AIDA

Advanced European Infrastructures for Detectors at Accelerators

Journal Publication

The C-RORC PCIe card and its application in the ALICE and ATLAS experiments

Engel, H (UNI-FRANKFURT)

13 February 2015

This work is part of AIDA Work Package **8: Improvement and equipment of irradiation and test beam lines**.

The C-RORC PCIe card and its application in the ALICE and ATLAS experiments

A. Borga[a] **F. Costa**[b] **G.J. Crone**[c] **H. Engel,**[d,1] **D. Eschweiler,**[e] **D. Francis,**[b] **B. Green,**[f] **M. Joos,**[b] **U. Kebschull,**[d] **T. Kiss,**[g] **A. Kugel,**[h] **J.G. Panduro Vazquez,**[f] **C. Soos,**[b] **P. Teixeira-Dias,**[f] **L. Tremblet,**[b] **P. Vande Vyvre,**[b] **W. Vandelli,**[b] **J.C. Vermeulen,**[a,1] **P. Werner**[b] **and F.J. Wickens**[i] **for the ALICE and ATLAS collaborations**

[a]*Nikhef National Institute for Subatomic Physics and University of Amsterdam, Amsterdam, Netherlands*

[b]*CERN, Geneva, Switzerland*

[c]*Department of Physics and Astronomy, University College London, London, U.K.*

[d]*Institut für Informatik, Johann Wolfgang Goethe-Universität Frankfurt, Frankfurt, Germany*

[e]*Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe-Universität Frankfurt, Frankfurt, Germany*

[f]*Department of Physics, Royal Holloway University of London, Surrey, U.K.*

[g]*Wigner Research Centre for Physics, Hungarian Academy of Sciences, Budapest, Hungary*

[h]*ZITI Institut für technische Informatik, Ruprecht-Karls-Universität Heidelberg, Mannheim, Germany*

[i]*Particle Physics Department, Rutherford Appleton Laboratory, Didcot, U.K.*

E-mail: hengel@cern.ch, j.vermeulen@nikhef.nl

ABSTRACT: The ALICE and ATLAS DAQ systems read out detector data via point-to-point serial links into custom hardware modules, the ALICE RORC and ATLAS ROBIN. To meet the increase in operational requirements both experiments are replacing their respective modules with a new common module, the C-RORC. This card, developed by ALICE, implements a PCIe Gen 2 x8 interface and interfaces to twelve optical links via three QSFP transceivers. This paper presents the design of the C-RORC, its performance and its application in the ALICE and ATLAS experiments.

KEYWORDS: Online farms and online filtering; Optical detector readout concepts; Data acquisition circuits; Data acquisition concepts

---

[1]Corresponding author.

Contents

1 Introduction

1.1 ALICE online architecture in Run 1 and Run 2

ALICE [1] is the heavy-ion experiment at the CERN LHC dedicated to the study of the physics of strongly interacting matter. It has been designed to cope with the high particle densities produced in central Pb-Pb collisions. The data captured from all 18 subdetectors are read out by the ALICE Data Acquisition (DAQ) system via around 500 serial optical links called Detector Data Links (DDLs) [2]. The data sent via DDLs from the cavern to the counting rooms is received in custom FPGA based DAQ Read-Out Receiver Cards (D-RORCs). These boards are installed in servers acting as Local Data Concentrators (LDCs). For each DDL an exact copy of the incoming data is forwarded within the D-RORC FPGA to another DDL towards the High-Level Trigger (HLT). A simplified overview of the read-out architecture is shown in figure 1.

The HLT is the first system in ALICE where data from all detectors is combined and reconstructed. This compute cluster is comparable in size to the DAQ cluster and additionally contains Graphics Processing Units (GPUs). The interface nodes are equipped with custom FPGA based HLT Read-Out Receiver Cards (H-RORCs), receiving the detector data via DDLs and performing first reconstruction steps. In addition to software based data processing on the nodes, the computing power of the HLT could significantly be enhanced by implementing pre-processing algorithms in the H-RORC firmware and offloading computations to GPUs [3]. Output nodes pass the processed data back to the DAQ system via H-RORCs and DDLs.

The HLT decisions for each event are readout by the DAQ, using the DDLs as for any other detector. The sub-events from the detector LDCs and the HLT decision are then sent over the Event Building Network for global processing and finally into long term storage.

The Read-Out Receiver Cards for DAQ and HLT have similar requirements, however they have been developed and maintained as independent projects. The H-RORC contains a Xilinx Virtex-4
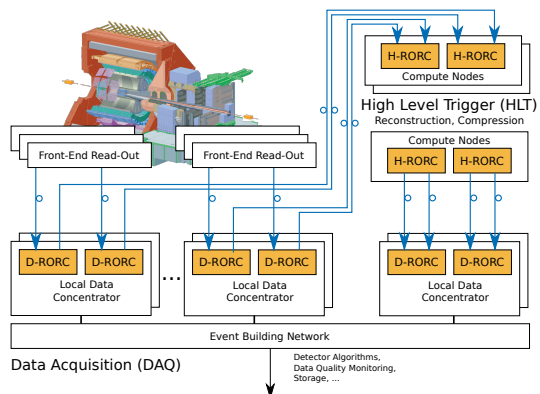
Figure 1. The ALICE online architecture with focus on the RORCs in DAQ and HLT.
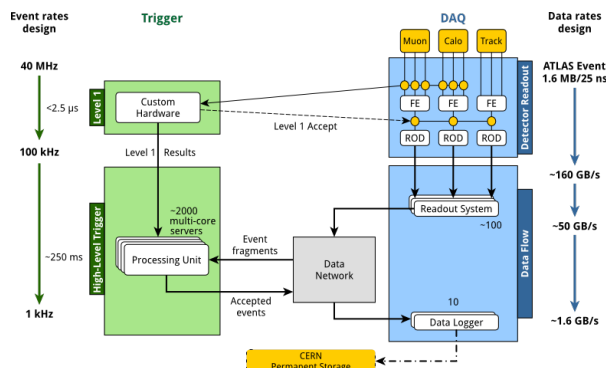

Figure 2. The ATLAS TDAQ system in Run 2.

FPGA and connects to DDLs via pluggable add-on boards hosting the optical links. The interface to the host machine is implemented with PCI-X. The D-RORCs have been used in two different revisions: one with PCI-X and one with PCIe interfacing to the host machine. These boards use Altera APEX or Stratix II FPGAs and have two optical interfaces per board. During Run 1 around 400 D-RORCs and around 240 H-RORCs were used in the DAQ and HLT systems.

The read-out architecture described will remain the same for Run 2. LHC luminosities after Long Shutdown 1 are expected to be in the range of $1 - 4 \times 10^{27} \ cm^{-2}s^{-1}$ with a center-of-mass energy of 5.1 TeV for Pb-Pb collisions. The expected data rates require that the read-out system as deployed during Run 1 is upgraded. The Time Projection Chamber (TPC) is replacing its Readout Control Unit with a redesign for higher detector bandwidth and increased output link rate (RCU2). The Transition Radiation Detector (TRD) is implementing a higher read-out link rate with the existing Global Tracking Unit (GTU) hardware. Therefore the original version of the DDL (also referred to as DDL1) has been upgraded to the DDL2 [4], which supports higher link rates. The increasing data rates and read-out changes also affect the systems of DAQ and HLT and in particular the Read-Out Receiver Cards.

Both types of RORCs used during Run 1 are limited in their optical read-out capabilities by the DDL1 link rates. Additionally, the PCI-X host interface is obsolete and increasingly rare in recent server PCs. These facts require a replacement of the Run 1 Read-Out Receiver Cards.

1.2 ATLAS: upgrade of the ReadOut system

The focus of the ATLAS experiment [5] at the LHC is the study of high-energy proton-proton collisions at high luminosities. The experiment makes use of a trigger system consisting of three levels to reduce the event rate to a manageable level. The first level consists of dedicated hardware. Data from events accepted by this level are transferred from the front-end electronics to the ReadOut Drivers (RODs). These are sub-detector specific modules, located in an underground service area adjacent to the cavern in which the experiment is installed. An important task of the RODs is to build event fragments and output these to the ReadOut System (ROS). For each first-level trigger accept each ROD outputs one event fragment. Each fragment contains an identifier, the L1Id, which is, apart from resets, monotonically increasing for consecutive fragments. A supervisor se-

lects a higher-level trigger processing node for handling the event and forwards the same L1Id and additional information, provided by the first-level trigger, to it. The additional information is used by the second-level trigger for requesting only part of the event data from the ROS. The L1Id is forwarded as part of each request for data associated with that L1Id via the Ethernet network connecting the nodes and the ROS. The ROS responds by sending the requested data. For Run 1 the second level of triggering was implemented using a dedicated set of server PCs. Upon acceptance by this level, full event building was performed by another dedicated set of server PCs known as the Event Builder,[1] which like the second-level trigger processors requested the event data from the ROS, but instead of a fraction all data were requested. Full events were then built and forwarded to the highest trigger level, known as the Event Filter and running on another dedicated set of server PCs. For Run 2 the same approach will be used, but all processing of an event, i.e. for second-level triggering, event building and Event Filter processing, will be done on the same processing node. As in Run 1 event fragments will be discarded in the ROS upon delete requests that are broadcast to the ROS by a supervisor. This occurs after a second-level trigger reject or after successful building of the full event (or of a partial event in case of certain types of events, in particular calibration events). A diagram of the structure of the Trigger and DAQ (TDAQ) system for Run 2, with data volumes and trigger rates indicated, is presented in figure 2.

The event fragments are transferred from the RODs to the ROS via dedicated point-to-point links in the form of optical fibers, using the S-link protocol [6] and running at either 160 MB/s or 200 MB/s maximum throughput. For Run 1 about 1600 of these links were deployed, this number increases to about 1800 for Run 2. The ROS as deployed during Run 1 was built from about 150 server PCs, with typically 4 ROBINs [7] installed per PC. ROBINs are PCI plug-in cards with three inputs for the point-to-point links via which the RODs output their data. Each PC was also equipped with a PCIe plugin card connecting via two ports to the data collection network, implemented with 1 Gb Ethernet technology. Each ROBIN contained a 64 MB paged memory buffer for each of the three inputs, a Xilinx Virtex-II FPGA, a PCI interface chip and a PowerPC processor keeping track, for each buffer and together with the FPGA, of the association between page number and L1Id of each fragment stored. Requests were forwarded by the PC to a ROBIN via its 64-bit 66 MHz PCI interface, requested data was written to the memory of the host via DMA.

The increase of the number of ROD-to-ROS links for Run 2 made a reduction of the rack space used per link desirable. Furthermore 64-bit PCI technology is becoming obsolete, motherboards with four PCI slots, similar to those installed in the ROS PCs used in Run 1, are not readily available for the current generation of CPUs (Ivy Bridge or Haswell architecture). A PCIe solution was therefore required. In addition the higher luminosity and collision energies of Run 2, the higher maximum average level-1 accept rate of 100 kHz (instead of about 70 kHz for Run 1), and updated trigger conditions will result in more data being sent to and requested from the ROS. Therefore it was decided to replace the ROS used in Run 1 by a more compact ROS with PCIe based ROBINs and capable of handling requests for event fragment data for at least 50% of the fragments received via the ROD-to-ROS links. With the CPU power available in modern server PCs it was considered feasible to move the tasks of the on-board processor of the ROBIN to the CPU of the ROS PC, simplifying the design of the ROBIN and also simplifying support, as both the software and the

---

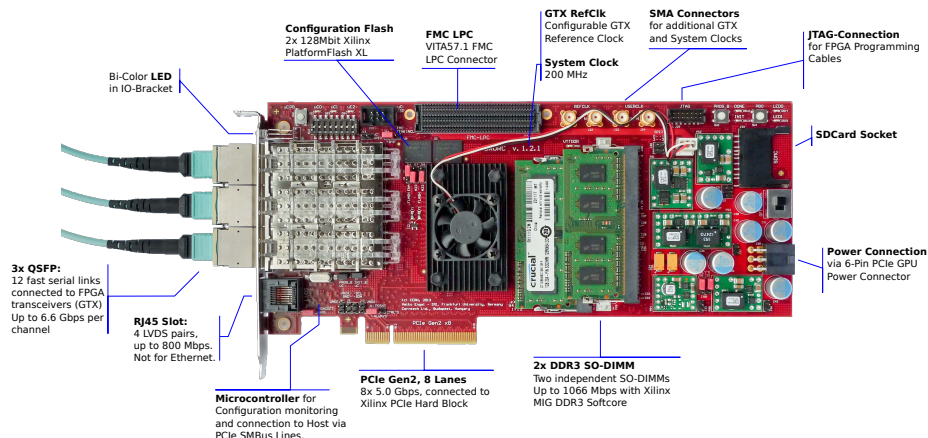[1]the PCs are also referred to as SFIs (SubFarm Inputs).

Figure 3. Photo of the C-RORC board with the major components and features annotated.
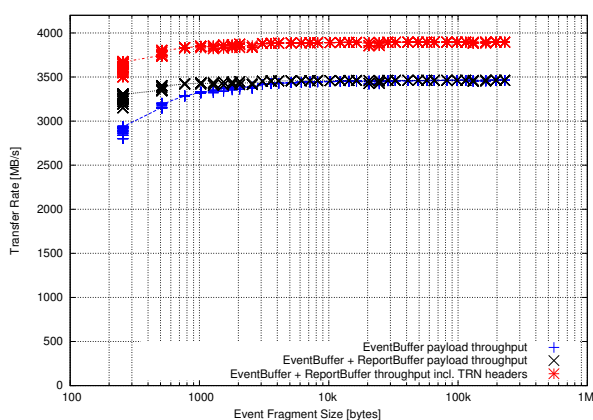

Figure 4. C-RORC DMA-to-host throughput.

development environment for the on-board processor no longer have to be maintained. This new version of the ROBIN is known as the RobinNP, "NP" refers to "No Processor". The custom board developed by the ALICE collaboration, the C-RORC, described in the next section, provides all functionality required for the RobinNP, as discussed in section 3.3.

2   The Common Read-Out Receiver Card (C-RORC)

The lack of suitable commercial platforms to replace the Run 1 Read-Out Receiver Cards deployed in ALICE led to the development of a custom board. Even though the development was driven by ALICE requirements, the target platform was kept as generic as possible. A photo of the final board with the major components annotated is shown in figure 3. The board is a full-width, full-height PCIe card according to the PCIe specification. The height of the components is kept within the specification to allow installation of boards into adjacent PCIe slots. The boards are powered from 6-pin GPU power cables.

The central component on the board is a Xilinx Virtex-6 FPGA. This FPGA already comes with a PCIe hard block for up to eight lane PCIe generation 2 (8x 5.0 Gbps). A measurement of the usable PCIe bandwidth with a maximum payload size of 256 byte per PCIe packet on a

recent IvyBridge server is shown in figure 4. This example uses a custom DMA engine and two DMA buffers as described in section 3.2. The transfer rate for the plain event payload to the host buffer is shown (lowest rates). The rate taking into account the transfer of the report words and the overall throughput at the PCIe transaction level packet interface including all transaction protocol headers (highest rates) are also shown. The throughput is quite close but not equal to the theoretical limit of 4 GB/s, as there is still a portion of bandwidth required for the link level protocols including crediting.

The board interfaces to 12 serial full duplex optical links via three QSFP modules, with each QSFP module connecting to four optical links. Break-out fibers are available to connect to the existing fiber installations. The serial links are directly connected to the transceivers of the FPGA (GTX), which limits the maximum serial link rate to around 6.6 Gbps. An on-board configurable reference clock oscillator makes it possible to use almost any link rate within the supported range. On-board DDR3 memory can be installed in two SO-DIMM sockets. The required memory controllers can be implemented in the FPGA and allow operation of single ranked modules up to 1066 Mbps and dual ranked modules up to 606 Mbps. Both interfaces have been tested with a variety of different modules up to 2×8 GB total capacity. FPGA configuration files can be stored in on-board synchronous flash memories for fast auto-configuration of the board upon power-on. Additionally, there is enough memory to store multiple FPGA configurations. A configuration microcontroller can be accessed by the host machine via SMBus even if the PCIe link is down. This allows implementation of a safe firmware upgrade procedure by always keeping a known-to-be-working configuration in the flash memory.

The large scale production of the boards was organized as a common effort between ALICE and ATLAS. Extensive hardware tests have already been conducted by the contractor. More application specific tests have been done by ALICE and ATLAS at CERN. At the time of this writing 359 boards have successfully been produced, tested and delivered to CERN, of which most have been installed in the ALICE DAQ and HLT and ATLAS DAQ systems.

3 Applications of the C-RORC in ALICE and ATLAS

With the C-RORC there is now a common hardware platform for three applications in two LHC experiments: ALICE Data Acquisition, ALICE High-Level Trigger and ATLAS TDAQ Read-Out System. Even though the platform is the same, each application has to interface to existing application-specific hardware and software infrastructure. For this reason firmware for each of the three applications is developed independently. Nevertheless, common building blocks are reused and approaches are shared. The following sections describe the applications in more detail.

3.1 ALICE data acquisition

The ALICE DAQ system handles the data flow from the detector to permanent data storage in the CERN computing center and is responsible for uploading configuration data to the detectors [8]. The interface to the DDLs in the DAQ Read-Out Receiver Card firmware is therefore providing two operating modes: *data taking* and *detector configuration*.

In *data taking* mode the receiving channel of each read-out link is used to transfer event data from the detector electronics to the DAQ farm. The transmitting channel is used for flow control. In
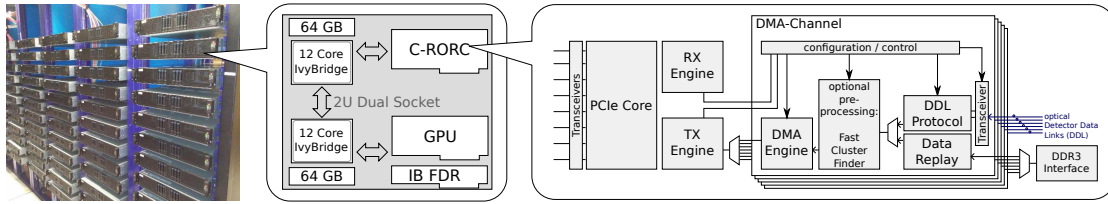
Figure 5. C-RORC Installation in the ALICE HLT for Run 2 and schematic drawing of the dataflow in the firmware.

detector configuration mode the transmitting channel is used to send configuration data to the front end electronics. The receiving channel is used for acknowledgments from the front end electronics.

The ALICE DAQ Run 2 setup is a mixed installation consisting of C-RORCs for all TPC, TRD and HLT-to-DAQ links. The previous D-RORC boards are still in use with the remaining detectors. The C-RORCs use six optical links to receive detector data and the other six links to send a copy of the data to the HLT. The copy process between the links is directly implemented in the RORC firmware. The DDL protocol has been ported to the higher DDL2 rates to support the detectors that upgrade their read-out for Run 2. The firmware interface to the host server via PCIe is based on a PLDA DMA engine [9] for six data channels. This is the same interface as already used for the D-RORC boards, which allows a common device driver and software interface for both types of boards.

The host memory for DMA operations is managed with the *physmem* driver and divided into page-like segments with known physical start addresses and lengths. These buffer descriptors are pushed into a FIFO in the RORC firmware and then used as start addresses for DMA transfers. For each descriptor used for a DMA transfer, the RORC writes an entry into a second DMA buffer in the host memory to inform the software of new data. The DAQ farm for Run 2 will consist of a cluster of around 130 servers with 10 Gb Ethernet interconnect, in which 59 C-RORCs are installed.

3.2 ALICE High-Level Trigger

In the ALICE HLT one C-RORC replaces three to six of the previous H-RORC boards, thus allowing a much denser integration of the optical links into the cluster. Up to 12 links per board are used to receive data from the DAQ system. The optical link protocol is identical to that used for ALICE DAQ: DDL at different link rates depending on the detector. For Run 2, 74 C-RORCs have been installed into 2U dual socket IvyBridge servers together with GPUs and 56 Gb InfiniBand interconnect. The overall HLT for Run 2 consists of 180 compute nodes, each with two 12-core CPUs and a GPU, and some infrastructure machines. A schematic picture of the node configuration and an overview of the dataflow inside the HLT C-RORC firmware is shown in figure 5.

The existing HLT data transport framework assumes one process per DDL. With 12 links per board this requires DMA engine firmware that is able to operate 12 DMA channels independently. This was not possible with any available commercial PCIe DMA core for the given FPGA architecture, so a custom DMA engine was developed. This DMA engine handles scatter-gather DMA descriptor lists provided by the host system and thus allows the standard Linux memory subsystem to be used for buffer allocation and mapping. The possibly scattered physical memory fragments are mapped into a contiguous virtual memory region by a user space device driver library. The
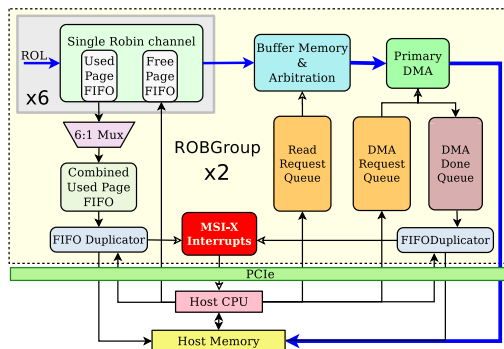
Figure 6. RobinNP firmware organization and flow of data from host CPU to the firmware (by means of programmed I/O) and from the firmware to the host memory (by means of DMA).

DMA buffers are used as ring buffers, with each DMA channel using two: *EventBuffer* and *ReportBuffer*. Detector data from the optical links are directly written into the *EventBuffer*. Once an event is fully transfered, an entry is written into the *ReportBuffer* containing the offset and length of the event in the *EventBuffer*. All hardware access is performed from user space using the Portable Driver Architecture (PDA, [10]) library together with a user space device driver. The PDA allows memory-mapping the DMA buffer twice to consecutive virtual memory addresses, which allows a transparent handling of the wrap-around effects of the ring buffers.

An essential part of the HLT firmware is the FastClusterFinder online pre-processing algorithm [11], which can be integrated into the dataflow to extract features of the raw TPC data while passing through the RORC. The FastClusterFinder can handle the full bandwidth of the DDL link and induces only a marginal additional readout latency of a few microseconds while saving a significant amount of CPU resources compared to the same processing steps in software. This was developed for DDL1 speed and and has now been tuned to support the higher optical link rates of the DDL2 protocol.

The ALICE HLT uses the DDR3 memory on the C-RORC only to replay previously recorded detector data into the system. Six DMA channels share one DDR3 SO-DIMM module. This allows the full HLT chain to be tested with real detector data without requiring DAQ or detector resources. The on-board DDR3 memory is not used during physics runs.

3.3 ATLAS Readout System

As mentioned in section 1.2, the C-RORC provides all the functionality required for the RobinNP: 12 ROD-to-ROS links can be connected to a single board, four times as many as to the ROBIN, and the PCIe interface has a throughput of at maximum about 15 times that of the PCI interface of a ROBIN. The resource requirements for implementation of RobinNP functionality are fully satisfied by the FPGA, which is also capable of handling the requirements with respect to data throughput. Furthermore up to 16 GB of buffer memory can be installed in the two SO-DIMM slots, while a ROBIN has 192 MByte of buffer memory. Higher speeds than the current 160 or 200 MB/s are also possible for the input links.

The C-RORC made it possible to build a new compact ROS that makes use of 98 2U-high server PCs, with per PC typically two C-RORCs installed and 24 ROD-to-ROS links connected (to

be compared to 12 links connected per 4U-high ROS PC for Run 1). Because of the factor of two increase of the number of links connected to a single PC, and because of the higher request fractions the networking infrastructure also had to be upgraded: instead of having two 1 Gb Ethernet links a ROS PC is now connected with four 10 Gb Ethernet links to the data collection network.

A schematic diagram of the RobinNP firmware and its interactions with the host PC is presented in figure 6. The firmware consists of two identical parts, referred to as ROBGroups, each connecting to six ROD-to-ROS links (labeled as ROL (ReadOut Link) in the diagram) and a common part implementing an eight lane Gen 1 PCIe interface and the DMA engine. The latter is the engine available from PLDA [9]. Each ROBGroup has one shared buffer memory, consisting of a 4 GByte DDR3 SO-DIMM module, which is logically subdivided in six partitions, one for each ROD-to-ROS link. Pages in the buffer memories are managed by multi-threaded software running on the ROS PC, a typical page size is 2 kByte. For each memory partition the PC provides information on free memory pages, via FIFOs implemented in firmware, to each of the 12 input handlers. Incoming fragments are stored in free pages. For every page used, information on the page number, L1Id and length of the fragment stored is entered in the Used Page FIFO of the input handler that handled the fragment. Per ROBGroup the information from each of these FIFOs flows into the "Combined Used Page FIFO", and is subsequently transferred to the memory of the PC by means of DMA by the "FIFO duplicator". The information is used by a dedicated thread for "indexing", i.e. information is stored on the relation between L1Id and the page (or pages if the fragment is larger than the page size) in which a fragment is stored as well as on the length of the fragment. Data requests received via the network cause a look-up of this information and forwarding of requests for reading data from the pages concerned. These data are then read by the FPGA from the DDR3 memory and passed to the DMA engine for transfer to the memory of the PC. For each ROBGroup a second FIFO duplicator transfers information concerning completed DMA transfers from a FIFO to the memory of the PC. This information is used for collecting the requested data, which is output via the network. Clear requests are also sent to the ROS via the network. These requests result in the identifiers of the pages concerned being recycled onto a free page stack and eventually back onto the Free Page FIFOs, thus allowing the data in memory to be overwritten. The communication between RobinNP and the PC is interrupt driven: the indexer thread is woken upon storage of new event data and the thread used for data collection is woken upon the completion of DMA transfers. Interrupt coalescence has been implemented in an innovative way: an interrupt only occurs if the buffer to which data is transferred from the FIFO with which the interrupt is associated is empty upon arrival of new data. During normal operation the PC does not need to read any data via PCIe from FIFOs in the FPGA, as all data is written under DMA control to the memory of the PC. In this way optimum utilisation of the available PCIe bandwidth is achieved.

At the time of writing the installation of the new ROS has just been completed. Each of the 98 installed ROS PCs has a single CPU motherboard equipped with an Intel E5-1650v2 six-core 3.5 GHz CPU and 16 GB of memory. The CPU connects directly to 40 PCIe Gen3 lanes, 32 lanes are connected to a riser card with four 8 lane connectors. In most of the PCs two connectors are used for two C-RORCs, the other two for two dual-port 10 Gb Ethernet NICs with optical transceivers. The operating system of the PCs is Linux (SLC6). This configuration has been shown to be able to satisfy the 50% readout fraction requirement at 100 kHz first-level trigger accept rate with two C-RORCs with RobinNP firmware installed [12].

4 Conclusion & outlook

This paper presents the C-RORC, a PCIe-based FPGA read-out board, which will be used in two of the major LHC experiments for three applications in data taking for Run 2. All parties strongly profited from the collaboration. The significant increase in production volume with respect to to deployment restricted to ALICE led to cost savings per board for both experiments. Usage experience, implementation methods and partly even source code could be shared between the developers of the different applications reducing the overall development time. All boards required for Run 2 have been successfully produced, tested, delivered and installed in the ALICE DAQ and HLT systems and in the ATLAS DAQ system.

Acknowledgments

References

[1] ALICE collaboration, *The ALICE experiment at the CERN Large Hadron Collider*, 2008 *JINST* **4** S08002.

[2] ALICE collaboration, *The technical design report of the trigger, data-acquisition, high level trigger, and control system*, CERN-LHCC-2003-062 (2003).

[3] S. Gorbunov and D. Rohr, on behalf of the ALICE collaboration, *ALICE HLT high speed tracking on GPU*, *IEEE Trans. Nucl. Sci.* **58** (2011) 1845.

[4] F. Costa, *DDL, the ALICE data transmission protocol and its evolution from 2 to 6 Gb/s*, submitted for publication in the proceedings of *Topical Workshop on Electronics for Particle Physics 2014 (TWEPP2014)*, 22–26 September 2014, Aix en Provence, France.

[5] ATLAS collaboration, *The ATLAS experiment at the CERN Large Hadron Collider*, 2008 *JINST* **3** S08003.

[6] H.C. van der Bij, R.A. Mclaren, O. Boyle and G. Rubin, *S-LINK, a data link interface specification for the LHC era*, *IEEE Trans. Nucl. Sci.* **44** (1997) 398.

[7] R. Cranfield et al., *The ATLAS ROBIN*, 2008 *JINST* **3** T01002.

[8] F. Carena et al., *The ALICE data acquisition system*, *Nucl. Instrum. Meth.* **A 741** (2014) 130.

[9] PLDA PCIe EZDMA IP core for Xilinx FPGAs, http://www.plda.com.

[10] D. Eschweiler and V. Lindenstruth, *The portable driver architecture*, in *Proceedings of the* 16th *Real-Time Linux Workshop*, Open Source Automation Development Lab (OSADL), October 2014.

[11] T. Alt, *A FPGA based pre-processor for the ALICE High-Level Trigger*. Ph.D. thesis, Goethe-University Frankfurt, Frakfurt, Germany, to be published.

[12] A. Borga et al., *Evolution of the ReadOut System of the ATLAS experiment*, PoS(TIPP2014)205.