PROCEEDINGS OF SCIENCE

Evolution of the ReadOut System of the ATLAS experiment

A. Borga[a] **G.J. Crone**[b] **B. Green**[c] **A. Kugel**[d] **M. Joos**[e] **J.G. Panduro Vazquez**[c] **J. Schumacher**[ef] **P. Teixeira-Dias**[c] **L. Tremblet**[e] **W. Vandelli**[*e] **J.C. Vermeulen**[a] **P. Werner**[a] **F.J. Wickens**[g]

[a] *Nikhef, National Institute for Subatomic Physics and University of Amsterdam, Amsterdam, The Netherlands*
[b] *Department of Physics and Astronomy, University College London, London, United Kingdom*
[c] *Department of Physics, Royal Holloway University of London, Surrey, United Kingdom*
[d] *ZITI Institut für technische Informatik, Ruprecht-Karls-Universität Heidelberg, Mannheim, Germany*
[e] *CERN, Geneva, Switzerland*
[f] *Department of Computer Science, University of Paderborn, Paderborn, Germany*
[g] *Particle Physics Department, Rutherford Appleton Laboratory, Didcot, United Kingdom*
E-mail: wainer.vandelli@cern.ch

ATL-DAQ-PROC-2014-012
21 June 2014

The ReadOut System (ROS) is a central and essential part of the ATLAS data-acquisition system. It receives and buffers event data accepted from all sub-detectors and first-level trigger subsystems. Event data are subsequently forwarded to the High-Level Trigger system and Event Builder via a GbE-based network. The ATLAS ROS will be completely renewed in view of the demanding conditions expected during Large Hadron Collider (LHC) Run 2 and Run 3. The new ROS will consist of roughly 100 Linux-based 2U-high rack-mounted server PCs, each equipped with 2 PCIe I/O cards and four 10GbE interfaces. The FPGA-based PCIe I/O cards, developed by the ALICE collaboration, will be configured with ATLAS-specific firmware, called RobinNP. They will provide connectivity to about 2000 point-to-point optical links conveying the ATLAS event data. This dense configuration provides an excellent test bench for studying I/O efficiency and challenges in current COTS PC architectures with non-uniform memory and I/O access paths. In this paper the requirements for Run 2 and the design choices for a system complying with or exceeding them are described. The results of performance measurements for different computer architectures, highlighting the effects of non-uniform resource distributions, are discussed. Finally the status of the project and outlook for operation in 2015 are presented.

Technology and Instrumentation in Particle Physics 2014
2-6 June, 2014
Amsterdam, the Netherlands

[*] Speaker.

1. Introduction

ATLAS [1] is one of the experiments installed at the LHC, CERN, Switzerland. In preparation for the data-taking period planned for 2015–2018 (Run 2), several upgrade and maintenance activities are taking place in the ongoing shutdown period. These are meant to improve overall ATLAS physics performance, enabling operation at a peak luminosity twice as high as in Run 1 and beyond the initial design goal. Among the many interventions, for the scope of this paper, it is important to mention the installation of an additional tracker layer and the reduction of multiplexing, and as a consequence extension, of off-detector electronics. The latter is needed to deal with higher data rates and larger event fragments.

In this paper, the upgrade of the ATLAS ReadOut System (ROS), a key data-acquisition component interfacing with the detector electronics, is discussed. After introducing the ROS functions, the requirements for Run 2 are detailed and the upgraded design is presented. Next the challenges of the ROS workload for commercial computing equipment are introduced and the results of measurements for different computer architectures are presented and discussed.

2. ATLAS Trigger and Data Acquisition (TDAQ)

The ATLAS Trigger and Data Acquisition (TDAQ) [2] system is responsible for selection and conveyance of interesting physics data while reducing the initial LHC collision frequency of 40 MHz to an average rate of stored physics events of 1 kHz.

The ATLAS TDAQ system for Run 2 is organized in a two-level selection scheme (figure 1), including a hardware-based first-level trigger (Level 1) and a software-based High-Level Trigger (HLT). The HLT operates over partial event information, driven by Level 1-tagged features, called Region-of-Interest (RoI). In Run 2, the rate of events accepted by the Level 1 trigger and filtered by the HLT is expected to be 100 kHz, 30% higher than the nominal 75 kHz rate of Run 1.

The TDAQ data-flow is centered around a push-pull architecture. Data fragments from the on-detector Front-End electronics (FE) are transmitted to the off-detector electronics (ReadOut Drivers - RODs) and then pushed via $\sim$2000 optical links into the ReadOut System (section 3). Fragments are then served as requested by the HLT processing tasks over an Ethernet network. Events accepted by the HLT are finally moved to a transient storage system (Data Logger). At the start of operation in 2015, the TDAQ HLT computing farm will include roughly 2000 multi-core servers executing more than 20000 processing applications.

The ATLAS TDAQ system for Run 2 does not use an explicit event building infrastructure [3]. The data collection is performed by a dedicated task, the Data-Collection Manager (DCM), operating on each node of the HLT computing farm. Event by event, the collection of data fragments is fully driven by the HLT algorithms. For accepted events, the DCM guarantees that the data collection process will be completed before the events are moved to the storage system.

The TDAQ system is based on in-house designed multi-threaded software, mostly written in C++ and Java and running in a Linux environment.

3. ReadOut System

As mentioned in section 2, the ReadOut System couples the detector sub-systems and the com-

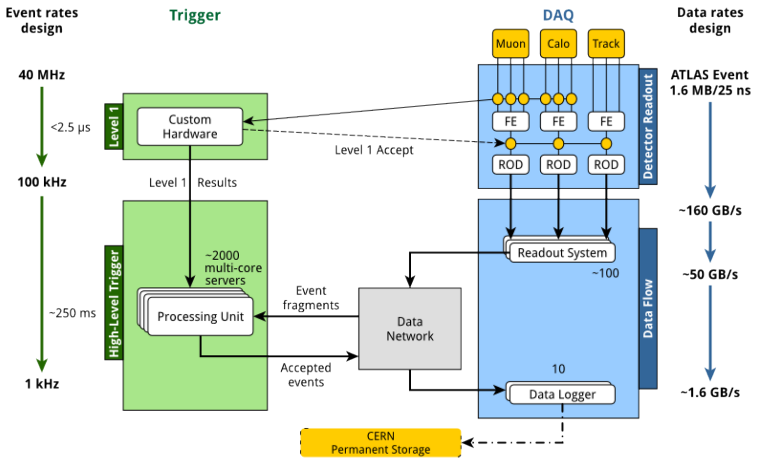


Figure 1: Functional diagram of the ATLAS Trigger and Data Acquisition system for Run 2. A detailed explanation of the different components is given in section 2.

mon ATLAS data-acquisition system. Besides interfacing the custom optical links to the Ethernet network, the ROS buffers the data fragments in internal memories until rejection or event building and serves data on-demand to the HLT tasks.

In Run 1, the ROS was implemented with ∼150 4U-high PCs [4]. Each PC was equipped with up to 4 (exceptionally 5) custom receiver cards, the ROBINs [5], and two copper GbE ports for data forwarding. The ROBIN was an FPGA-based PCI card providing three optical receivers capable of operating at 2 Gbps and compatible with the S-LINK [6] protocol. Incoming data fragments were stored in on-board memory (64 MB/link) and an embedded processor took care of data bookkeeping and request management. The readout fraction, defined as the fraction of received data fragments which are forwarded to the HLT, is a key parameter summarizing the performance figure and requirements of a ROS PC. The latest model of ROS PC in Run 1 was capable of achieving a readout fraction of 10–15%. Depending on the operating conditions, performance was either limited by the ROBIN embedded processor capabilities or by the network connectivity.

4. Readout System Upgrade: Motivations and Requirements

Going from Run 1 to Run 2, the Readout System functions remain unchanged. However operational and technological aspects suggested the need for an upgrade. In particular:

- As discussed in section 1, changes on the detector side increased the number of readout links by 25%, from ∼1600 to ∼2000. As a consequence, the ROS will either require more rack space or denser packing.
- Other parameters being equal, the event complexity and therefore the HLT data-access needs increase with the luminosity. In order to be compatible with the trigger scenarios for Run 2 and offer enough operational margin, the ROS PC must be capable of sustaining a readout fraction of 50%.
- The size of the memory buffer per link in the ROS constrains the average HLT processing time before event building and, as a consequence, the HLT farm size in terms of parallel

processes. Considering the HLT processing time dependency on luminosity and the recent trends in multi-core CPUs, a substantial increase in the buffering capabilities is essential.

- A denser solution and a larger readout fraction imply a higher throughput per ROS PC. Hence moving from GbE to 10GbE is also required.
- Parallel PCI, upon which the ROBIN is based, is an ageing technology, less and less common in COTS computing equipment. A PCIe-based solution would guarantee a longer-term availability of compatible computing platforms.
- Looking forward to the ATLAS upgrade plans in preparation for Run 3, forward compatibility with new generations of faster optical links would be a long-term advantage.

The ROS upgrade project discussed in the following sections, the Generation III (Gen III) project, aims to fulfill these goals.

5. Generation III ROS Design

The ALICE Common ReadOut Receiver Card (C-RORC) [7] was identified as an ideal platform for the development of the Gen III ROS optical receiver and buffer. It provides:

- 3 QSFP cages with transceivers for connecting up to 12 serial bi-directional optical links operating at up to 6 Gpbs;
- one Xilinx Virtex-6 FPGA;
- two SODIMM RAM modules, up to 8 GB each, compatible with DDR3-1066;
- a PCIe interface supporting configurations up to Gen2 x8.

Dedicated ATLAS firmware for the C-RORC card, called RobinNP (Robin No Processor), was developed. Based on the original ROBIN firmware, it includes innovative elements. The C-RORC does not have an on-board embedded processor: the data and request management tasks are instead off-loaded to the CPU on the host PC. This requires a low latency path between the C-RORC hardware and the software operating on the host. Special hardware entities implemented in the FPGA, the FIFO duplicators, automatically transfer information from the C-RORC to the host memory avoiding PCIe read operations.

In addition, unlike the ROBIN, the RobinNP firmware uses MSI-X interrupts to notify the host of task completion and data availability. This allowed the redesign of the high-level software operating on the ROS computers to better perform on modern multi-core CPUs.

6. Generation III ROS Computer Architecture

Initial measurements with the C-RORC card operating the RobinNP firmware and the rack space requirements indicated that the optimum ROS PC configuration should house two C-RORC cards, therefore interfacing 24 optical links. At a Level 1 trigger rate of 100 kHz, the maximum average fragment size for the S-LINK is ~1.6 kB. A Gen III PC operating in the worst-case scenario involving the maximum readout fraction (50%) and the largest average fragment size will

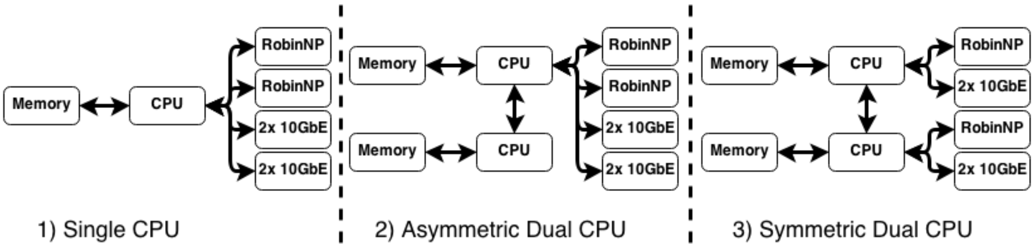


Figure 2: Functional schemes of the computer architectures compared in the performance studies discussed in section 7. Arrows represent the available communication paths.

output over the network roughly 16 Gbps. This is largely compatible with the PCIe interface capabilities: the effective application-level throughput for a PCIe Gen1 x8 interface, as implemented in the RobinNP firmware, is ∼12 Gbps [8], hence ∼24 Gbps for two cards. For a PC with this configuration, at least two 10GbE ports are required to forward data to the HLT. For redundancy reasons each PC will be equipped with four 10GbE ports, using two dual-port network cards.

Modern commercial CPUs include embedded memory and PCIe controllers. In particular in multi-CPU servers, this leads to non-uniform memory (NUMA) and I/O modules (NUIOA) access patterns. This can be clearly seen in the central and right parts of figure 2: a process might require one or two steps to reach an external resource depending on their relative locations. As shown in the left part of the figure, this problem does not apply to a single CPU architecture where all resource accesses are uniform[1].

The workload of the Gen III ROS is very I/O intensive, streaming up to 16 Gbps from the C-RORC cards to the host memory and from the host memory to the network cards. A priori, a single-CPU seems more appropriate for such a workload. On the other hand, a dual-CPU solution can provide more integrated computing power, in particular in the form of additional cores and hence parallelism. It was therefore decided to study the Gen III ROS performance on different computer architectures with the goal of defining the best platform for this system.

7. Performance Studies

7.1 Experimental Setup

Three sample systems, matching the architectures presented in figure 2, were used. The respective CPU characteristics are summarized in table 1. Each sample was equipped with two cards operating the RobinNP firmware and four 10GbE ports, as for the baseline Gen III ROS configuration.

The performance of the sample systems was evaluated in laboratory conditions. Data fragments with configurable size were either provided by data-source PCs via 24 S-LINK fibres or by means of internal data generators built-in in the RobinNP firmware. Data-sink PCs equipped with a sufficient number of 10GbE ports were used to issue data requests in configurable patterns. Dedicated emulation software was executed on the data-source and data-sink machines while the ROS sample under test executed the standard ROS application.

[1] Single-socket CPUs with NUMA characteristics were not considered in this study.

Architecture Type	Number of CPUs	CPU Model	CPU Clock [GHz]	Number of SMT cores
Single CPU	1	Intel E5-1650V2	3.5	6
Asymmetric dual CPU	2	Intel E5-2690	2.9	8
Symmetric dual CPU	2	Intel E5-2643	3.3	4

Table 1: CPU characteristics of the test PCs. Simultaneous MultiThreading (SMT) provides hardware support for multiple (two for the concerned CPU models) threads per core.

7.2 Measurement Results

While measurements were performed with different data fragment sizes and data access patterns, the results presented in this section refer to the worst case operational conditions introduced in section 6. Figure 3 summarizes the results of the computer architecture comparison. Different approaches, as detailed in the following, were used to study the system behaviour.

Initially the single-CPU sample and the symmetrical dual-CPU one were compared. As shown in figure 3a, the single CPU motherboard provides the performance level needed for Run 2. The symmetric dual-CPU configuration does not provide substantial improvements. This indicates that the additional parallelism offered by the dual-CPU system cannot be fully exploited by the ROS workload. This could be due to software inefficiencies or to the fact that potential advantages are offset by the additional communication and scheduling complexity of the non-uniform architecture.

While it is possible to envisage software schemes potentially making better use of a dual-CPU system, any performance improvement with respect to a single CPU configuration would be outweighed by the cost difference. As reported in figure 3a, CPUs capable of operating in a dual-socket system are normally significantly more expensive than similar single-socket CPUs.

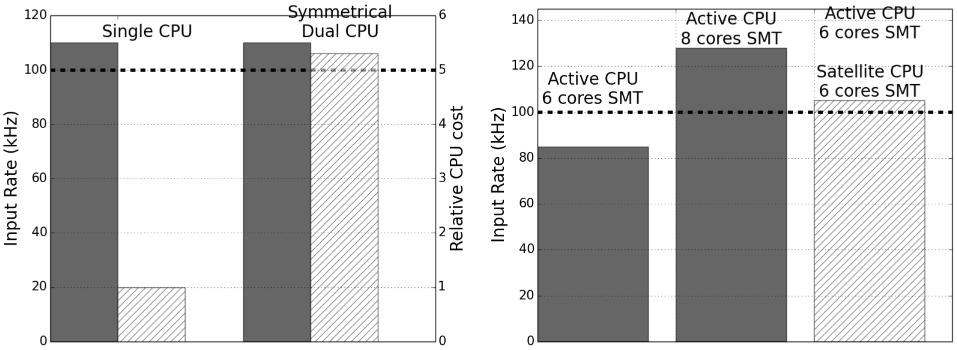
The asymmetric dual-CPU sample computer allows investigation of the results of the single-CPU and symmetric configurations. Indeed the system can be seen as a single-CPU architecture (*primary* CPU) with an additional *satellite* CPU which does not provide direct I/O paths. In order to exploit this scheme, the Linux CPU hotplug capabilities [9][10] were used. CPU hotplug allows enabling and disabling of individual computing cores at run time. While typically employed in virtualized environments, in this study CPU hotplug enables in situ evaluation of the performance dependency on the number of cores and the effects of a satellite CPU.

For a single-CPU system, the performance strongly depends on the number of available computing cores. This was also verified in the asymmetric dual-CPU system, by disabling all satellite CPU cores and only selectively the primary CPU cores. This is shown by the solid bars in figure 3b.

Enabling additional cores in the satellite CPU leads to marginal improvements or even performance reduction (dashed bar in figure 3b). This indicates that the additional I/O cost introduced by the CPU-to-CPU communication cancels the potential performance gains of the additional computing cores.

7.3 Generation III ROS Performance

Based on the results of the investigations discussed in section 7.2, it was concluded that a single CPU system is the best computer architecture for the third generation ROS computers. It should be



(a) (b)

Figure 3: (a) Solid bars show the performance of the single-CPU computer and a dual-CPU computer. Dashed bars represent the relative CPU release-cost normalized to the Intel E5-1650V2 model. The dashed line shows the ATLAS Run 2 target Level 1 trigger rate. (b) Performance comparison of different core configurations on a dual-CPU asymmetric computer. The dashed line shows the ATLAS Run 2 target Level 1 trigger rate.

noted that operational requirements further constrained the final choice of PC to be used, e.g. the motherboard had to support the use of I2C across PCIe to allow in-situ loading of new firmware in the C-RORC. The current performance of the chosen system, whose CPU characteristics are summarized in table 1, are presented in figure 4. At 50% readout fraction, the system capabilities exceed or match the required 100 kHz Level 1 trigger rate for all accessible average fragment sizes. For average fragment sizes smaller than ∼0.7 kB, the Gen III ROS can sustain a readout fraction of 100%. For larger fragments, 100% readout operation is still possible for less than 24 input optical links. This capability will be exploited for specific trigger-related detector systems from which data will be requested by the HLT for most of the events. It should be noticed that, for large fragment sizes, the system throughput approaches the effective PCIe interface bandwidth limit (section 6).

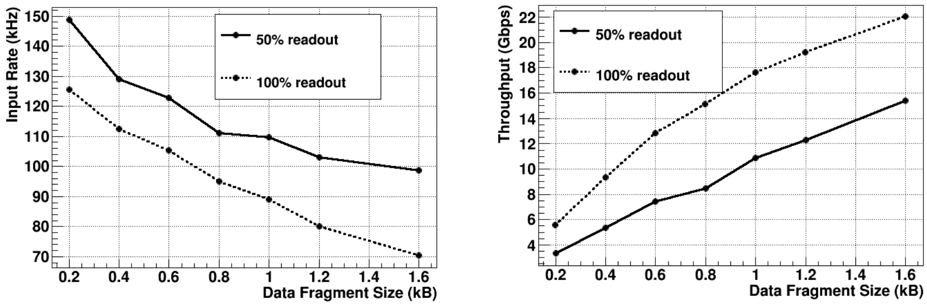
Even if the baseline performance figures are already compatible with the ATLAS requirements, ongoing optimization of the RobinNP firmware and ROS software are expected to yield additional improvements and therefore more flexibility and operational margin.

8. Conclusions

The third generation of the ATLAS ReadOut system is shifting from the development to the deployment phase. While the system functions are unchanged with respect to Run 1, the Gen III ROS is introducing a new technology landscape and boosting the operational capabilities.

Overall the system provides a factor ∼3 performance improvement, from 10–15% to 50% readout fraction, with respect to the ROS Gen II. Due to the higher system density, the capabilities of the individual PCs are boosted by a factor 6 going from Gen II to Gen III.

The performance for this intensive I/O workload was evaluated for different computer architectures, concluding that a single-CPU system is sufficient and provides the best performance per unit cost. Virtualization and power management tools, like CPU hot-plug and dynamic frequency



(a) (b)

Figure 4: Performance figure of the chosen Gen III ROS PC hardware as a function of the average data fragment size for two different data-request patterns. The input event rate (a) is equivalent to the maximum acceptable Level 1 trigger rate. In (b) the corresponding network throughput is shown.

scaling (not discussed in this paper), were used to investigate and compare different computer architectures.

The Gen III ROS system is expected to be fully operational by the end of 2014, in time for the start of LHC Run 2.

References

[1] ATLAS Collaboration, *The ATLAS experiment at the CERN Large Hadron Collider*, *JINST* **3** (2008) S08003.

[2] ATLAS Collaboration, *ATLAS, High-Level Trigger, Data Acquisition and Controls*, CERN/LHCC/2003-022, CERN Geneva 2003. `https://cds.cern.ch/record/616089`

[3] N. Garelli (on behalf of the ATLAS Collaboration), *The Evolution of the Trigger and Data Acquisition System in the ATLAS Experiment*, *J. Phys.: Conf. Ser.* 513 (2014) 012007.

[4] G. Crone et al., *The ATLAS ReadOut System Performance with first data and perspective for the future*, *Nucl. Instr. and Meth. A* 623 (2010), 534-536.

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

[6] H.C. van der Bij, R.A. Mclaren, O. Boyle, G. Rubin, *S-LINK, a data link interface specification for the LHC era*, *IEEE Trans. Nucl. Sci.* **44** (1997), 398-402.

[7] H. Engel,U. Kebschull (For the ALICE Collaboration), *Common read-out receiver card for ALICE Run2*, *JINST* **8** (2013) C12016.

[8] A. Goldhammer, J.J. Ayer, *Understanding Performance of PCI Express Systems*, Xilinx WP350, 2008. `http://china.xilinx.com/support/documentation/white_papers/wp350.pdf`

[9] Z. Mwaikambo et al., *Linux kernel hotplug CPU support*, in proceedings of *Linux Symposium*, Vol. 2, 2004. `http://www.linuxsymposium.org/archives/OLS/Reprints-2004/Reprint-Russell-OLS2004.pdf`

[10] `https://www.kernel.org/doc/Documentation/cpu-hotplug.txt`