PROCEEDINGS OF SCIENCE

The ATLAS Dataflow System in Run 2: Design and Performance

Othmane Rifki* **on behalf of the ATLAS Collaboration**
University of Oklahoma
E-mail: othmane.rifki@cern.ch

The ATLAS Dataflow system is composed of distributed hardware and software responsible for buffering and transporting event data from the Readout system to the High Level Trigger and to the event storage. By building on the experience gained during the successful first run of the LHC, the ATLAS Data Acquisition (DAQ) architecture has been simplified and upgraded to take advantage of state of the art technologies resulting in a maximized efficiency and improved performance. This proceeding describes the new architecture of the ATLAS DAQ system and highlights its performance during Run 2 of the LHC.

38th International Conference on High Energy Physics
3-10 August 2016
Chicago, USA

*Speaker.

PoS(ICHEP2016)240

1. Introduction

The ATLAS detector [1] is a multipurpose particle detector at the Large Hadron Collider (LHC) at CERN, Switzerland. After a 2-year shutdown for maintenance and upgrade, the LHC resumed operations starting Run 2 of the LHC in 2015. The ATLAS trigger and data acquisition (TDAQ) system was upgraded to simplify its architecture and increase its flexbility due to the increased energy and instantaneous luminosity (rate of proton collisions), and the addition of new detector systems. In Run 2 the recorded particle interactions, i.e. events, have a larger size and need to be processed at higher rates which required an upgrade of the dataflow component of the ATLAS TDAQ system. This system has been re-shaped in order to maximize the flexibility and efficiency of the data selection process leading to a different architecture of the ATLAS dataflow. In this proceeding, the Run 2 challenges motivating the upgrade will be covered along with a description of the new dataflow architecture and its performance.

2. Run 2 Challenges

The ATLAS TDAQ system reduces the proton interaction rate from 40 MHz to the ATLAS data storage capacity of about 1.5 kHz. A hardware First Level Trigger (L1) reduces the rate to 100 kHz and a software High Level Trigger (HLT) selects events for offline analysis. The function of the DAQ system is to efficiently buffer, transport, and record the events that were selected by the trigger system. Its peformance is affected by the instanteneous luminosity that leads to busy events with multiple proton-proton interactions occuring in each bunch crossing, referred to as pileup. The high pileup results in a higher data volume collected by the detector that needs to be processed at the required rate to avoid exerting back-pressure on the L1 system. In Run 2, the LHC has exceeded the designed instantaneous luminosity of 10^{34} cm^{-2} s^{-1} leading to pileup of $<\mu>=30$ or more as shown in Figure 1. The L1 accept rate has also increased from 75 kHz in Run 1 to 100 kHz in Run 2 and the average output rate of the data logger system has increased from 400-600 Hz in Run 1 to about 3 kHz with 1.5 kHz for physics data. Moreover, there were new detectors that were added in Run 2 (Insertable B-layer (IBL), L1 topological trigger, Fast Tracker (FTK))[3] leading to an increase of 20% in the number of readout channels. To be able to deliver more rate to the High Level Trigger (HLT), the upgrade also targeted the Readout System (ROS)[4]. For the same reason the two level of the HLT system were collapsed into a single level which made the system more flexible allowing for incremental data retrieval and analysis. The dataflow network system was re-designed to increase its capacity and simplify its architecture[5].

3. ATLAS Dataflow Design

In Run 1, the farm was subdivided to several slices, with each slice managed by a dedicated supervisor. This layout has been dropped in favor of global management by a single farm master operating at 100 kHz referred to as the HLT supervisor (HLTSV). The Region of Interest Builder (RoIB) that assembles the RoIs previously implemented on a VMEbus system is now integrated with the HLTSV and the RoI building done in software. The change in the HLT architecture from two to one level required re-writing the HLT software and algorithms in such a way that

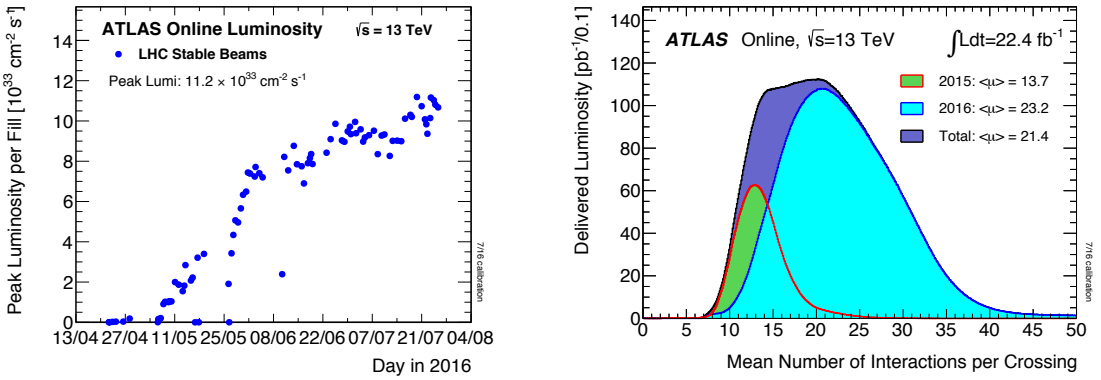

Figure 1: Run conditions during Run 2: ATLAS online luminosity (left), ATLAS online pileup (right) [2].

each node in the farm can perform all processing steps. The handling of these processing steps is done by a single Data Collection Manager (DCM) process running on each HLT node to manage the L1 RoIs, the dataflow between the ROS and the HLT processing units (HLTPU), the event building processes, and the data logging. In the new architecture, the computing resources are managed more efficiently by balancing the utilization of all cluster nodes depending on the active HLT algorithms and by sharing the HLT code and services to reduce memory and resource usage.

The dataflow network was simplified and upgraded to handle a larger data volume. A single network is used for RoI based access from the ROS, event building in the HLT processing nodes, and sending data for logging. A 10 GbE connectivity has been adopted throughout the dataflow system resulting in a factor of four increase in bandwidth between the data loggers and the permanent storage, and a 4×10 GbE output from each ROS PC to the core routers. The HLTSV and the HLT racks are all connected directly to each of the two core routers via 2×10 GbE connection. Each HLT rack is hosting up to 40 nodes connected by 2×1 GbE to the top-rack switches. The capacity of the routers can accommodate an increase in the number of HLT server racks and ROS PCs by a factor of two, which will be needed when the system scales as run conditions change. The core routers also provide load balancing and traffic shaping protocols [5] to distribute the data throughout the system more evenly. A duplication of core routers provide link redundancy at every level in case of link or switch failures.

To take advantage of multi-core architectures, the dataflow software is using multi-threaded software design for CPU consuming operations. The Input/Output of the dataflow is based on asynchronous communication using industry standard libraries such as the Boost::ASIO library. All the ATLAS software suite was switched to exclusively 64 bit operation in 2016.

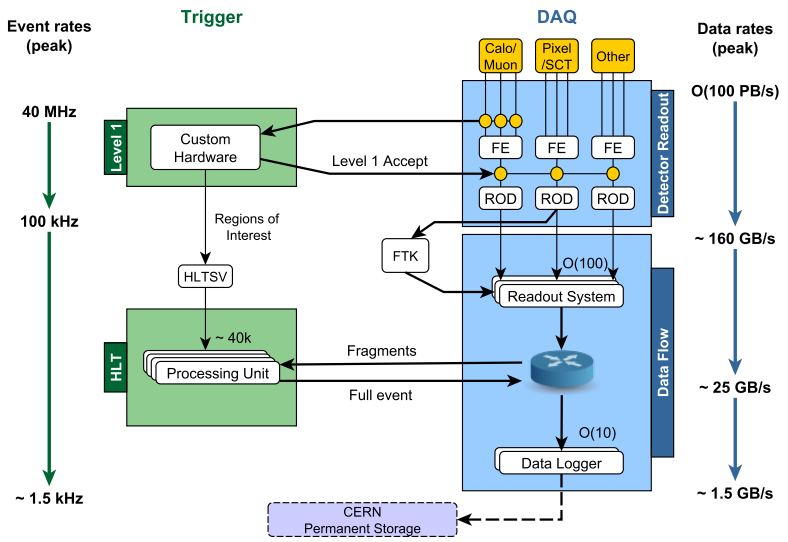

Figure 2: ATLAS TDAQ architecture.

4. Region of Interest Builder

The first step of the HLT processing is to run on the RoIs found by the L1 hardware trigger. These RoIs are collected and distributed to the HLT farm by the RoIB [6] which was the latest change to the ATLAS dataflow. The evolution of the RoIB system from a crate of custom VME-based electronics (VME-RoIB) to a commodity PC hosting a custom PCI-Express card (PC-RoIB) has been undertaken to increase the system performance, flexibility, and ease of maintenance. The functionality of the VME-RoIB previously possible only in FPGAs has now been implemented in a multi-threaded C++ software library. For each proton-proton collision that is accepted by the L1 trigger, the RoIB receives an RoI record from the custom inputs via S-Link. The RoIB assembles these records into a single record which is then forwarded to the HLTSV. The HLTSV then distributes these single records to the HLT farm. The RoIB is also responsible for monitoring the data integrity of the incoming fragments and diagnostic performance of the system.

As shown in Figure 3, the performance of the PC-RoIB with realistic running ATLAS conditions is improved over the VME-RoIB particularly at high RoI sizes.

Figure 4 shows that the memory usage of the HLTSV is at the level of 5% and that the RoIB event assembly does not depend on pileup conditions.

5. Performance in Run 2

The reliable operation of the TDAQ system directly impacts the efficiency of the ATLAS experiment in recording the collisions delivered by the LHC. As a result, high data-taking efficiency is crucial for the ATLAS physics program. The ATLAS recorded efficiency in 2016 is over 90%, as shown in Figure 5 with a negligible fraction of data loss due to the DAQ system. The new dataflow architecture is scaling well with the increased instantaneous luminosity during 2016 data-taking

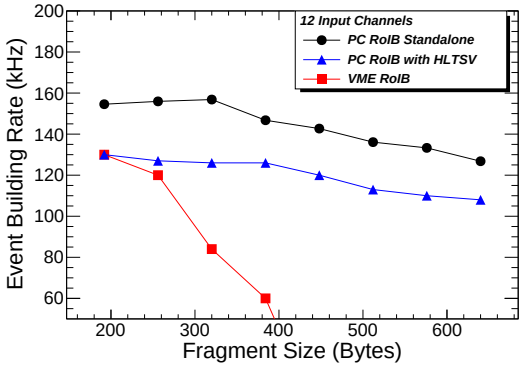

Figure 3: The event building rate as a function of the RoI record size in Bytes. The rates are shown for a standalone application that implements a minimal interface for event building, the integrated RoIB software into an HLTSV process running within the full ATLAS TDAQ software suite, and for comparison the VME-RoIB performance.

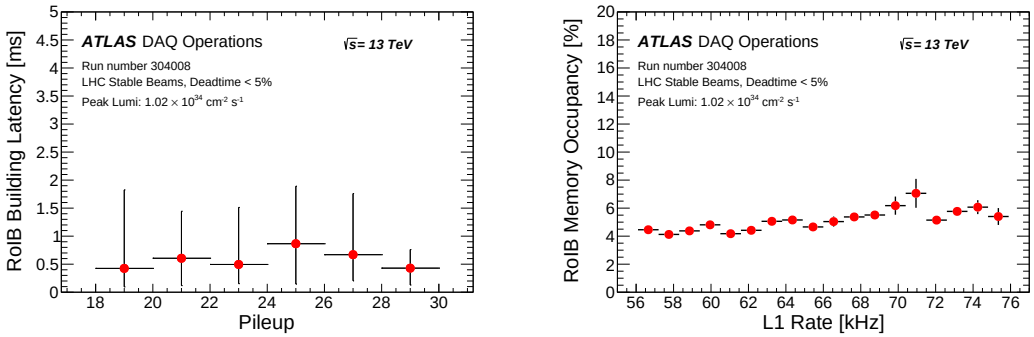

Figure 4: RoIB performance: RoIB building latency as a function of pileup (left), RoIB memory occupancy as a function of L1 rate (right).

and is capable of handling larger pileup and thus larger event sizes. For illustration, Figure 6 shows the evolution of the average processing time per event and the event size where there is relatively mild increase as a function of pileup which will within the system capacity.

6. Conclusion

The dataflow system of ATLAS was re-shaped for Run 2 in order to handle the more demanding run conditions expected throughout the run. The new redesign profitted from the technological progress that took place in the last few years. As a result, the new system is considerably simplified, more performant, and scalable. Moreover, there is more headroom in performance to cope with more challenging run conditions of the LHC to ensure that ATLAS DAQ continues delivering physics data with high efficiency.

7. Acknowledgement

We would like to thank the U.S. Department of Energy grant DE-SC0009956 for support of

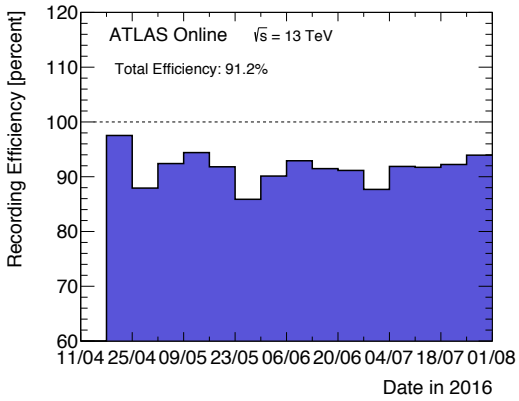

Figure 5: ATLAS recorded efficiency [2].

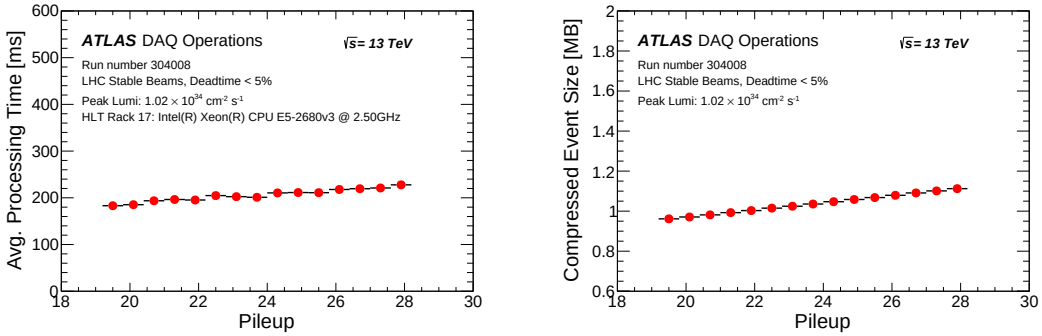

Figure 6: Performance in Run 2: Average processing time as a function of pileup (left), compressed event size as a function of pileup (right).

the speaker in the work presented in this proceeding.

References

[1] ATLAS Collaboration. *Journal of Instrumentation*, 3(08):S08003, 2008.

[2] Luminosity public results in run 2. `https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2`. Accessed: 2016-08-01.

[3] ATLAS Collaboration. Technical Design Report for the Phase-I Upgrade of the ATLAS TDAQ System. Technical Report CERN-LHCC-2013-018, Sep 2013.

[4] W. P. Vazquez et al. The atlas data acquisition system: from run 1 to run 2. *Nuclear and Particle Physics Proceedings*, 273 - 275:939 – 944, 2016.

[5] A Negri et al. Evolution of the trigger and data acquisition system for the atlas experiment. *Journal of Physics: Conference Series*, 396(1):012033, 2012.

[6] B. Abbott, R. Blair, G. Crone, B. Green, J. Love, J. Proudfoot, O. Rifki, W.P. Vazquez, W. Vandelli, and J. Zhang. The evolution of the region of interest builder for the atlas experiment at cern. *Journal of Instrumentation*, 11(02):C02080, 2016.