THE ATLAS DATA ACQUISITION SYSTEM IN LHC RUN 2

M. E. Pozo Astigarraga, on behalf of the ATLAS Collaboration

CERN, CH-1211 Geneva 23, Switzerland

E-mail: eukeni.pozo@cern.ch

The LHC has been providing proton-proton collisions with record intensity and energy since the start of Run 2 in 2015. In the ATLAS experiment the Data Acquisition is responsible for the transport and storage of the more complex event data at higher rates that the new collision environment implies. Data from events selected by the first level hardware trigger are subject to further filtration from software running on a commodity server farm. During this time the data are transferred from detector electronics across 1900 optical links to custom buffer hardware hosted across 102 commodity server PCs, and then across the system for processing by high bandwidth network at an average throughput of 30 GB/s. Accepted events are transported to a data logging system for final packaging and transfer to permanent storage, with an average output rate of 1.5 GB/s. The whole system is actively monitored to maximise efficiency and minimise downtime. Due to the scale of the system and the challenging collision environment the ATLAS DAQ system is a prime example of the effective use of many modern technologies and standards in a high energy physics data taking environment, with state of the art networking, storage and real-time monitoring applications.

Keywords: Data acquisition, networks, storage, computing.

1. Introduction

The Large Hadron Collider (LHC) located at CERN is a circular particle accelerator providing proton-proton collisions 40 million times per second. Protons circulate in bunches that interact at several points along the LHC ring where different particle detectors study the results of the interactions. ATLAS (A Toroidal LHC ApparatuS) is one of the general purpose detectors [1].

The ATLAS Data Acquisition (DAQ) system was designed to operate during the LHC Run 1 period (2009-2013). Due to the upgrades performed on the LHC during the Long Shutdown (2013-2015), the DAQ system had to be upgraded to cope with the new operational conditions. Table 1 shows how several LHC parameters have changed significantly for the Run 2 period (2015-2019). In particular, the average number of concurrent proton-proton interactions in a given bunch crossing, the pileup, has increased by more than 50%.

	Run 1	**Run 2**
Center-of-mass energy	8 TeV	13 TeV
Peak luminosity	$7 \times 10^{33} cm^{-2} s^{-1}$	$1.74 \times 10^{34} cm^{-2} s^{-1}$
Average event pileup	20.7	32.2
Peak event pileup	40	60
Proton bunch crossing rate	20 MHz	40 MHz

Table 1: Change of LHC operating conditions between Run 1 and Run 2 (as of September 2017)

Higher pileup implies that there are more hits in the detector making it more difficult to identify signal interactions from backgrounds, requiring additional resources for handling and processing by the trigger and DAQ systems. Furthermore, higher bunch crossing and trigger rates motivated the upgrade of the ATLAS DAQ system in Run 2. New developments in networking, storage and computing technologies allowed the DAQ system to improve the Run 1 system performance and meet the technical requirements of the Run 2 conditions. The result is a more robust system where the need for expert support has been reduced.

In the present document, we introduce in Section 2 the key data flow components and give an overview of their role and performance. Relevant developments in the DAQ control and monitoring software are also be presented in Section 3. Finally, the conclusion and future plans are described in Section 4.

2. Dataflow in the ATLAS Data Acquisition System

The DAQ system needs to operate at the Level-1 (L1) trigger output rate. The L1 trigger receives muon and calorimeter information in order to reduce the 40 MHz event rate to a maximum value of 100 kHz. It makes use of custom hardware to make a decision in 2.5 μs. The events passing the filter are read-out by the subdetector specific devices known as Read-Out Drivers (RODs).

The DAQ system receives and buffers event fragments from the RODs at around 160 GB/s and transports a subset of them to the High Level Trigger (HLT), which further reduces the event rate down to an average value of 1 kHz. The HLT is a set of software algorithms running on a computer farm made of around 2.000 nodes hosting up to 40.000 CPU cores. On each of these cores runs a Processing Unit (PU), which independently processes a given physics event. Finally, the accepted events are stored temporarily before sending them to permanent storage and ultimately to offline analysis.

The different components participating in the read-out, transport and storage of the events are known as dataflow components. In Figure we depict the relationship between these components and the L1 and HLT trigger systems.

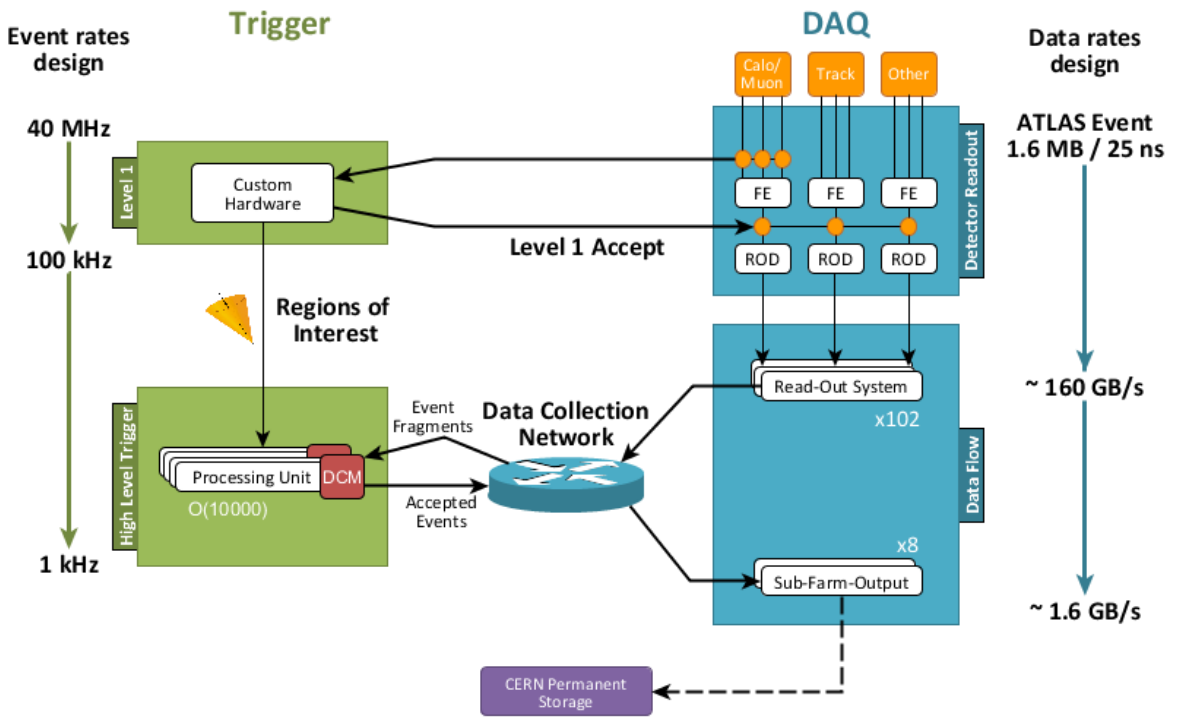

Figure 1. ATLAS Trigger and DAQ systems in Run 2

2.1. The Read-Out System

The Read-Out System (ROS) receives and buffers event fragments from the RODs and serves them to the HLT farm upon request of the PU in charge of the event filtering. For this purpose, a new custom read-out card developed by ALICE experiment but running custom ATLAS firmware has been used for Run 2. The ALICE card has been renamed as Robin-NP [2].

The Robin-NP is hosted in a commodity server connected to the data collection network. It receives event fragments at the L1 trigger output rate on 12 custom links and hosts them in internal buffers (8 GB). When data are requested, the fragments are moved to the server memory, processed by the server CPU and shipped by the Linux TCP/IP stack over four 10 Gigabit Ethernet links. The event fragments that are not read-out by the HLT farm are deleted directly from the Robin-NP buffers without passing through the system memory.

In total, close to 1900 custom links are connected to 102 ROS servers hosting one or two Robin-NP cards each. The ROSes need to handle an average of 160 GB/s of event data and serve a subset of it to the HLT farm.

2.2. Region of Interest Builder and High Level Trigger Supervisor

The Region of Interest Builder (RoIB) is a component that receives and assembles Regions of Interest (RoI) from fragments produced by L1 trigger sources. These are the L1 calorimeter, L1 muon, L1 topological trigger and the Central Trigger Processor. The RoIs contain the information about candidate trigger objects. The RoI records are passed to the High Level Trigger Supervisor (HLTSV) application, which in turn seeds the PUs in order to start the filtering of the event.

The L1 fragments are received by the RoIB at the L1 trigger rate on custom links connected to a Robin-NP card. With respect to Run 1, where the RoIB was implemented on custom hardware running on VME boards, the system has notably been simplified with the usage of the Robin-NP [3]. The robustness of the new system has increased removing the need for a dedicated on-call expert.

Besides the RoI assignment to the PUs, the HLTSV also sends the clear event fragments message to the ROS buffers once the event has been accepted or rejected. This is accomplished by sending a single multicast message to all the ROS servers at once. In case of problem during the event processing, the HLTSV can re-assign the RoI record to another PU after a configurable timeout.

2.3. The Data Collection Manager

In the Run 2 system the PUs rely on a specialized component for the event fragment collection from the ROSes known as Data Collection Manager (DCM). The DCM is a single-threaded C++ application serving all the PUs running on a given host. It uses *Boost ASIO* library to deal with the asynchronous I/O and *Boost Interprocess* for the Inter-Process Communication with the PUs.

The DCM implements a traffic shaping mechanism to control the amount of traffic injected in the network at a given moment, avoiding in this way instantaneous network saturation. The DCM maintains a set of credits that are consumed when fragments are requested and restored when the fragments are received. If no credits are available at a given moment, the requests are queued until enough credits are restored. It has been demonstrated that with the right number of credits in the DCMs the network throughput and the system performance can be maximized [4].

In addition, the DCM supports the duplication of accepted events going to different output streams (e.g. for calibration purposes) and data compression before event logging.

2.4. Event logging: the Sub-Farm Output

The event logging is performed by the Sub-Farm Output (SFO). The SFO system provides up to 48h of temporary storage for the accepted events. This temporary logging allows separating the detector operation from the availability of the CERN permanent storage system, which is not under control of the experiment. Background jobs copy event files to the permanent storage locations and delete them from the local disk only when they are safely on tape.

In total there are four Direct Attached Storage units connected to eight servers with multiple redundant data paths for fault tolerance and resiliency (see Figure). The SFOs offer a maximum concurrent read and write rate of 8 GB/s. In average, the system operates at a lower rate of around 3.2 GB/s [5].

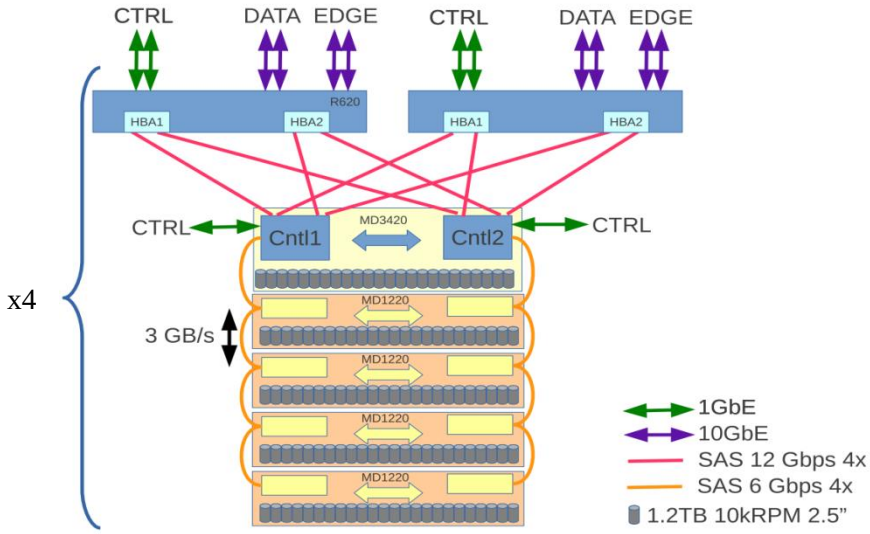

Figure 2. Internal architecture of a SFO unit

2.5. The Data Acquisition Network

The Data Acquisition Network connects all the previous components using a high throughput Ethernet network. In total, more than five hundred 10 Gigabit ports provide the necessary bandwidth and redundancy. In a typical ATLAS run, the aggregated throughput is above 30 GB/s at the beginning of the run and decays softly following the luminosity delivered by the LHC.

The ROS and RoIB server nodes located in the experiment underground area are connected with 150m long links to the surface core-network routers (Brocade MLXe devices). Each top-of-rack (ToR) switch aggregating HLT nodes within a rack is, in turn, connected with two 10 Gigabit uplinks to the routers. The SFO servers are connected in the same way with two 10 Gigabit uplinks each. In addition, the SFO output links are aggregated in the border router providing the long distance connectivity to the CERN data centre.

The main characteristic of the data acquisition network is the core router aggregation in a cluster as depicted in Figure . Using a Brocade proprietary protocol known as *Multi Chassis Trunking* it is possible to peer the two main routers and expose them to the rest of the network as a single device with twice the ports and packet switching capacity. This feature allows the transparent handling of any hardware failure in the network components, increasing the system efficiency and reducing the need of on-call support while the experiment operates.

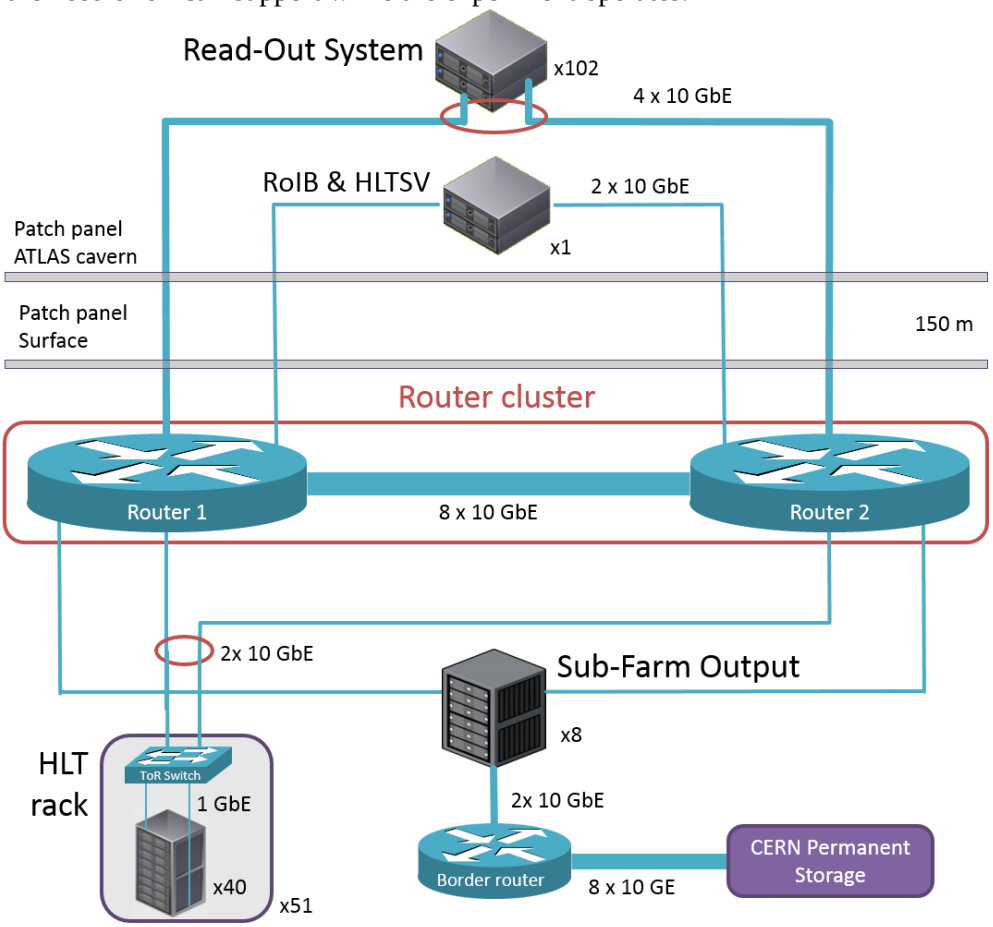

Figure 3. ATLAS data acquisition network

Ethernet networks suffer from packets drops. Nevertheless, in a data centre environment it is highly desirable that packets drops do not occur under any circumstance. In the DAQ particular case, the PUs need to wait for the data to be retransmitted by the TCP protocol with the consequent loss in CPU efficiency for the system. In practise, even if enough bandwidth is installed to cope with the highest estimated throughput, it is not possible to avoid instantaneous ToR switch buffer oversubscription due to the nature of the traffic [6]. This can produce a collapse of the network throughput, known in the literature as *TCP incast* [7]. Indeed, many sources, the ROS servers, send event fragments corresponding to a given event to the same destination, the DCM, in a reduced time frame.

In order to mitigate the TCP *incast* several strategies can be applied: first, the DCM's traffic-shaping mechanism described in Section 2.3 acts as an application level flow-control mechanism. Second, it is possible to use deep-buffer ToR switches with enough capacity to deal with any brief burst of packets addressed to a given output switch port. Third, following the market trend on Ethernet data centre switches, we can use dynamic shallow buffers provided that the total burst needed to be handled by the switch in the worst-case scenario does not exceed the total available capacity. All these mechanisms are applied separately where needed depending on the installed network device model.

3. Control and monitoring in the ATLAS Data Acquisition System

One of the main responsibilities of the DAQ system is to provide the software infrastructure to configure, monitor and control a data taking session. More than 40.000 applications running on different platforms and connected to a dedicated control network need to be managed efficiently by the DAQ software.

In such a big environment, hardware and software failures will happen and the system needs to deal with them quickly to minimize the impact on the data taking and to keep the system efficiency high. In addition, it is desired that the operators of the experiment are capable of dealing with problems that are not necessarily under their domain of expertise.

In Run 2, a significant effort has been made to incorporate new techniques for automating many of the procedures that would require expert intervention otherwise. The introduction of a Complex Event Processing (CEP) engine has proven to be of great help to achieve this objective. New monitoring tools have also been developed in order to provide an enhanced view of the system. We give next an overview of such improvements.

3.1. Complex Event Processing engines in the DAQ system

Complex Event Processing refers to a method for finding complex patterns in streams of process and system monitoring data and taking associated action. It is possible to establish different types of relationships between incidents, such as correlation, aggregation or temporal causality in a well-defined sliding window.

The DAQ system uses a CEP engine based on an Open Source solution known as *Esper* [8]. *Esper* provides a rich Event Processing Language with an SQL-like syntax, which allows easily defining rules and associated actions.

Two different applications have been developed for Run 2 operation. The first one, the Central Hint and Information Processor (CHIP) [9], has become a key component for the ATLAS run control software. CHIP interacts directly with the main process of the hierarchy under which all the ATLAS online applications are running. CHIP performs anomaly detection and error management, and executes recovery actions automatically in a timely manner. Hundreds of recovery actions are taken by CHIP during a typical physics run.

The second application, Shifter Assistant [10], is a tool developed to assist the ATLAS operators. It promptly notifies shifters about problems and failures, establishing their severity and providing pertinent information from different data sources. CHIP also recommends actions to perform for every issue detected, reducing the need of expert intervention when facing known or predictable issues.

3.2. Monitoring in the DAQ system

The DAQ software infrastructure provides a wide set of monitoring tools used by many of the ATLAS subsystems. Online application monitoring, error reporting, histogram publications or event data sampling are a few examples of the monitoring framework functionality.

In Run 2, a new tool was introduced for the persistent storage of the online monitoring data, known as P-BEAST: A Persistent Back-End for AtlaS Tdaq [11]. P-BEAST is a time-series database keeping an important fraction of operational data from ATLAS. It was developed using low-level primitives from *Google Protocol Buffers* [12], needed for data interoperability, compaction and compression.

Up to 500.000 attributes per second are stored during a data taking process and they are made accessible at a later time, programmatically or using a *Grafana*-based dashboard [13]. Most of the Run 2 operational data are currently stored in P-BEAST.

4. Conclusion and future plans

We have described the ATLAS DAQ system, its hardware components and key software elements, as they have been designed to operate under Run 2 LHC conditions. The experience accumulated during the Run 1 operation and the technology evolutions contributed to a successful

upgrade. The usage of the latest computing, networking and storage technologies was necessary to achieve the required design objectives and allowed increasing the system robustness.

In addition, the introduction of CEP engines in the system has helped keeping the system efficiency high and reducing the load on experts and operators of the experiment. New P-BEAST application stores persistently thousands of values of operational data allowing the experiment experts to access them later.

For the future, new challenges will arise as LHC increases luminosity. First in Run 3, and later in the so-called High-Luminosity LHC, the DAQ system will need to be upgraded to sustain new rates and bigger event sizes. The direction taken is to bring commercial technologies closer as possible to the detector, replacing today's custom electronics where possible. This reduces the effort needed to build future upgrades and benefit from the remarkable progress that industry is doing on the field of data center technologies.

References

[1] The ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider // Journal of Instrumentation 2008: vol. 3 p. S08003

[2] Borga A. et al. Evolution of the ReadOut System of the ATLAS experiment // Proceedings of Science 2014: TIPP2014 205.

[3] Abbott B. et al. The evolution of the Region of Interest builder for the ATLAS experiment at CERN // Journal of Instrumentation 2016: vol. 11 p. C02080

[4] Colombo T. and the ATLAS Collaboration. Data-Flow Performance Optimisation on Unreliable Networks: The ATLAS Data-Acquisition Case // Journal of Physics: Conference Series 2015: 608 012005.

[5] Le Goff F. and Vandelli W. Automated load balancing in the ATLAS high-performance storage software // Available at: https://cds.cern.ch/record/2270626/files/ATL-DAQ-PROC-2017-015.pdf (accessed 26.10.2017)

[6] Jereczek G. et al. Analogues between tuning TCP for data acquisition and datacenter networks // Proc. IEEE ICC, 2015, DOI: 10.1109/ICC.2015.7249288

[7] A. Phanishayee et al. Measurement and analysis of TCP throughput collapse in cluster-based storage systems // Proc. of the 6th USENIX Conference on File and Storage Technologies (FAST'08) 2008: pp. 12:1–12:14.

[8] EsperTech Inc. *Esper* // Available at: http://www.espertech.com/esper/ (accessed 26.10.2017)

[9] G Anders et al. Intelligent operations of the data acquisition system of the ATLAS experiment at LHC // Journal of Physics 2015: Conference Series 608 012007

[10] Santos A. et al. A Validation System for the Complex Event Processing Directives of the ATLAS Shifter Assistant Tool // Journal of Physics 2015: Conference Series 664 062055

[11] Sicoe A.D. et al. A persistent back-end for the ATLAS TDAQ online information service (P-BEAST) // Journal of Physics 2012: Conference Series 368 012002

[12] Grafana Labs. *Grafana* //Available at: https://grafana.com/ (accessed 26.10.2017)

[13] Google. *Protocol Buffers* // Available at: https://developers.google.com/protocol-buffers/ (accessed 26.10.2017)