Available online at www.sciencedirect.com

ScienceDirect

Nuclear and Particle Physics Proceedings 273–275 (2016) 939–944

www.elsevier.com/locate/nppp

# The ATLAS Data Acquisition System: from Run 1 to Run 2

William Panduro Vazquez

*Department of Physics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom*

on behalf of the ATLAS Collaboration

**Abstract**

The experience gained during the first period of very successful data taking of the ATLAS experiment (Run 1) has inspired a number of ideas for improvement of the Data Acquisition (DAQ) system that are being put in place during the so-called Long Shutdown 1 of the Large Hadron Collider (LHC), in 2013/14. We have updated the data-flow architecture, rewritten an important fraction of the software and replaced hardware, profiting from state of the art technologies.

This paper summarizes the main changes that have been applied to the ATLAS DAQ system and highlights the expected performance and functional improvements that will be available for the LHC Run 2. Particular emphasis will be put on explaining the reasons for our architectural and technical choices, as well as on the simulation and testing approach used to validate this system.

*Keywords:* ATLAS, TDAQ, Upgrade

## 1. Introduction

The ATLAS experiment [1], based at the Large Hadron Collider (LHC) at CERN, Switzerland, is undergoing a 2-year maintenance and upgrade process in preparation for a new phase of data taking starting in 2015 (Run 2). The run conditions during this new phase will place greater requirements on the data acquisition (DAQ) system. These range from the increased energy and rate of LHC collisions, potentially leading to larger particle interactions (events) to process, to detector and readout improvements requiring the system to process more data and higher rates than before.

The focus of this paper will be the upgrade of the ATLAS data acquisition system, in order to meet the challenges of Run 2 as summarised above. The main features to be discussed will be the upgrades of the readout system (ROS) and high-level trigger as well as the dataflow network through which they interact. Upgrade work is still ongoing but is scheduled to be completed by the end of 2014.

## 2. ATLAS Trigger and Data Acquisition System in Run 1

The structure of the trigger and data acquisition (TDAQ) system in Run 1 is shown in Figure 1. The system is responsible for receiving and interpreting sensor signals from the ATLAS detector and converting them at high rate into datasets which can be analysed in search of interesting physics phenomena.

Event data are initially read out via purpose-built electronics (referred to as the front-end). These systems perform initial pulse shaping, analogue-to-digital conversion and aggregation of the signals received from on-detector sensors. Portions of these data, from the calorimeter and muon systems, are then fed to what is known as the level-1 (L1) trigger. This system, also implemented using custom electronics, makes fast decisions as to whether to further process or discard an event. At this stage in Run 1 the overall event rate was reduced from 20 MHz to a maximum of 75 kHz.

If an incoming event passes level-1 selection a signal

http://dx.doi.org/10.1016/j.nuclphysbps.2015.09.146

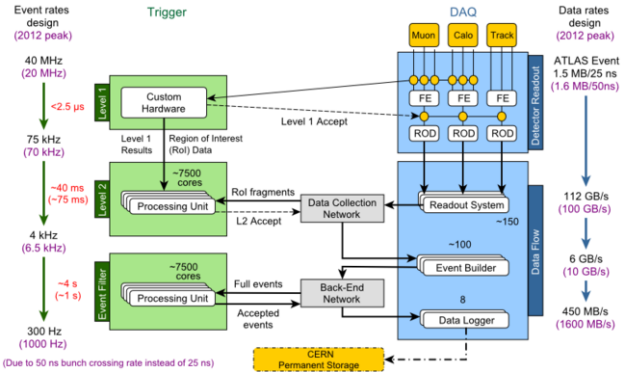


Figure 1: ATLAS Trigger and Data Acquisition System in Run 1.

is sent back to the front end, causing the data associated with the event to be read out for all components of the detector. These are relayed via optical fibres (approximately 1600 for Run 1 and 1800 for Run 2) using the 160 MB/s S-link protocol [2] from dedicated relay hardware (known as readout drivers, or RODs) to the readout system (ROS). The ROS is the first part of the DAQ chain to make partial use of off-the-shelf hardware. During Run 1, data were stored in buffers implemented on custom PCI expansion boards (known as ROBINs [3]) hosted in commercially available server PCs.

In order to avoid the overhead of transferring unwanted data, the event components used as part of the level-1 decision are at this point used to construct regions-of-interest (ROI) [4] to guide further event reconstruction and selection. These regions are based on geographical locations within the detector in which interesting signals are identified and assembled via dedicated hardware. A processing farm, known in Run 1 as the 'level-2' (L2) trigger, then processes the ROI information and samples other data from the indicated events from the buffers in the ROS PCs. The data are transferred via a high speed Ethernet-based network (the data collection, or DC, network) and subjected to software based selection algorithms.

In Run 1, events passing level-2 selection were then sent for full assembly via the PCs in the 'event builder' (EB) farm. The peak output event rate of L2 was 6.5 kHz. The EB farm requests full readout of the selected event from the ROS and then passes the data to the final 'event filter' (EF) farm via a second high speed network (the back-end, or BE, network) where final, more complex, selection algorithms are applied. Events passing this final stage are relayed to the data logging system, where they are written to permanent storage.

The L2, EB, EF and data logging stages are all implemented on commercially available server PCs using entirely software-based selection algorithms. Collectively the L2 and EF systems are referred to as the 'high level trigger' (HLT). The peak data rate recorded to disc after the HLT in Run 1 was of order 1 kHz. This translated to ~10-15% of all data reaching the ROS being read out for further analysis.

## 3. Run 1 Performance and Upgrade Motivation

The TDAQ system performed well in Run 1. Downtime due to problems with system components was kept to a minimum, leading to an overall data-taking efficiency of 94.9% [5]. The remaining inefficiency mainly came from irreducible 'dead-time' within the front-end readout electronics, whereby there is a fixed time window after processing a signal in which a given sensor is unable to process new input.

The requirements placed on the system evolved during Run 1 as a function of the increased collision rate (luminosity) provided by the LHC as well as the complexity of these events. The primary source of complexity is an effect known as 'pile-up'. This is where multiple proton-proton interactions occur during a single bunch crossing, which the level-1 trigger is unable to separate, resulting in high detector occupancy and making it more difficult to cleanly select interesting physics processes within the event. The overall effect of this is an increase in the volume of data needing to be processed through the system, while achieving the processing rates required to effectively handle all of the events without a backlog. The effect of pile-up on some significant DAQ system parameters is presented in Figure 2.

The throughput challenge was partially addressed throughout Run 1 by an ongoing program of improvement of HLT event selection algorithms, optimised network management and dataflow, and upgrade of HLT farm machines and ROS PCs. While this played a significant role in allowing the system to reach the desired performance throughout the run, it became clear that similar incremental changes would prove insufficient for Run 2.

The plan for the 2013/2014 shutdown period included the installation of new detector and trigger components [7], each requiring readout paths of their own. The effect of this has been an increase in the required number of links between the RODs and ROS from 1600 in Run 1 to near 1800. The original plan was also to increase the rate of data accepted by the L1 trigger system from 75 kHz to 100 kHz, as well as increasing the average output rate written to disc of 1 kHz, as opposed

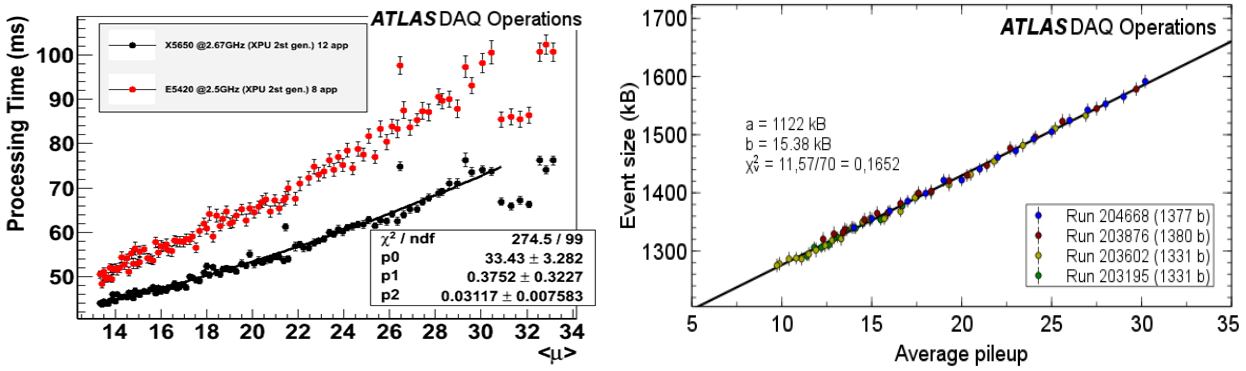


Figure 2: The effect of pileup <$\mu$> (the average number of proton-proton interactions per bunch crossing in a given data taking run) on (left) L2 processing time and (right) event size. [6]

to 400-600 Hz in Run 1. Taken together with the potentially increasing event size due to pileup, the motivation was in place for a system-wide upgrade.

Performance studies [7] showed that the existing ROS could only just meet the above requirements while exhausting almost all spare hardware, leaving little contingency for expansion within Run 2. The space needed to satisfy the required growth in number of ROS PCs was also a potential problem. Furthermore, concerns existed over hardware obsolescence, with the ROBINs using increasingly uncommon 64-bit PCI, as well as having fixed processing potential due to a non-upgradeable onboard power PC chip for event management. The size of the memory buffers on these boards (64 MB per optical link) was sufficient to hold data for expected event sizes for the time required by the HLT farm to complete processing. However, should the farm be expanded further in size this might cease to be the case. As such, a solution with greater expandability in future was desirable. The new requirements also posed a challenge for the HLT itself, which was to handle significantly more complex decisions and increased data rates.

In order to overcome the challenges of Run 2 it was decided that both the ROS and HLT systems required a major overhaul. The consequence of this was also a matching overhaul of the dataflow networks connecting the two systems, as well as the final event building and logging components. These networks would be required to handle increasing volumes of data while also facilitating the implementation of new standards and techniques to enable the successful upgrades of the ROS and HLT. In the following sections the upgrades to the three major components (ROS, network and HLT) will be discussed in detail. Finally, results of performance studies demonstrating the effectiveness of the upgrade will be presented. A diagram of the revised system is shown in Figure 3.

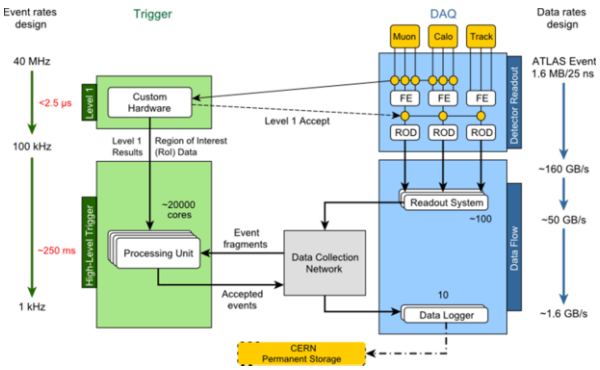


Figure 3: ATLAS TDAQ System in Run 2.

## 4. ROS Upgrade

The upgrade of the ROS focussed on increasing the density of the system (i.e. the number of links that can be handled in the same amount of server space) as well as increasing the data rates and volumes to be processed. Furthermore, the overall buffering capacity was upgraded to allow for future increases in requirements due to expansion of the HLT. Taking all of this into account, the design requirements for the new system were to buffer an input data rate of 100 kHz up to an average size of 1.6 kB per input link while also being able to read out 50% of this data to the HLT with no loss of performance. This is to be compared to a 75 kHz input rate with 10-15% readout in Run 1 up to the same event size.

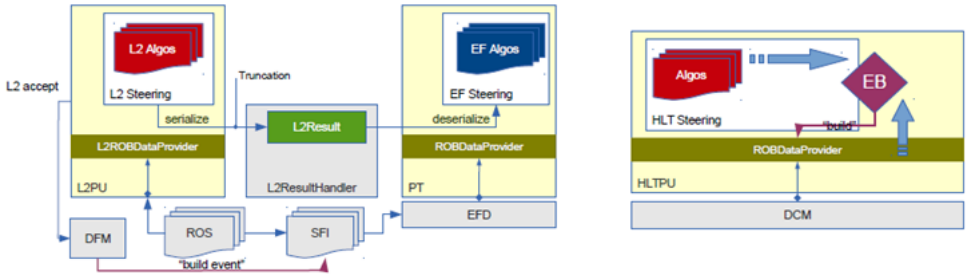


Figure 5: Run 1 HLT conceptual structure (left) alongside the Run 2 merged structure (right). Note that the ROS is still present but excluded from the right hand side of the diagram. ROI based event building continues in Run 2 but the results are handled directly by the combined HLT farm

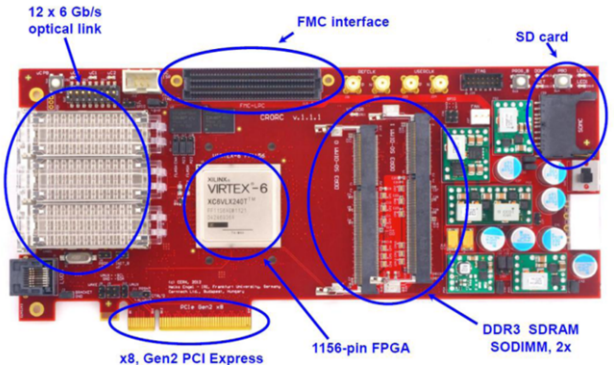


Figure 4: ATLAS RobinNP / ALICE C-RORC hardware.

### 4.1. Hardware and Firmware Upgrade

The primary element of the upgrade was the move away from the old ROBIN boards to new custom hardware known as the RobinNP. The RobinNP follows the same design philosophy as the ROBIN, but is able to handle four times as many input links in the same volume, while providing a factor of six increase in output bandwidth. The RobinNP functionality is implemented in firmware on a dedicated PCIe board originally developed by the ALICE Collaboration [8] and shown in Figure 4. The hardware features three quad-optical transceivers (QSFPs), as well as a high performance Xilinx Virtex 6 series FPGA and up to 16 GB of on-board memory capacity (on the RobinNP 8 GB is installed) through two SO-DIMM slots.

The choice of PCIe as the interface bus with the host system for the RobinNP was taken to ensure maximum compatibility with potential future hosts. Also, though the RobinNP itself only implements a Gen1x8 bus, the hardware allows Gen2x8. The current bus allows an in-practice maximum output bandwidth of 1.6 GB/s, compared to a requirement of 960 MB/s for 50% readout of 12 input links each providing a maximum nominal throughput of 160 MB/s. However, should ATLAS require a greater readout fraction in the future this can be achieved through an upgrade of the RobinNP firmware to use a Gen2x8 bus, effectively doubling the output bandwidth and allowing 100% readout with no new hardware expenditure.

Another key feature of the RobinNP is the decision to remove event processing and management away from an on-board chip and into the CPU of the host PC. Thus future performance improvements can be achieved by the relatively cheap upgrade of the host CPU compared to the cost of re-working custom hardware.

The RobinNP is hosted by a new generation of server-class machines, chosen to occupy half the vertical profile of their predecessors (2U vs the original 4U). Each machine hosts a single 6-core Xeon Ivy-Bridge grade 3.5 GHz six core CPU [9] and 16 GB of RAM. The network capacity of the new machines is also a significant increase on the older versions. Each machine now supports four 10 GbE optical Ethernet ports, which can be run in individual or bonded configurations. This is to be compared to two 1 GbE connections in the previous generation. Typically each new ROS PC hosts two RobinNP cards, though some variation may occur depending on individual detector sub-system requirements.

### 4.2. Software Upgrade

Alongside the changes to ROS hardware, a major overhaul of all dataflow software has been performed. This included not only the integration of the processing and management features previously performed on-board the ROBIN but also the replacement of the old protocols used for interaction with the HLT by implementation of industry standard asynchronous I/O based on the boost software library [10].

These changes were implemented within a broader redesign of the threading model of the system to reduce

wasted CPU cycles with improved inter-thread communication (based on pipelining and signalling rather than parallelism and polling) and exchange of command information with the hardware through use of an innovative interrupt coalescence protocol.

## 5. High-Level Trigger Upgrade

The upgrade of the HLT focused firstly in increasing processing capacity through upgrading existing servers within the farm and secondly on a major conceptual change to significantly improve efficiency.

The first component of the upgrade is an ongoing program that should see the number of available cores increase from 15,000 in Run 1 to at least 20,000 during Run 2. Further upgrades to core number are also possible during the run as required. More significantly, the entire L2 - EF structure has been simplified and the two processing steps conceptually merged as shown in Figure 5. Thus where previously L2 algorithms on a dedicated farm seeded processing in a separate EF farm, in Run 2 there will be one common HLT farm with each node capable of performing all processing steps.

The merger of the two HLT steps has been achieved through a complete rewrite of many of the individual algorithms that previously ran on the two farms. A single 'data collection manager' (DCM) process running on each HLT node orchestrates the data flow from the ROS through to the HLT processing units, event building processes and finally the data logging system (Sub-Farm Output, or SFOs).

The benefits of the merger can be felt in several areas. Firstly, there is no longer a need to transfer event data from one farm to another, or worse re-request the data from the ROS, resulting in a significant throughput saving. Secondly, whereas previously the split resulted in particular cores in the farms only ever being tasked with particular processes, the new system allows all cores to perform all HLT processes. This allows much more efficient resource distribution and load balancing. The increased capacity and flexibility of the system will also allow event building to occur faster than the previous maximum rate of 7 kHz.

## 6. Dataflow Network Upgrade

The conceptual changes to the HLT, as well as the increased data logging rate and throughput requirements of the upgraded ROS, have also mandated a major upgrade of the dataflow network. The most significant aspect to this is the obsolescence of the network layer

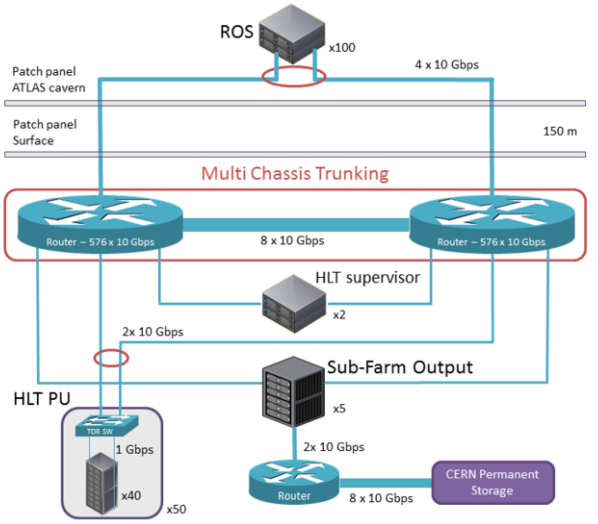


Figure 6: Dataflow network in Run 2, demonstrating a new single data collection layer with 10 GbE connectivity throughout the backbone and with the new ROS. The HLT Supervisor assigns events based on ROI data to the HLT PUs, which are software processes running on the HLT farm processors. Finally, event passing the whole selection are routed to permanent storage via the Sub-Farm Output (SFO)

transferring data from L2 to EF via the event builder, due to the logical merge of these functions into one farm. This leaves a single dataflow network.

The redesigned network is shown in Figure 6. In Run 1 10 GbE connectivity was implemented between the top level data collection and back end switches, the racks housing the ROS PCs and the concentrators serving individual HLT nodes and SFOs. The new system has eliminated the back-end network, but extended 10 GbE connectivity to individual ROS PCs as well as to the new generation of SFOs. The connection between the SFOs and the permanent data storage system, known as CASTOR, has also been upgraded to four times its previous bandwidth. Each ROS PC now has 2x10 GbE connections between it and each core router (i.e. a total of 40 GbE output per PC). Each HLT supervisor node, as well as the HLT systems themselves, are connected directly to each core router via 10 GbE connections.

The overall capacity of the routers allows for almost a factor of two increase in throughput above what is expected at the start of Run 2, thus allowing for a large increase in the number of HLT server racks and ROS PCs without the need for further overhaul of the network. This flexibility is expected to allow the system to scale to accommodate even the most extreme evolution in performance requirements during Run 2. Finally, new load balancing and traffic shaping protocols [5] will allow better distribution of data throughout the system.

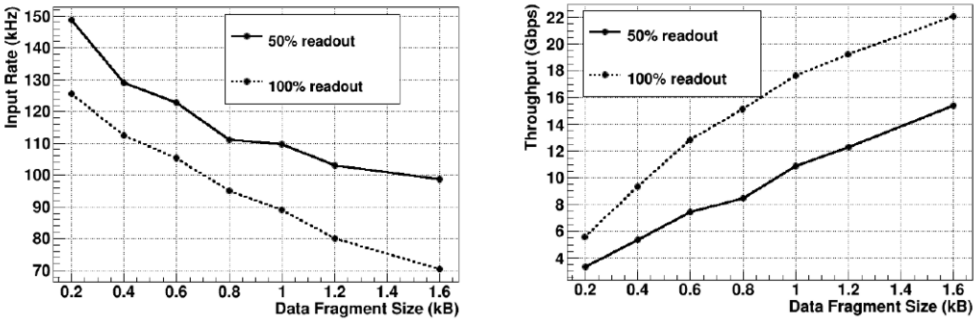


Figure 7: Results of performance benchmarks for next generation ROS PC hosting two RobinNPs for two different readout scenarios. On the left is the maximum possible L1 trigger rate and on the right the measured output bandwidth from the PC, both plotted against increasing event fragment size [6].

## 7. Upgrade Performance Studies & Status

The upgrade to the ROS was completed at the start of the final quarter of 2014. Detailed performance studies with the upgraded TDAQ system are currently under way. Preliminary results (as presented in Figure 7) suggest that the upgraded ROS meets the performance requirements for Run 2. The system as is being deployed also provides a significant improvement in density. Whereas in Run 1 there were 150 PCs servicing 1600 links the new system will service 1800 links with only 98 PCs. The relative ease of further upgrades to the system, from the RobinNP output bandwidth to the host CPUs, as well as the savings due to density, should allow the system to scale effectively during Run 2 and meet the changing requirements throughout the run period.

The HLT-merging process and dataflow network upgrade have been successfully completed, with the combined farm already undergoing continuous testing and use for calibration, performance and re-start studies as ATLAS prepares to return to data taking [11]. The work to add cores to the farm will continue into and throughout 2015 as per system requirements.

## 8. Conclusions

The TDAQ system of the ATLAS experiment has undergone a comprehensive upgrade during the 2013/2014 shutdown period in order to meet the increased requirements of LHC Run 2. These requirements stem partly from increased rates and event sizes due to the collision environment, and partly from the enhanced capabilities of the detector.

The upgrade work has ranged from renewed hardware, providing denser and higher performance components, but also comprehensively rewritten software. The algorithms governing selection and dataflow in the HLT have been revised and merged. The architecture of the software governing the ROS has also been substantially re-written, with enhanced multi-threading and industry standard libraries for interprocess communication, as well as an innovative high performance hardware interface.

Work is currently ongoing to finalise the upgrade by the end of 2014, leaving several months for testing and validation before the envisaged return of LHC collisions in the second quarter of 2015. Performance studies suggest the upgrades leave the system in a good position to meet the challenges of Run 2, and with the ability to scale effectively throughout the run period.

## References

[1] ATLAS Collaboration, 2008 JINST 3 S08003.
[2] O. Boyle, et al., S-LINK, a Data Link Interface Specification for the LHC Era, Proceedings of the Xth IEEE Real Time Conference (1997).
[3] R. Cranfield, et al., The ATLAS ROBIN, 2008 JINST 3 T01002.
[4] R. Blair, et al., The ATLAS High Level Trigger Region of Interest Builder, 2008 JINST 3 P04001.
[5] A. Negri, Evolution of the Trigger and Data Acquisition System for the ATLAS experiment, 2012 J. Phys.: Conf. Ser. 396 012033.
[6] `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/ApprovedPlotsDAQ`.
[7] R. Bartoldus, et al., ATLAS TDAQ System Phase-I Upgrade Technical Design Report, CERN-LHCC-2013-018.
[8] H. Engel, U. Kebschull, Common read-out receiver card for ALICE Run2, 2013 JINST 8 C12016.
[9] Intel Corporation, E5-1650 v2 processor, `http://ark.intel.com/products/75780`.
[10] C. Kohlhoff, Boost.Asio, `http://www.boost.org`.
[11] F. Pastore, The ATLAS Trigger System: Past, Present and Future, Proceedings of the 37th International Conference on High Energy Physics (2014).