Assessment of the ALICE O2 readout servers

Filippo Costa[1,*], *Sylvain* Chapeland[1,**], *Konstantinos* Alexopoulos[1,***], and *Ulrich* Fuchs[1,****]

[1]CERN, Experimental Physics Department, Geneva, Switzerland

Abstract. The ALICE experiment at CERN LHC (Large Hadron Collider) is undertaking a major upgrade during LHC Long Shutdown 2 in 2019-2020. The raw data input from the detector will then increase a hundredfold, up to 3.4 TB/s. In order to cope with such a large throughput, a new Online-Offline computing system, called O2, will be deployed. The FLP servers (First Layer Processor) are the readout nodes hosting the CRU (Common Readout Unit) cards in charge of transferring the data from the detector links to the computer memory. The data then flow through a chain of software components until they are shipped over network to the processing nodes. In order to select a suitable platform for the FLP (First Level Processor), it is essential that the hardware and the software are tested together. Each candidate server is therefore equipped with multiple readout cards (CRU), one InfiniBand 100G Host Channel Adapter, and the O2 readout software suite. A series of tests are then run to ensure the readout system is stable and fulfils the data throughput requirement of 42Gbps (highest data rate in output of the FLP equipped with three CRUs). This paper presents the software and firmware features developed to evaluate and validate different candidates for the FLP servers. In particular we describe the data flow from the CRU firmware generating data, up to the network card where the buffers are sent over the network using RDMA. We also discuss the testing procedure and the results collected on different servers.

1 Introduction

The ALICE[1] experiment at CERN LHC (Large Hadron Collider) is undertaking a major upgrade during LHC Long Shutdown 2 in 2019-2020. In order to cope with the high data throughput expected by the detector, 3.4 TB/s, a new data acquisition farm has been designed. Five hundred FPGA-PCIe readout cards, called CRU (Common Readout Unit), have been produced and two hundred servers, called FLP (First Level Processor), had to be purchased to be equipped with the readout cards. Different candidates have been identified as possible readout server. The readout requirements for RUN3 are very demanding and the server must be able to handle high data throughput up to a maximum of 100 Gbps (input and output) for long periods of data taking. A series of tests were carried out to verify the correct behavior of the server equipped with multiple readout cards, up to a maximum of three.

[*] e-mail: filippo.costa@cern.ch
[**] e-mail: sylvain.chapeland@cern.ch
[***] e-mail: kostas.alexopoulos@cern.ch
[****] e-mail: ulrich.fuchs@cern.ch

The CRU hosts an INTEL ARRIA X FPGA and it has a PCIe GEN3 x16 interface towards the server. In order to be selected, the candidate server had to satisfy various hardware and software criteria. Mechanical installation, heat dissipation, BIOS settings and memory installation are only a few requirements that defined the selection procedure for the readout server. Dedicated software and firmware have been written to test different FLP candidates and to decide the best one for ALICE requirements. In the following sections the testing procedures are described.

2 Candidate selection

The server had to be 3 U height at maximum and to allow for the installation of three CRU cards and one 100 Gbps InfiniBand network card. The initial list of candidates is listed in Tab. 1.

Table 1. FLP candidates.

SERVER MODEL	CRU
ASUS ESC4000-G4	3
DELL PowerEdge R740	3
DELL PowerEdge R7425	2
HP ProLiant DL380	3
SUPERMICRO X11DPG-QT	3

2.1 Hardware tests

The first test was a simple hardware compatibility test. After installing the three CRUs and the network card in the server, it had to be possible to boot and recognize properly all the components installed. Every CRU card is detected in the system as two PCIe devices gen3 x 8, as single PCIe gen3 x16 interface is not available nowadays in FPGA. Many modern servers support bifurcation, feature that allows to have two PCIe endpoint on a single slot card. Although supported it is not always enabled and available in the list of options in the BIOS, forcing the company to provide a customized BIOS version for the user. During our tests we had to contact ASUS to obtain a modified version of the BIOS to enable bifurcation in the system. When executing *lspci* command on Linux, the system had to show a total of seven PCIe devices, two PCIe endpoints for each CRU cards and one for the 100 Gb NIC (Network Interface Card). The DELL PowerEdge R7245 (AMD CPU based) when booting and equipped with 3 CRUs was reporting errors on the CRU cards installed in a specific PCIe slot. Changing CRU cards or PCIe riser didn't fix the issue. The HP Proliant DL380 was loading the Operating System, but the cards were not showed properly and the access to the card was slower than usual. As these 2 servers showed instability when running the system with three cards, they were excluded from the list of possible candidates.

2.2 Thermal test

One vital parameter for the proper functioning of the CRU card is the FPGA core temperature. The ARRIA X has several high-speed serial links, with a bandwidth up to 17.4 Gbps. The CRU implements 48 GBTs [2] as communication protocol with the detector electronics to collect data and transfer clock, trigger and slow control. The CRU communicates with the software through the PCIe interface for control and to perform DMA. When all these interfaces are used in parallel the temperature of the FPGA rises fast. When the FPGA temperature reaches 80 °C a microprocessor, installed on the CRU, cuts the power to the board. Installed on top of the FPGA there is a passive heat sync, the dissipation of the heat relies completely on the mechanics and airflow of the server and the speed of the fans. Low profile server (2 U height) like the DELL or the ASUS are GPU optimized and have dedicated airflow channels directed on the PCIe interfaces. This provides excellent heat dissipation. In other cases when the server is not optimized for GPU usage, like the SUPERMICRO X11DPG-QT (3 U height), there is no dedicated airflow channel for the PCIe cards. The fans installed in the machine actually create turbulence inside the server and this results in very low performances when dissipating the heat of the CRUs. During the tests of the servers with 3 CRU cards, the DELL and the ASUS showed very similar results, while the airflow in the SUPERMICRO was not sufficient to keep the temperature of the FPGA below the required operational value (60 °C). The tests showed also that the amount of air received from the three cards was not the same as shown in Tab. 2. For this reason, the FLP tender specified the minimum airflow of 3 m/s required on the CRU. The standard SUPERMICRO should be complemented with an additional dedicated element to guarantee sufficient airflow to the CRU.

Table 2. Airflow speed at the output of the FAN and of the CRU.

SERVER	FAN (m/s)	CRU (m/s)	CRU MAX TEMP (C)
ASUS	11	4.5	52
DELL	11	4.5	53
SUPERMICRO	11	4 - 2.4 - 2	68

3 Software tests

One of the main requirements for the candidate FLP was to be able to operate properly during long period of stable data taking. For this reason a series of tests have been prepared to simulate realistic data taking scenarios and to study the behavior of the platform fully equipped. The server indeed had to be able to receive data from the three readout cards and store the information in the memory. The same data had to be sent to the network using the 100 Gb InfiniBand card using RDMA. To be as close as possible to a realistic scenario, the actual readout software [3] and CRU firmware were used, applying modifications and tuning when needed.

3.1 Raw DMA throughput

As a first test the servers had to show nominal DMA performance on all PCIe devices when collecting data from the CRU. Before testing the server in real running conditions we exercised the CPU and the memory configuration of the different machines to verify whether the maximum throughput could be obtained for long period of data taking, up to a maximum of 8 hours. A CRU card is capable of generating a total throughput of 110 Gbps. The different servers were tested against this value, installing three readout cards in each server and running the same readout software and firmware. All the servers equipped with three CRUs showed similar performance absorbing a total DMA throughput of 330 Gbps (Fig. 1). In order to achieve this result BIOS configuration and memory allocation had to be changed and optimized. Modern CPUs offer many options for the software to run in an efficient way and profit from the new processor features. However, some of them can increase the time to access the PCIe bus resulting in a lower PCIe throughput of the cards. Therefore, we had to disable options like: Adjacent Cache Line Prefetch, Hardware Prefetcher, DCU stream prefetcher, DCU IP Prefetcher and Sub NUMA cluster. Another important parameter to obtain the maximum throughput is the occupancy of the memory channels in the server. Each Intel CPU provides six memory channels, and we started the tests installing a limited amount of memory modules. As soon as we started collecting data we noticed an unbalanced DMA throughput between the cards that was changing when adding more modules in the system. We noticed that leaving memory channels unequipped or having an asymmetric memory configuration between the two CPUs was degrading considerably the performances of the server, up to a point where it was not possible to run any software anymore, as showed in Tab. 3. The best performance was achieved when all the memory channels of the two CPUs were equipped with memory modules, six for each CPU in case of Intel processor.

Table 3. Memory configuration and CRU DMA throughput. Every CRU is detected as 2 PCIe cards, called end-point (EP0-EP1), because of the bifurcation. The table shows the performance of each end-point. Given the configuration of the system there is 1 CRU (CRU0) connected to a CPU and 2 CRUs (CRU1 and CRU2) connected to the second CPU.

memory modules (8 GB)	DMA performance Gbps (EP0 - EP1)
6	CRU0(54.0, 54.0) - CRU1(52.5, 52.5) CRU2(55.6, 55.6)
8	server unstable (CPU stuck)
10	CRU0(55.6, 55.6) - CRU1(30.0, 30.0) CRU2(31.4, 31.4)
12	CRU0(55.6, 55.6) - CRU1(55.0, 55.0) CRU2(55.4, 55.4)

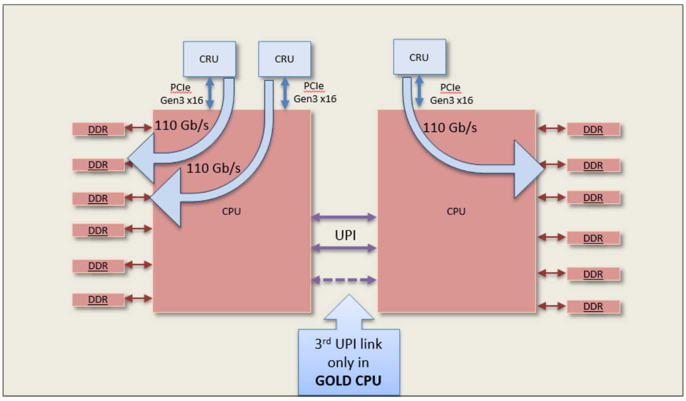

Figure 1. DMA data flow in a dual-socket system with three readout cards installed. The picture describes the data flow from the CRUs up to the memory of the server (DDR).

3.2 Realistic data flow performance

When developing software for a dual-socket server and using PCIe cards it is important to configure the NUMA node properly, otherwise the performance of the overall system can degrade easily, resulting in up to 50% throughput drop. The main goal is to configure the software that runs on CPU0 to perform DMA on the memory and PCIe cards connected to the same CPU. In the ideal situation the memory buffer allocated on the memory connected on CPU0 are passed to the PCIe card, for DMA transfer, connected to the same CPU0. Using four PCIe card in a dual-socket server, by design two PCIe cards are connected to CPU0 and two cards to CPU1 and the memory is divided by two processors. In this configuration communication between the two sockets is unavoidable. It is very important to avoid overflowing the Intel UPI (Ultra Path Interconnect) links between the two processors as this will have an impact on performance and on latency in accessing the PCIe bus. Some detectors require data processing in the FLP before transferring the data out from the server (Fig. 2). For this reason we decided to test and compare two types of Intel Cascade Lake CPUs: Silver 4210 and Gold 6230a CPU, the latter providing more computing power. During our DMA tests we noticed that the server equipped with Silver CPU was dropping packets compared to the Gold CPU that was working fine. The CPU usage to perform DMA was similar on the Gold and Silver, the main difference is an additional lane in the UPI bus for the Gold CPU. During the test we noticed that the PCIe bus access time was increased in the server equipped with Silver CPU and the buffer capacity of the firmware of the CRU board had to be doubled, fixing the issue. We concluded that although the total throughput was lower than the maximum allowed, the heavy traffic on the UPI bus increased the latency in accessing the PCIe bus, causing the CRU to drop packets. This effect was detected during the tests including the InfiniBand in the data flow and sending data over RDMA. In this case the network cards had to fetch data crossing the UPI interconnect, as two CRUs were connected to the other CPU as showed in Fig. 3.

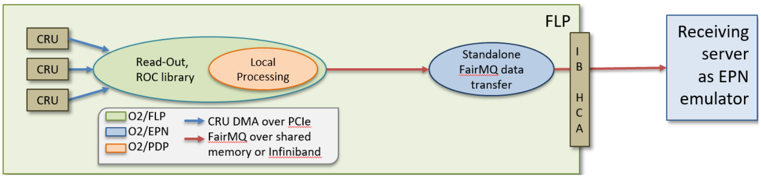

Figure 2. Realistic FLP data flow, from CRU to NIC card, including data processing. O2 FLP/EPN/PDP are the different project groups that develop software running on the FLP to handle the data stream.

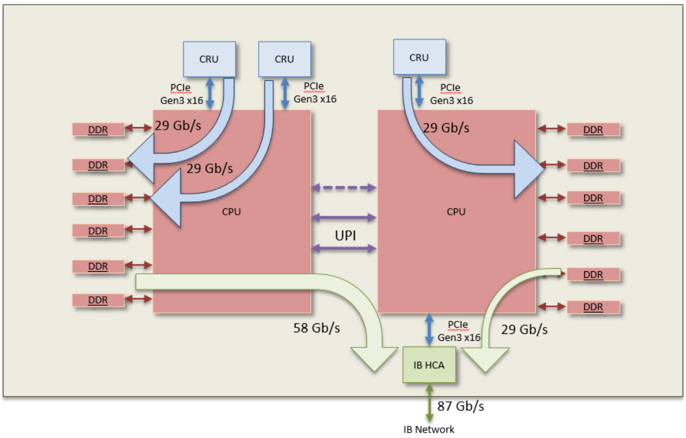

Figure 3. DMA and RDMA data flow, crossing two CPUs and different memory banks. The data is sent out to the network using an IB HCA (InfiniBand Host Channel Adapter) card

4 Conclusion

The increased complexity of servers and new FPGA readout cards used in High Energy Physics experiments, turned the selection of a server into a delicate task, requiring a lot of firmware and software preparation, thus time. It is very important to plan well in advance the tests to be performed on different candidates. Dual-socket systems, bifurcation, memory configuration, NUMA nodes, all play an important role in determining the overall performance of the server. We described the series of tests performed for ALICE on different server candidates. After testing mechanics, BIOS configuration, air flow and overall data throughput of the system the results led to the selection of the DELL PowerEdge R740 equipped with 96 GB of RAM and two flavours of Intel Cascade Lake CPU: Silver 4210 and Gold 6230. Given the higher price of the Gold CPU compared to the Silver one, we decided to purchase the Gold CPU only for the detectors that require local processing in the FLP before sending data to the EPN.

References

[1] T.A. Collaboration, K. Aamodt, A.A. Quintana, R. Achenbach, S. Acounis, D. Adamová, C. Adler, M. Aggarwal, F. Agnese, G.A. Rinella et al., Journal of Instrumentation **3**,S08002 (2008).

[2] M. Barros Marin and S. Baron and S.S. Feger and P. Leitao and E.S. Lupu and C. Soos and P. Vichoudis and K. Wyllie, Journal of Instrumentation **10**, C03021–C03021 (2015).

[3] S. Chapeland, F. Costa, EPJ Web of Conferences **214**, 01041 (2019).