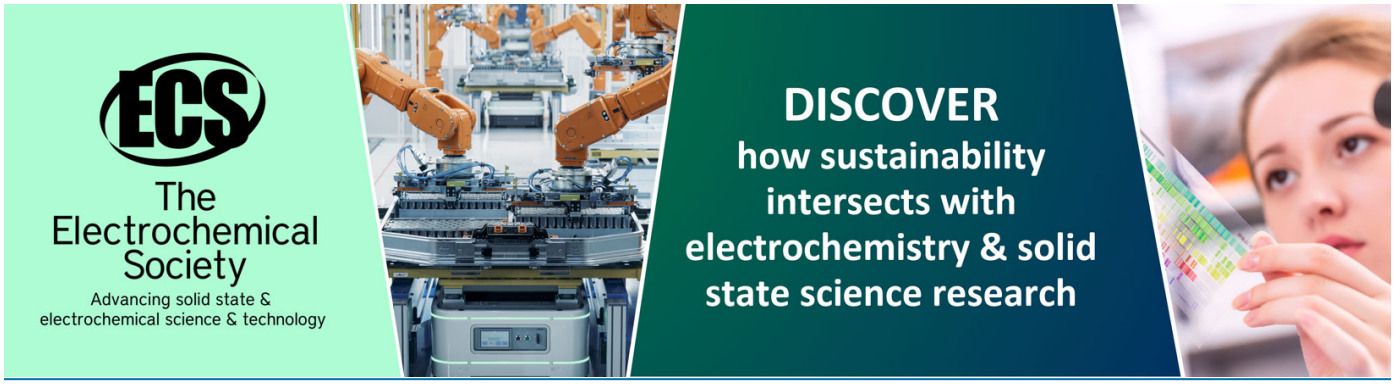
The ATLAS Data Acquisition System in LHC Run 2



IOP Conf. Series: Journal of Physics: Conf. Series **898** (2017) 032017 doi:10.1088/1742-6596/898/3/032017

The ATLAS Data Acquisition System in LHC Run 2

William Panduro Vazquez[1] on behalf of the ATLAS Collaboration

[1]Department of Physics, Royal Holloway, University of London, Egham, Surrey, TW20 0EX

E-mail: j.panduro.vazquez@cern.ch

Abstract. The LHC has been providing pp collisions with record luminosity and energy since the start of Run 2 in 2015. The Trigger and Data Acquisition system of the ATLAS experiment has been upgraded to deal with the increased performance required by this new operational mode.

The dataflow system and associated network infrastructure have been reshaped in order to benefit from technological progress and to maximize the flexibility and efficiency of the data selection process. The new design is radically different from the previous implementation both in terms of architecture and performance, with the previous two-level structure merged into a single processing farm, performing incremental data collection and analysis. In addition, logical farm slicing, with each slice managed by a dedicated supervisor, has been dropped in favour of global management by a single farm master operating at 100 kHz. This farm master has also been integrated with a new software-based Region of Interest builder, replacing the previous VMEbus-based system. Finally, the Readout system has been completely refitted with new higher performance, lower footprint server machines housing a new custom front-end interface card.

Here we will cover the overall design of the system, along with performance results from the start-up phase of LHC Run 2.

1. Introduction

The ATLAS experiment [1], based at the Large Hadron Collider (LHC) at CERN, Switzerland, underwent a 2-year maintenance and upgrade process in preparation for LHC Run 2, starting in 2015. The new data-taking period saw the LHC colliding particles at record energy and intensity, as well as the integration of new ATLAS detector components, all of which placed greater demands on the performance of the ATLAS data acquisition (DAQ) system. In Run 2 to date a peak luminosity of 1.37×10^{34} cm^{-2} s^{-1}, higher than the LHC design value, has already been achieved. By the end of the data-taking period in 2018 it is possible that this value could reach 2×10^{34} cm^{-2} s^{-1}. One of the main factors behind this increased intensity is the increase in the number of interactions occurring per LHC bunch crossing, known as pileup. Events featuring increasing pileup are larger and take longer to process, with a commensurate increase in the amount of buffering capacity required, while also posing an algorithmic challenge. While the pileup increase has been mitigated somewhat by the decrease in the time between LHC bunch crossings to 25 ns the value is expected to increase over time. The pileup values observed in Run 2 up to September 2016 are shown in Figure 1.

Alongside the increasing demands of the LHC collision environment, ATLAS sought to increase the event rates processed by various stages of the trigger system to optimise sensitivity to interesting physics signatures. The event rate from the Level 1 hardware trigger system

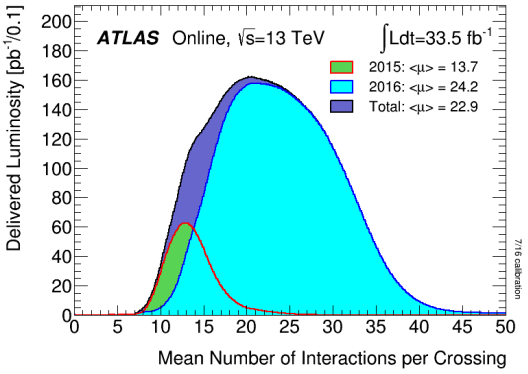


Figure 1. Pileup distribution observed for Run 2 up to September 2016 [2].

increased from 70 kHz in Run 1 to a target of 100 kHz, with the rate accepted by the HLT increasing from 600 Hz to 1.5 kHz. Commensurate with this, and the larger event size due to pileup, the expected throughput from the system to permanent storage increased from 1.0 GB/s to 1.5 GB/s.

The focus of this paper will be the upgrade of the ATLAS DAQ 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), Region of Interest Builder (RoIB), High-Level Trigger (HLT) and Data Logger (Sub-Farm Output, or SFO), as well as the dataflow network through which they interact. An overview of the upgraded system is presented in Figure 2.

2. High Level Trigger

The upgrade of the HLT focused on a major conceptual change to significantly improve efficiency. Alongside this, there is a separate programme to increase processing capacity through replacing older servers within the farm with newer models at the end of their operational life.

The hardware replacement programme has seen the number of motherboards increase to approximately 2000 by the end of 2016. The newest machines, making up 50% of the farm, are based on modern Intel Haswell [3] architecture-based CPUs. More significantly, the structure of the farm has been simplified. What were previously two separate farms [4] handling Regions of Interest and full event filtering were merged into one system.

The merger of the two HLT steps was achieved through a complete rewrite of many of the individual algorithms that previously ran on the two farms, with functionality combined, as demonstrated schematically in Figure 3. 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 SFOs.

The benefits of the merger can be felt in several areas. Firstly, data throughput requirements are reduced as there is no longer a need to transfer event data from one farm to another, or re-request the data from the ROS. Secondly, whereas the split system resulted in particular cores in the farms only ever being tasked with given processing types, potentially leading to under-utilisation, the new system allows all cores to perform all HLT processes. This enables much more efficient resource distribution and load balancing. The increased capacity and flexibility of the system also makes it possible to build events gradually by requesting data during processing, rather than having a dedicated event building step. With the old architecture such a step was necessary, and was implemented with dedicated machines. The performance of these systems limited the event building rate to 7 kHz. With the new architecture limitations may come from readout or network bandwidth, but at a much higher threshold than the previous system. No

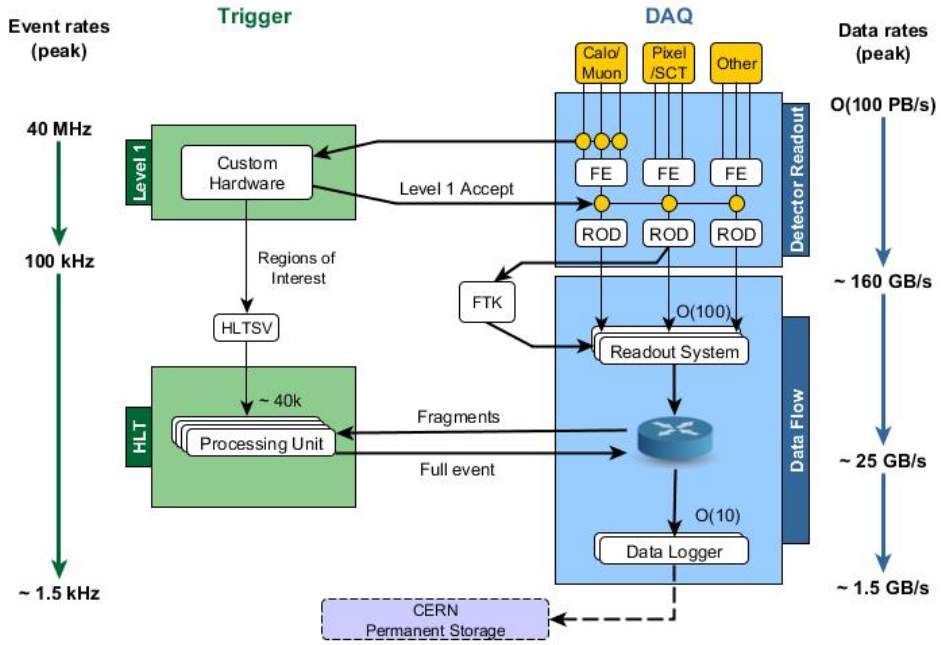


Figure 2. The ATLAS DAQ System in LHC Run 2. Events passing the Level 1 hardware trigger (top left) are passed to the HLT (bottom left) via the farm supervisor node (HLTSV), now including assembly of Regions of Interest. Simultaneously to this, event data from the detector front-end electronics systems are sent to the Readout System (ROS) via optical links from the Readout Drivers (RODs) in response to a Level 1 trigger accept signal. These data are then buffered in the ROS and made available for sampling by algorithms running in the HLT. Once the HLT accepts an event it is sent to permanent storage via the Data Logger.

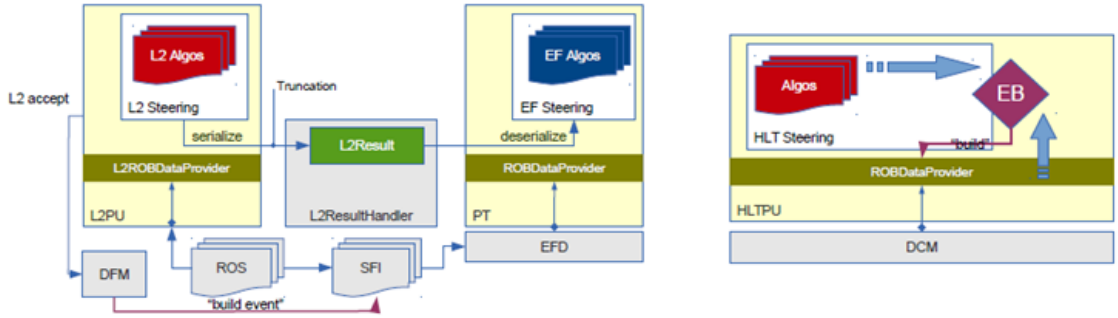


Figure 3. Conceptual diagram showing the merger of logical blocks within the combined High Level Trigger. The old architecture is shown on the left and the new one on the right.

saturation has yet been observed in regular operations.

3. Readout System (ROS)

The ROS is responsible for buffering event data passing Level 1 trigger selection for the period of time taken for the HLT to perform its more detailed processing. The ROS consists of a set of

commodity server machines hosting custom-built I/O cards. These receive data from detector front-end electronics over one of approximately 1850 optical links and store them in internal memory buffers. During processing the HLT will request data from the ROS as needed before either accepting an event, and sending it to the SFOs for permanent storage, or requesting its deletion from the ROS buffers.

The upgrade of the ROS [5] 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 farm size. The new system is able to buffer an input data rate of 100 kHz up to an average event fragment size of 1.6 kB per input link while also being able to send 50% of the 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.

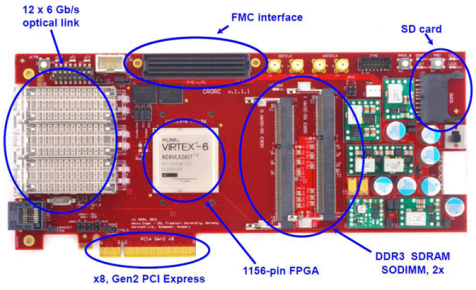


Figure 4. ATLAS RobinNP / ALICE C-RORC hardware.

The primary element of the ROS upgrade, installed for the start of Run 2, was the design and implementation of an upgraded version of the custom I/O board, known as the RobinNP. The new component consists of ATLAS-specific firmware installed on a PCIe board originally developed by the ALICE Collaboration [6] and shown in Figure 4. The hardware features a high performance Xilinx Virtex 6 series FPGA and 8 GB of on-board DDR3 RAM. The current PCIe bus implementation allows an in-practice maximum output bandwidth of 1.6 GB/s, potentially upgradable to 3.2 GB/s should requirements evolve.

The RobinNP follows the same design philosophy as its predecessor the ROBIN [7], but is able to leverage its upgraded hardware to handle four times as many input links in the same volume, while providing a factor of six increase in output bandwidth. Data processing and management features previously performed on-board the ROBIN have been migrated to the software running on the host PC to improve maintainability and reduce hardware complexity. Transfers of event metadata from RobinNP to host are handled via a novel mechanism based on automatic mirroring of firmware structures into software via Direct Memory Access (DMA), with signalling to the host via coalesced interrupts.

The new boards are hosted in 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 [8] 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 10GbE optical ports, which are run in a bonded configuration. This is to be compared to two GbE connections in the previous generation. Typically each new ROS PC hosts two RobinNP cards, servicing up to 24 input links, though some variation may occur depending on individual detector sub-system requirements.

The performance of the upgraded Readout System has comfortably satisfied the ATLAS requirements for Run 2, with significant headway within which to accommodate evolving

conditions looking toward LHC Run 3. Figure 5 demonstrates the impact of the increased memory capacity by comparing the limits imposed by the Run 1 system on HLT farm size (a), as well as the evolution of buffer occupancy throughout a run (b).

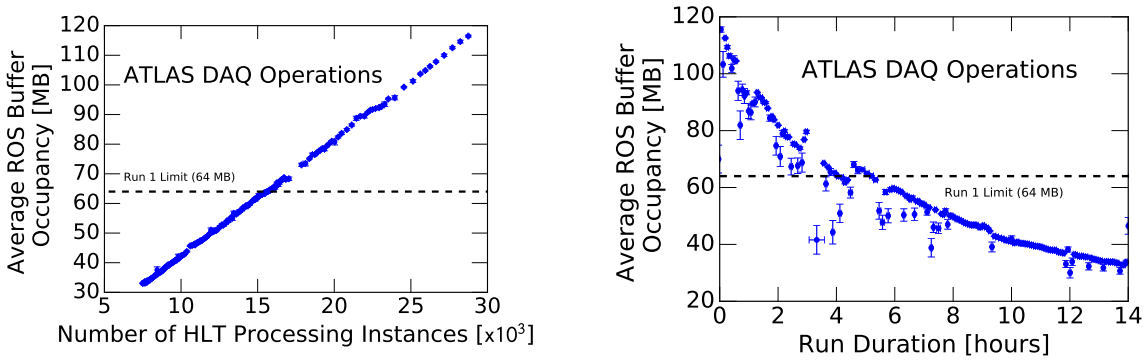


Figure 5. (a) Average ROS buffer occupancy for a single readout channel plotted against number of HLT application instances in use in the HLT farm. As can be seen the Run 1 limitation of 64 MB would have limited the farm size to approximately 15,000 HLT application instances, compared to the currently active 40,000. Each channel now has 0.67 GB of buffer space available, meaning no limitations on farm size in the foreseeable future. (b) The evolution of buffer occupancy for a typical single readout channel during an ATLAS run. Under these conditions the Run 1 system would be unable to cope with required buffering load for a significant portion of the run. Data for both plots were taken from a typical ATLAS run from October 2016, with a peak luminosity at start of run of $1.25\times10^{34}\text{cm}^{-2}\text{s}^{-1}$

4. Region of Interest (RoI) Builder

Event processing in the ATLAS HLT is driven by the study of detector Regions of Interest (RoI) flagged by the Level 1 hardware trigger system. RoI information from the different Level 1 system components are then assembled into a single entity via a dedicated system before being assigned to a free HLT node via a farm supervisor system for processing. In Run 1 this Region of Interest Builder (RoIB) system was implemented in custom hardware based on VMEbus technology, communicating with multiple supervisor nodes. Experience indicated that this system would place performance limitations on the complexity of RoI information which could be exploited in Run 2, while also becoming subject to maintenance and obsolescence concerns given the age of the custom hardware.

In Run 2 the RoIB system was replaced with an integrated design based on ROS technology [9]. RoI data from the Level 1 system components are now received by a RobinNP card hosted in a standard ROS server PC and automatically transferred to the host for further processing. The building of RoIs then proceeds entirely in software, before transfer to a single farm supervisor process, now running on the same host, for assignment to a HLT node. The evolution of the system is presented in Figure 6.

The upgraded RoIB was successfully deployed at the start of 2016 and has already made it possible to handle more detailed RoI information at high rate, while also enabling an expansion of monitoring capability and downstream error reporting. The design also carries with it sufficient spare capacity to handle any realistic evolution in performance requirements.

5. Data Logger (SFO)

The data logger, referred to for historical reasons as SFO, is responsible for aggregating large volumes of event data accepted by the HLT for further processing into a standardised file format

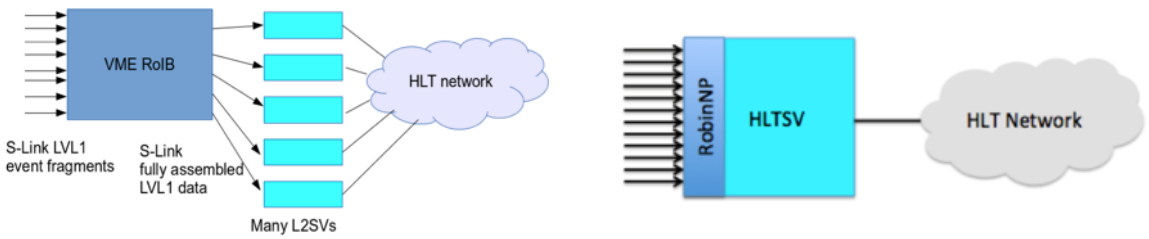


Figure 6. Schematics of RoIB system in Run 1 (left) and Run 2 (right).

for transfer to the ATLAS Tier0 processing centre for further analysis and permanent storage using the EOS standard [10]. For Run 2 the SFO system was upgraded with state-of-the-art storage and CPU technology and a reworked multi-threading processing model. The new system contains six dedicated nodes with an overall buffer capacity of 340 TB, meaning that up to 24 hours of collision data can be stored locally before transfer to EOS. The system is extensible, with new nodes able to be integrated to scale up performance as needed, but the new installation is already performing beyond the design throughput of 1.5 GB/s, as shown in Figure 8.

6. Networking

The conceptual changes to the HLT, as well as the increased data logging rate and throughput requirements of the upgraded ROS, also mandated a major upgrade of the ATLAS dataflow network. The most significant aspect to this is the obsolescence of the network layer transferring data between the two component farms of the HLT via the event builder machines, due to the logical merger of these functions into one farm. This leaves a single dataflow network, as shown in Figure 7(a).

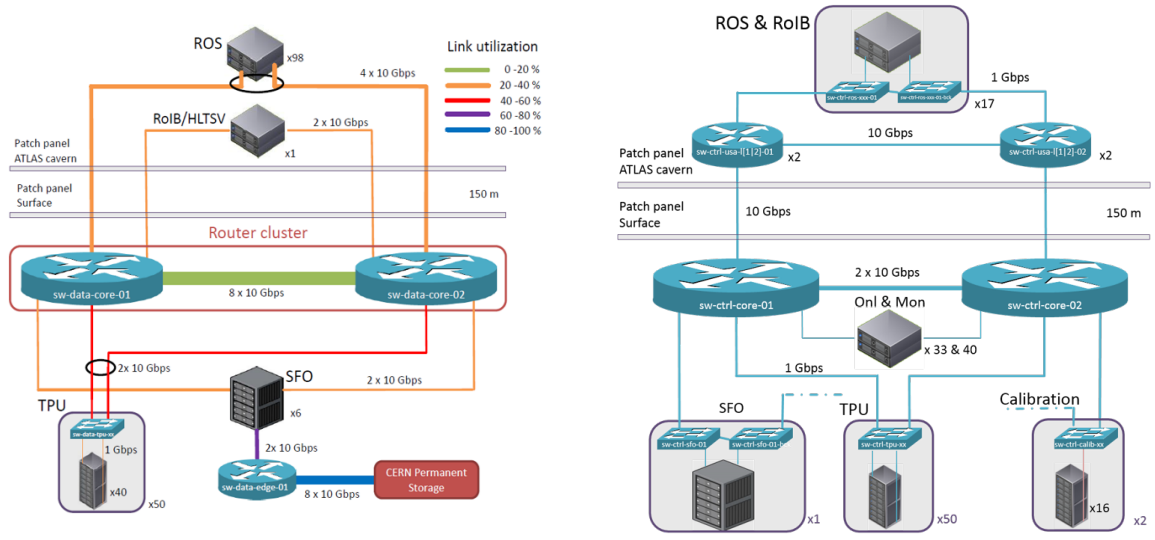


Figure 7. (a) Dataflow network in Run 2, demonstrating a new single data collection layer with 10GbE connectivity throughout the backbone and with the new ROS. (b) Control network in Run 2, implementing active-backup redundancy at all critical levels.

The Run 1 system saw 10GbE connectivity 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 extends 10GbE connectivity to individual ROS PCs

as well as to the new generation of SFOs and HLTSV. The connection between the SFOs and the permanent EOS data storage system has also been upgraded to four times its previous bandwidth. Each ROS PC now has 2×10GbE connections to each core router (i.e., a total of 40 GbE output per PC). Dataflow across the backbone routers is logically combined using Multi-Chassis Trunking [11] technology, providing each ROS PC with active-active redundancy in the case of link failure. Each HLT rack was also upgraded with a deep buffer top-of-rack switch to significantly reduce event building time by reducing the rate of packet drops and retransmissions.

The overall capacity of the routers allows for almost a factor of two increase in throughput above what was 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 [12] make possible better distribution of data throughout the system. The impact of the redesign of the HLT and dataflow network overhaul can be seen in Figure 8, which shows the HLT input bandwidth in a typical ATLAS run for different collision luminosities.

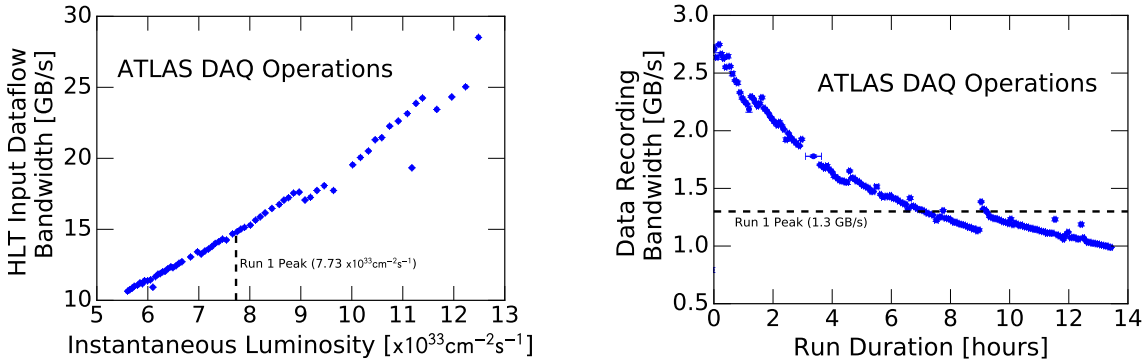


Figure 8. (a) HLT input bandwidth as a function of instantaneous luminosity for a typical ATLAS run. The largest instantaneous luminosity delivered in Run 1 is shown by the vertical dashed line. Comparison with the previous system is difficult given the different designs, but for reference it can be noted that the largest achievable bandwidth into the old Level-2 farm was of order 5 GB/s with the Event Filter farm accepting 10 GB/s [4]. (b) Evolution of SFO output bandwidth over a typical ATLAS run. The maximum performance value observed in Run 1 is shown by the horizontal dashed line. Data for both plots were taken during a run in October 2016, with a peak luminosity at start of run of $1.25 \times 10^{34}\text{cm}^{-2}\text{s}^{-1}$

Alongside the dataflow network, ATLAS maintains a separate lower bandwidth control network layer, shown in Figure 7(b), for the propagation of command information and non-bulk dataflow communication. The control network was also overhauled ahead of Run 2 to facilitate the deployment of active-backup redundancy at the level of the core routers, top-of-rack switches and infrastructure links. The new system no longer requires dedicated on-call support, allowing resources to be allocated elsewhere.

7. Software

Alongside the hardware changes described in this paper, a major redesign of all dataflow software was performed in preparation for Run 2. The central data request and messaging interface between the HLT and the ROS was re-implemented using an industry standard asynchronous I/O based on the Boost [13] software library. Multi-threaded processing models throughout the

system were overhauled to exploit improvements in C++11, as well as Intel Threaded Building Blocks [14] technology.

At a higher level a wider overhaul took place of control, configuration and monitoring software encompassing the full range of DAQ system components [15] [16] [17] [18]. During 2016 a review also took place of software build and revision management procedures, leading to the adoption of new Git [19] and CMake [20] based workflows. A full migration was also made to 64-bit software throughout the experiment, facilitated by the replacement of the final instances of legacy VMEbus crate control hardware.

8. Conclusions and Outlook

The ATLAS DAQ system has undergone an extensive program of improvements in order to meet the challenges of LHC Run 2. Key among the improvements are an overhauled readout system with improved buffering capacity and throughput; updated Region of Interest and data logging systems and a redesigned HLT farm architecture, merging two previously separate layers. Connecting all of these systems is a high performance dataflow network based on 10GbE technology, making use of modern multi-chassis trunking techniques for improved redundancy.

The upgraded DAQ system was successfully integrated in time for Run 2, and has delivered high performance and availability in all subsequent data taking. In 2016 the measured uptime fraction during LHC colliding beams with all ATLAS detectors configured was 91.7%, with the majority of the downtime coming from factors beyond the DAQ system, such as the ongoing integration of new detector components and the challenging readout environment for detector front-end electronics provided by the record luminosities. Measurements to date indicate that there is enough spare capacity in the DAQ system to satisfy ATLAS performance requirements through to the end of Run 2 and beyond.

References

[1] ATLAS Collaboration 2008, *JINST* **3** S08003.
[2] ATLAS Collaboration 2016, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2
[3] Intel Corporation, E5-2680 v3 processor, http://ark.intel.com/products/81908
[4] ATLAS TDAQ Collaboration 2016 *JINST* **6** P06008.
[5] Borga A *et al*, Evolution of the ReadOut System of the ATLAS experiment PoS TIPP2014 (2014) 205.
[6] Engel H and Kebschull U, Common read-out receiver card for ALICE Run2 2013 *JINST* **8** C12016.
[7] Cranfield R *et al*, The ATLAS ROBIN 2008 *JINST* **3** T01002.
[8] Intel Corporation, E5-1650 v2 processor, http://ark.intel.com/products/75780
[9] Abbott B *et al*, The evolution of the Region of Interest builder for the ATLAS experiment at CERN 2016 *JINST* **11** C02080.
[10] CERN 2013 http://eos.readthedocs.io/en/latest/
[11] Brocade Communications Systems, http://www.brocade.com/content/dam/common/documents/content-types/whitepaper/brocade-multi-chassis-trunking-wp.pdf
[12] Negri A, Evolution of the Trigger and Data Acquisition System for the ATLAS experiment, 2012 J. Phys.: Conf. Ser. 396 012033.
[13] Kohlhoff C, Boost.Asio http://www.boost.org.
[14] Intel Corporation, Threaded Building Blocks, https://www.threadingbuildingblocks.org/
[15] Lee C *et al* 2016, ATLAS TDAQ System Administration: Master of Puppets, Proceedings of CHEP 2016 J. Phys.: Conf. Series.
[16] Fazio D *et al* 2016, Integrated monitoring of the ATLAS online computing farm, Proceedings of J. Phys.: Conf. Series.
[17] Soloviev I *et al* 2016, A web-based solution to visualize operational monitoring data in the Trigger and Data Acquisition system of the ATLAS experiment at the LHC, Proceedings of CHEP 2016 J. Phys.: Conf. Series.
[18] Aleksandrov I *et al* 2016, The Resource Manager the ATLAS Trigger and Data Acquisition System, Proceedings of CHEP 2016 J. Phys.: Conf. Series.
[19] Software Freedom Conservancy, Git, https://git-scm.com/
[20] Kitware, CMake, https://cmake.org/