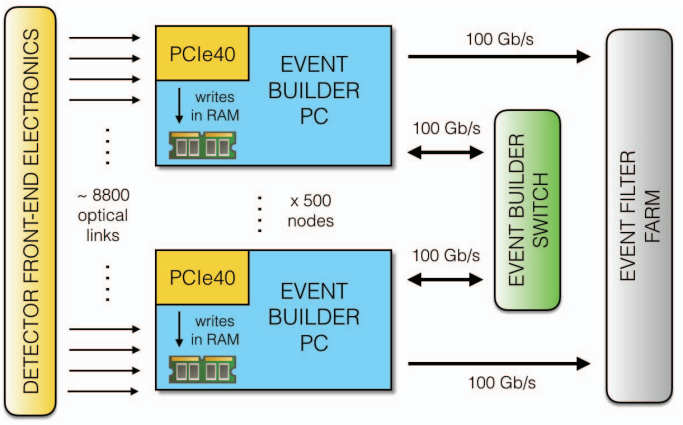

Fig. 1. The architecture of the upgraded LHCb readout-system.

II. Testbed Setup

Measurements on low-power architectures are performed in collaboration with the Computing On SoC Architectures project (COSA) [6], that provided the testbed. The tested architectures are the Intel Atom C2750 [7] and the Intel Xeon D-1540 [8], two low-power x86 processors designed by Intel for its server line. For each architecture, the testbed is composed of two nodes of the same type connected back to back with InfiniBand Host Channel Adapters (HCAs). The available InfiniBand HCAs were the QLogic QDR and the Mellanox FDR. The QLogic QDR InfiniBand HCA requires a PCI-2 slot with 8 lanes and it is declared a maximum bandwidth of 27.2 Gb/s. The Mellanox FDR InfiniBand HCA requires a PCIe-3 with 16 lanes and foresees a maximum bandwidth of 54.3 Gb/s. Depending on the PCIe capacity of each architecture, different InfiniBand HCAs were used. The results are compared to those obtained on a testbed composed of two Intel Xeon based standard servers. These servers mount a dual socket Intel Xeon E5-2683 v3 [9] processor, that supports PCIe Gen 3 interconnects with a maximum of 40 lanes. The Thermal Design Power (TDP) of the Intel Xeon E5-2683 v3 is of 120 W. Due to the high PCIe capacity of this kind of processor, the Xeon servers support both the two available InfiniBand HCAs.

Manuscript received May 31, 2016.

D. Cesini, A. Ferraro, A. Falabella, F. Giacomini and U. Marconi are with INFN, Italy.

M. Manzali is with University of Ferrara, Italy, and INFN, Italy.

N. Neufeld, S. Valat and B. Voneki are with CERN, Switzerland.

978-1-5090-2014-0/16/$31.00 ©2016 IEEE

Each test consists in running the Event Builder software in a 2-nodes setup. Besides the bandwidth, the other monitored parameters are the temperature and the power consumption before and while running the software. The power consumption measurements refer to the overall power consumption of each node including processor, motherboard, memory, disk and InfiniBand network card.

III. Intel Atom C2750

The Intel Atom C2750 is a low-power System on Chip (SoC) designed by Intel for microservers and manufactured on 22nm technology. It supports PCIe-2 interconnects with a maximum of 16 lanes. The TDP of the Intel Atom C2750 is of 20 W. The motherboard that is mounting this processor provides a slot PCIe-2 with 8 lanes. In this testbed the two Atom nodes are connected together with QLogic QDR InfiniBand cards, due to the missing PCIe-3 support required by the more powerful Mellanox FDR InfiniBand cards.

Results obtained running the Event Builder software on the Xeon servers and on the Atom nodes with QDR InfiniBand connectivity are shown in Figure 2, where the bandwidths are plotted as a function of time. The figure unveils the performance inefficiency of the Atom nodes, that reach an average bandwidth of 15.37 Gb/s with respect to the 23.19 Gb/s reached by the Xeon servers. On the other hands the power consumption measurements turn the tables: the Atom nodes consume about the 18,73% with respect to the Xeon servers (28.93 W against 154.44 W), as illustrated in Figure 3.

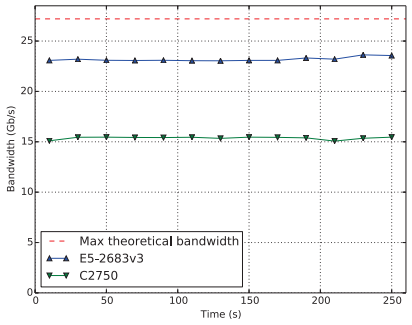

Fig. 2. Bandwidth running the EB on nodes with QDR cards.

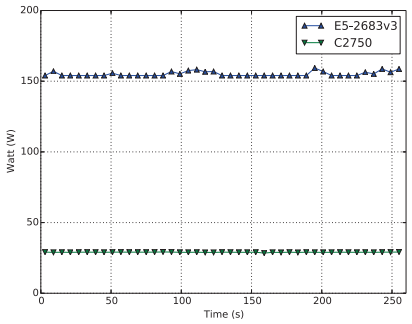

Fig. 3. Power consumption running the EB on nodes with QDR cards.

IV. Intel Xeon D-1540

The Intel Xeon D-1540, also called XeonD in the following, brings the performance of Intel Xeon processors into a lower-power SoC. It is a low-power SoC designed for servers and manufactured on 14nm technology. Being more a server than a pure low-power architecture, the Intel Xeon D-1540 has a higher power consumption with respect to the Intel Atom C2750, with a TDP of 45 W. The motherboard that is mounting this processor provides a PCIe-3 slot with 16 lanes. In this testbed the two XeonD nodes are connected together with Mellanox FDR InfiniBand cards.

The measurements of the bandwidth running the Event Builder software include the same tests performed with the Atom and an additional test which adds a pure computation process running over 4 cores on each node during the second half of the run. The aim of this added test is to investigate the possibility of the Event Builder to perform computations such as a pre-analysis of the events prior to send them to the Event Filter Farm.

The bandwidths as a function of time are shown in Figure 4, where the black vertical line indicates the computation process start time. Both the Xeon servers and the XeonD nodes reach about the 99.12% of the theoretical bandwidth allowed by the FDR InfiniBand cards, keeping the bandwidth stable even with a computation process running. These performances were achieved with very different power consumptions, with the XeonD nodes that require about a third of the energy consumed by the Xeon servers, as illustrated in Figure 5.

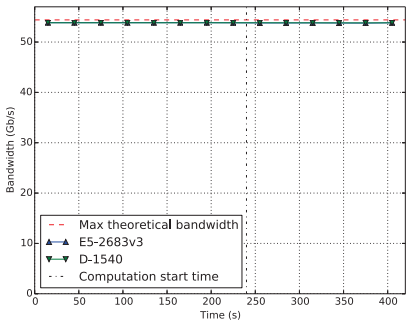

Fig. 4. Bandwidth running the EB on nodes with FDR cards.

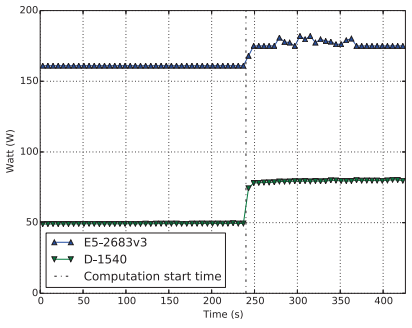

Fig. 5. Power consumption running the EB on nodes with FDR cards.

V. Conclusion

Each node of the Event Builder foreseen by the upgrade of the LHCb experiment requires at least two PCIe-3 slots of 16 lanes, one for the PCIe40 readout board and the other one for the network card needed to communicate with the event-building network. At the time of writing, none of the tested low-power architectures is suitable for the event-building of LHCb, due to the lack of enough PCI-e capability.

Despite that, the Intel Xeond D-1540 seems a really interesting processor for high-throughput data acquisition purposes. It ensures all the functionalities of the Intel Xeon family but reducing costs and power consumption. Tests show that with the adopted testbed this processor achieves comparable results in terms of bandwidth with respect to the standard server. Further studies on upcoming processors belonging to the Xeon D-1500 family should be performed in case Intel will upgrade the PCIe capacity of these products.

On the other hand, the Intel Atom C2750 can't compete with the high performance provided by standard Intel Xeon processors. Aside from the missing PCIe-3 support, tests show that this processor cannot achieve the maximum bandwidth allowed by the used HCA, even with the right PCIe slot. For these reasons this processor family cannot be a suitable candidate for the Event Builder. However, its extremely low power consumption makes it still interesting for data acquisition purposes where the bandwidth and computational requirements are less strict than those of the event-building of the LHCb experiment.

References

[1] M. Bellato, G. Collazuol, I. D'Antone, P. Durante, D. Galli, B. Jost, I. Lax, G. Liu, U. Marconi, N. Neufeld, R. Schwemmer and V. Vagnoni, *A PCIe Gen3 based readout for the LHCb upgrade*, Journal of Physics: Conference Series (2104).

[2] LHCb Collaboration, *LHCb Trigger and Online Upgrade Technical Design Report*, CERN-LHCC-2014-016 (2014).

[3] A. Otto, D. Campora, N. Neufeld, R. Schwemmer, F. Pisani, *A first look at 100 Gbps LAN technologies, with an emphasis on future DAQ applications*, 21st International Conference on Computing in High Energy and Nuclear Physics (2015).

[4] R. Buyya, T. Cortes, H. Jin, *An Introduction to the InfiniBand Architecture*, Wiley-IEEE Press (2002).

[5] Open Fabric Alliance (OFED), https://www.openfabrics.org/index.php.

[6] INFN COSA Project - COMPUTING ON SOC ARCHITECTURE, http://www.cosa-project.it/home.html.

[7] Intel Atom Processor C2750, http://ark.intel.com/it/products/77987/Intel-Atom-Processor-C2750-4M-Cache-2_40-GHz.

[8] Intel Xeon Processor D-1540, http://ark.intel.com/it/products/87039/Intel-Xeon-Processor-D-1540-12M-Cache-2_00-GHz.

[9] Super Micro X10SDV-F motherboard, http://ark.intel.com/it/products/81055/Intel-Xeon-Processor-E5-2683-v3-35M-Cache-2_00-GHz.