Evolution of the ATLAS Trigger and Data Acquisition System

M E Pozo Astigarraga, on behalf of the ATLAS collaboration

CERN, CH-1211 Geneva 23, Switzerland

E-mail: mpozoast@cern.ch

Abstract. ATLAS is a Physics experiment that explores high-energy particle collisions at the Large Hadron Collider at CERN. It uses tens of millions of electronics channels to capture the outcome of the particle bunches crossing each other every 25 ns. Since reading out and storing the complete information is not feasible (~100 TB/s), ATLAS makes use of a complex and highly distributed Trigger and Data Acquisition (TDAQ) system, in charge of selecting only interesting data and transporting those to permanent mass storage (~1 GB/s) for later analysis. The data reduction is carried out in two stages: first, custom electronics performs an initial level of data rejection for each bunch crossing based on partial and localized information. Only data corresponding to collisions passing this stage of selection will be actually read-out from the on-detector electronics. Then, a large computer farm (~17 k cores) analyses these data in real-time and decides which ones are worth being stored for Physics analysis. A large network allows moving the data from ~2000 front-end buffers to the location where they are processed and from there to mass storage. The overall TDAQ system is embedded in a common software framework that allows controlling, configuring and monitoring the data taking process. The experience gained during the first period of data taking of the ATLAS experiment (Run I, 2010-2012) has inspired a number of ideas for improvement of the TDAQ system that are being put in place during the so-called Long Shutdown 1 of the Large Hadron Collider (LHC), in 2013/14. This paper summarizes the main changes that have been applied to the ATLAS TDAQ system and highlights the expected performance and functional improvements that will be available for the LHC Run II. Particular emphasis will be put on the evolution of the software-based data selection and of the flow of data in the system. The reasons for the modified architectural and technical choices will be explained, and details will be provided on the simulation and testing approach used to validate this system.

1. Introduction

ATLAS (A Toroidal LHC ApparatuS) is a multipurpose particle detector located in the Large Hadron Collider (LHC) at CERN [1]. When the first proton beams began to circulate in the LHC, by the end of 2009, the period known as Run 1 commenced. From that moment until the beginning of 2013, the ATLAS detector operated successfully recording proton-proton collisions at a center of mass energy of 8 TeV. In 2013, the Long Shutdown 1 period, or LS1, was scheduled in order to do the necessary maintenance and upgrade operations on the different systems of the apparatus. In particular, the Trigger and Data Acquisition (TDAQ) system of the ATLAS detector experienced important changes some of which have been described in [2] and [3].

The ATLAS TDAQ system is responsible for the readout, selection and sending to permanent storage of the physics events and plays a fundamental role in the ATLAS operation. This document presents the current status of the Data Flow, Read-Out System and Data Collection network during the

LS1 evolution. In the first quarter of 2015, a new period named Run 2 will begin in the LHC, with collisions expected at a center of mass energy of 13 TeV, inaugurating a new era for the HEP experiments.

2. The Data Flow upgrade

Protons bunches traveling nearly at the speed of light collide in the center of ATLAS several millions of times per second. In order to reduce the data volume generated in the ATLAS subdetectors multiple stages of event filtering are performed. The Level 1 (L1) Trigger performs a first rejection based on simple calorimetry and muon tracking information reducing the rate by more than a factor hundred. The events passing the L1 Trigger are read-out from the subdetectors' Front-End electronics and the data is stored in large memory buffers in the Read-Out System for further real-time analysis in the so called High Level Trigger (HLT) farm. A set of applications, libraries and communication protocols transporting and delivering the information from the Read-Out System (ROS) to the processing units in the HLT farm is known as the Data Flow.

During Run 1, two separated computer farms were in charge of the filtering of the events, performing the Level 2 (L2) and Level 3 (L3) trigger respectively. In the first farm, known as Data Collection farm, the triggering algorithms filtered the events at 75 KHz based on a partial reconstruction of the collision data. Once an event passed the L2 filter, a dedicated set of nodes, known as Sub-Farm-Input (SFI), made the full reconstruction of the events (1.5 MB) by pulling the fragments from the Read-Out System. In the second farm, the Event Filtering farm, the processing nodes filtered the events at 3 KHz rate executing the HLT algorithms on the full event pulled from the SFI nodes. Finally, the events retained were pushed to the Sub-Farm-Output (SFO) for temporary storage until the CERN Storage System could accept them for permanent storage (200 Hz).

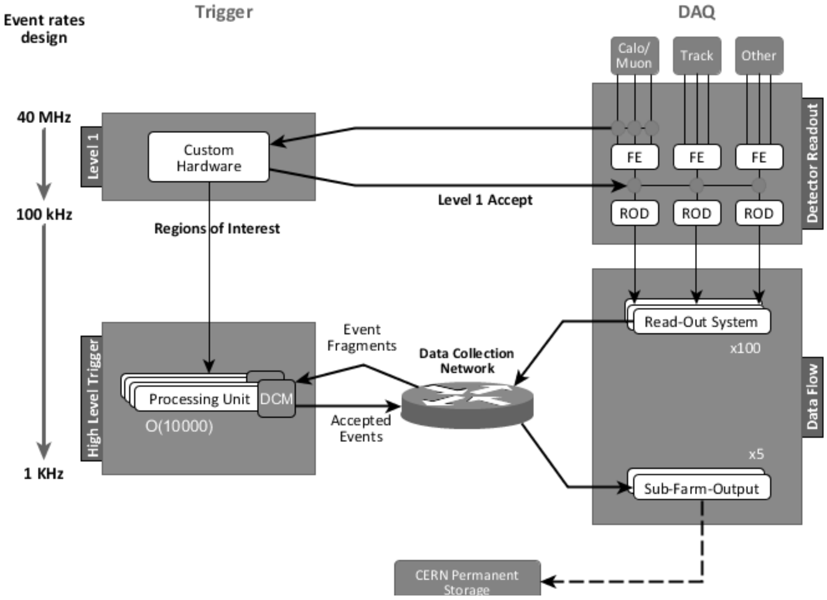

Figure 1. Atlas Trigger and DAQ with expected rates in Run 2

The Data Flow has been completely reviewed during LS1 in order to cope with the higher physics event rates and sizes expected in Run 2: detector read-out of ~2 MB events at 100 KHz. The HLT algorithms use often more data than originally foreseen for the Level 2 selection but very rarely need the full event to take the final decision. Driven by the HLT Supervisor nodes, a new Data Flow component running in each of the processing nodes, the Data Collection Manager (DCM), requests the event fragments progressively to the Read-Out-System as they are needed by the HLT algorithms. In the same way as for the original design, the events passing the HLT filter are sent to the Sub-Farm-Output where they are stored until they can be sent to the CERN permanent storage for Offline analysis.

This new architecture maximizes the flexibility of data access by the HLT algorithms, thus optimizing the selection efficiency; it profits from the advances in technology that allow the deployment of higher throughput and port density networks.

3. The Read-Out System upgrade

The Read-Out System receives and stores event fragments after the L1 accept. The fragments are requested by the processing nodes in the HLT farm for filtering and removed from the memory buffers only when a decision has been taken.

During Run 1, a hundred and fifty PCs ensured the correct behavior of the Read-Out-System. Each PC hosted from four to five PCI cards named ROBINs, and each ROBIN had three Read-Out-Links (ROLs) which connected the detector read-out to the ROS PCs. The event fragments stored in the ROBIN's buffers were requested by the processing nodes of the HLT farm thanks to the two 1 Gbps Ethernet links on a dual-port network interface card.

For Run 2, the design goal for the Read-Out System is 50% of read-out at 100 KHz. The increase in luminosity expected for Run 2 required a 20% increase in number of ROLs [4]. In addition, the limited memory capacity of the ROBINs and the obsolescence of the electronics made necessary a review and upgrade of the cards. For this purpose, a new PCIe card named RobinNP was developed with substantial improvements on the resources and performance characteristics. The connectivity of the ROS PCs to the HLT farm is provided by four 10 Gbps Ethernet ports on two dual-port network interface cards.

4. The Network architecture upgrade

The DAQ system makes use of two independent computer networks: the Data Collection Network and the Control Network.

4.1. The Data Collection Network

The Data Collection (DC) network connects the HLT processing nodes with the Read-Out System and the Sub-Farm-Output. The DC network has evolved to provide network connectivity as required by the new Data Flow architecture: the old Data Collection and Back-End networks have been merged into a single Ethernet network, the ROS PCs have been directly connected to the network cores and the Sub-Farm-Input removed (see Figure **2**. Data Collection network).

On the ROS side, the removal of the aggregation layer had important consequences on the traffic patterns. This change was motivated by the current state of the technology that allowed only the aggregation from 10 to 40 Gbps Ethernet. Because of the high cost of this type of network devices compared to the low aggregation factor they provide (4:1), it was decided that the aggregation layer could be removed. Each ROS PC has currently quadruple active-active bonded connections to the network routers and because the over-dimensioning of the link capacity it can support almost transparently the failure of any of these links.

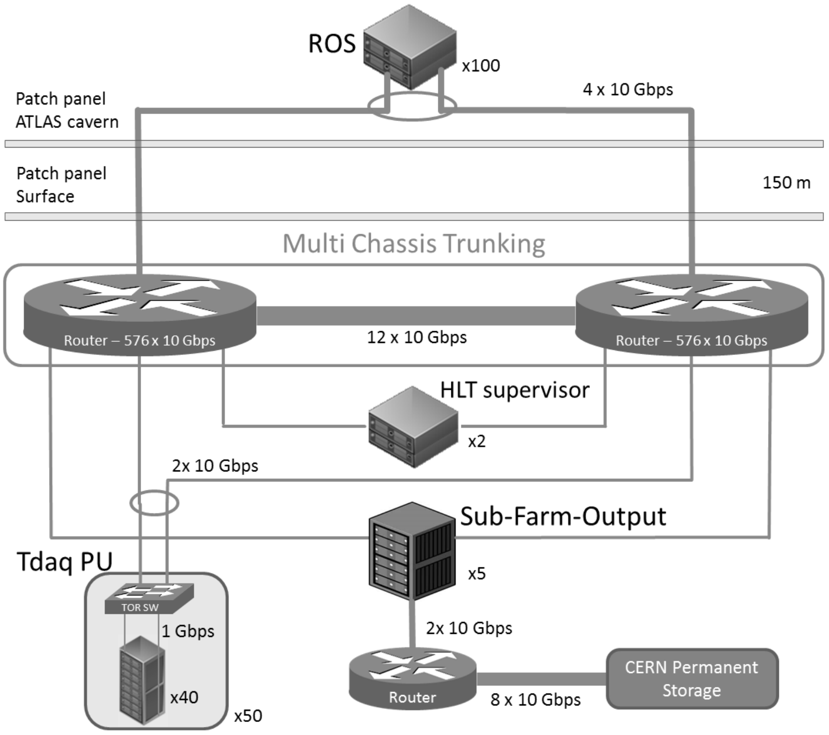

Figure 2. Data Collection network

The main change on the network side came from the replacement of the two network core routers that provided redundant connectivity to the different sub-systems. These devices were at the end of their lifetime and have been replaced by newer generation ones which are optimized to operate at 10 Gbps Ethernet. The internal architecture of the new routers incorporates a major improvement in the packet forwarding capacity by introducing the Virtual Output Queuing (VOQ) technique, which assigns different packet queues to each output port. Thanks to the VOQ technique, the packets addressed to a given output port can be forwarded independently of the traffic going to the other output ports, removing the head-of-line blocking phenomenon[5]. The consequence is that the device backplane has now a 100% packet forwarding efficiency.

The router manufacturer proposed several types of 10 Gbps Ethernet linecards for the routers with eight and twenty-four ports and different CAM memory sizes. The final purchase of the linecards, representing 60% of the total price, was delayed to the second half of the LS1 period in order to evaluate the various models. A simple software tool was developed to simulate the traffic throughput and patterns expected in the final system allowing the measurement of the switching capacity of the linecards and the backplane capacity of the routers. In the Figure 3. 8 port linecard (left) Vs 24 port linecard (right) maximum full-duplex traffic throughput we can find the comparison between the throughput for the eight and twenty-four port linecards when bidirectional TCP traffic is forwarded by the linecards' ASIC. The eight port linecard showed a performance of 93% of the theoretical values while the twenty-four port one performed at 87%. In order to measure the backplane capacity of the router, traffic was sent to the output ports in different linecards showing a 93% of efficiency for both models. Further tests with the ROS PCs were carried out using the Data Acquisition framework and modifying the traffic burstiness in the Data Flow software. This was possible by tuning the traffic shaping mechanism in the Data Collection Managers (DCM) that controls the number of event fragments that are consecutively requested to the Read-Out-System. All these tests proved that the linecard with twenty-four 10 Gbps Ethernet ports was the best cost-effective solution and that due to the unidirectional nature of the traffic, seven ports in a group of eight could be used safely in non-blocking mode, even if the specifications stated that only six should be used.

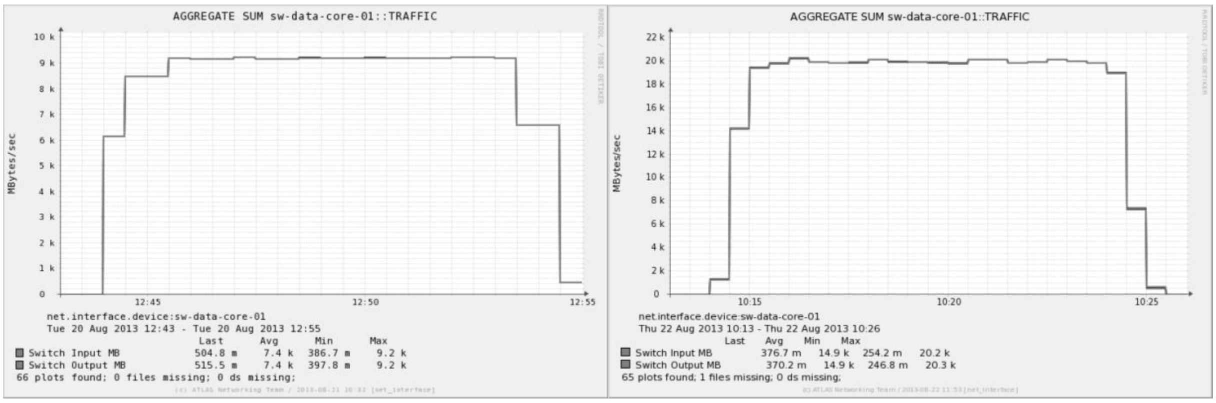

Figure 3. 8 port linecard (left) Vs 24 port linecard (right) maximum full-duplex traffic throughput

An additional enhancement of the new generation linecards is the deep packet buffers capable of absorbing few hundreds of milliseconds of packet bursts, covering most of the worst case scenarios. As a consequence, the network packets drops occur almost exclusively to the Top-Of-the-Rack switches in the processing nodes racks.

An interesting feature of the new network routers known as *Multi Chassis Trunking* provides both load balancing and link redundancy to the network (see Figure **2**. Data Collection network). This is achieved creating aggregated links on the devices connected to the routers which are perceived as a single virtual network device. A proprietary protocol running between the two routers ensures the Routing and Forwarding table synchronization between the two network cores needed for an optimal forwarding of the traffic.

Finally, several Top-Of-the-Rack switches for processing nodes racks were studied during LS1; this allowed us to clearly identify the required characteristics for future replacements. Many different candidate switches were evaluated taking into consideration the new traffic patterns and the traffic shaping policies that the DCMs can apply to the fragment requests. The results of the tests showed that without any traffic shaping, a switch with a buffer size bigger than a full event size per output port is needed to avoid packet drops (~2 MB per output port). However, applying an intelligent traffic shaping policy on the DCM, it was possible to obtain reasonably good results with much smaller shared buffers per device. Nowadays the data center switch market is moving into that direction and cheaper switches can be obtained. This subject is still under study and a decision about a suitable switch replacement will be taken in the following months.

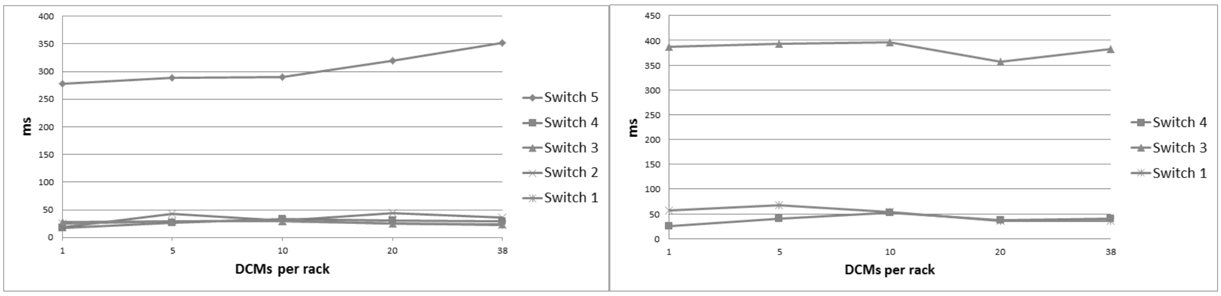

Figure 4. Event Building time (ms) with traffic shaping (left) and without (right)

4.2. The Control Network

The Control Network is a parallel network to the DC network in which the traffic for the control, configuration and monitoring of the TDAQ infrastructure flows. There are not strong performance requirements for the control network so the main upgrade activities concerned the redundancy and fail-over techniques. As part of the redundancy improvements, an Active-Backup setup has been installed for all important nodes in the system: Online and Monitoring PCs, ROS, HLT supervisor and SFOs. Several fail-over tests have been carried out to validate the technical choices, trying to anticipate any potential issue.

5. Conclusions

The right functioning of the TDAQ system has a direct impact on the operation of the ATLAS experiment and the achievement of its Physics goals. In this paper we have shown how the TDAQ architecture has been reshaped during LS1 period in order to profit from the technological progress and to maximize the flexibility and efficiency of the data selection process.

The preliminary tests carried out so far have demonstrated the correct behaviour of the deployed system and validated our technical choices. The overall TDAQ has nevertheless not yet been completely installed, thus more measurements and tests will be needed before the start of data taking with the LHC in 2015.

References

[1] ATLAS Collaboration. 2008. *Journal of Instrumentation* 3 S08003

[2] A. Krasznahorkay et al. 2013. The evolution of the Trigger and Data Acquisition System in the ATLAS experiment. ATL-DAQ-PROC-2013-018, https://cds.cern.ch/record/1604503

[3] N. Garelli et al. 2013. The Evolution of the Trigger and Data Acquisition System in the ATLAS Experiment. ATL-DAQ-PROC-2013-029, https://cds.cern.ch/record/1622797

[4] W. Vandelli et al. 2014. Evolution of the ReadOut System of the ATLAS experiment ATL-DAQ-PROC-2014-012 https://cds.cern.ch/record/1710776

[5] N. McKeown, A. Mekkittikul; V. Anantharam, J.Walrand. 1999. "Achieving 100% Throughput in an Input-Queued Switch". *IEEE Transactions on Communications* 47 (8): 1260–1267