Evolution of the Trigger and Data Acquisition System for the ATLAS experiment

A Negri\textsuperscript{1} on behalf of the ATLAS TDAQ collaboration\textsuperscript{[1]}
\textsuperscript{1}University of Pavia and INFN
E-mail: Andrea.Negri@cern.ch

Abstract. The ATLAS experiment at the Large Hadron Collider at CERN relies on a complex and highly distributed Trigger and Data Acquisition (TDAQ) system to gather and select particle collision data at unprecedented energy and rates. The TDAQ is composed of three levels which reduces the event rate from the design bunch-crossing rate of 40 MHz to an average event recording rate of about 200 Hz. The first part of this paper gives an overview of the operational performance of the DAQ system during 2011 and the first months of data taking in 2012. It describes how the flexibility inherent in the design of the system has been exploited to meet the changing needs of ATLAS data taking and in some cases push performance beyond the original design performance specification. The experience accumulated in the TDAQ system operation during these years stimulated also interest to explore possible evolutions, despite the success of the current design. One attractive direction is to merge three systems - the second trigger level (L2), the Event Builder (EB), and the Event Filter (EF) - within a single homogeneous one in which each processing node executes all the steps required by the trigger and data acquisition process. Appealing aspects of this design are: a simplification of the software architecture and of its configuration, a better exploitation of the computing resources, the caching of fragments already collected for L2 processing, the automated load balancing between L2 and EF selection steps, the sharing of code and services on HLT nodes. Furthermore, the full treatment of the HLT selection on a single node allows more flexible approaches, for example "incremental event building" in which trigger algorithms progressively enlarge the size of the analyzed region of interest, before requiring the building of the complete event. To spot possible limitations of the new approach and to demonstrate the benefits outlined above, a prototype has been implemented. The preliminary measurements are positive and further tests are scheduled for the next months.

1. Introduction

The Large Hadron Collider \cite{2} was designed to provide proton–proton collisions at the center-of-mass energy of 14 TeV with a luminosity of $10^{34}$ cm$^{-2}$s$^{-1}$. In 2011 LHC provided 31 weeks of proton–proton collisions at $\sqrt{s} = 7$ TeV and one month of ion–ion interactions. In the proton run the luminosity steadily increased during the year reaching a peak value of $3.42 \times 10^{33}$ cm$^{-2}$s$^{-1}$. This year LHC is operating at $\sqrt{s} = 8$ TeV and the expected peak luminosity is $6.68 \times 10^{33}$ cm$^{-2}$s$^{-1}$. At the time of this conference (May 2012), LHC already reached about 83% of this value. Since this luminosity is delivered with a bunch separation twice the nominal value (25 ns), the average number of interactions per bunch-crossing ($\mu$) is doubled. The fast evolution of the collider performance and the high average pileup present a challenge for the Trigger and Data Acquisition system (TDAQ) of the ATLAS detector \cite{3}. 

Published under licence by IOP Publishing Ltd
2. The ATLAS TDAQ system

The ATLAS TDAQ architecture [4] is outlined in Fig. 1. It is organized in three levels of online event selection. Starting from the nominal bunch crossing rate of 40 MHz, the rate of selected events is reduced to an average rate of a few hundred Hertz for permanent storage. The first level (L1) is built from dedicated hardware and makes use of coarse information coming from muon and calorimeter systems. The L1 trigger has a maximum accept rate of 75 kHz and a latency of 2.5 $\mu$s. The High-Level Triggers (HLT) are distributed software systems executed on commodity computers and have access to the detector data at full granularity. The Level 2 (L2) accesses only a few percent of the available data in the so-called regions-of-interest (RoI). They are identified by the L1 and assembled by a dedicated hardware component, the RoI builder [5]. The L2 has a design output rate of 3 kHz and the average processing time is about 40 ms. The Event Filter (EF) operates on the complete events built in the Event Builder system (EB), which is composed of a farm of about 50 nodes. The EF has a design output rate of $\sim$200 Hz and the average processing time is about 4 s. In both L2 and EF, the HLT selection is managed by the HLT Steering [6], that schedules the HLT algorithms corresponding to the trigger menu prescriptions.

On L1 accept, data are sent from the front end electronics to detector-specific modules (ROD), which build event fragments and push them via $\sim$1600 optical links [7] to dedicated memories [8]. The latter are in charge of storing the data during L2 and EB latencies and are hosted on PCI boards mounted on $\sim$150 PCs (ROS). The backbone of the TDAQ system is composed of two Gigabit Ethernet networks: the data collection (DC) and the back end (BE). The DC network allows the L2 and EB systems to collect the data from the ROEs, while the BE network is used to move built events from the EB to the EF and accepted events from EF to the data logger system.
3. Operational performance 2011

In 2011 the ATLAS TDAQ coped extremely well with the evolving requirements. It recorded 5.25 \( fb^{-1} \) of data with the remarkable overall data taking efficiency of 94.9%. Since LHC operated with a bunch separation of 50 ns, the L1 operated at an input rate of 20 MHz. The maximum output rate was limited to \( \sim 65 \) kHz to prevent excessive dead time. As shown in Fig. 1, the EB and Data logger systems operated well beyond the design. Therefore the disk size of the five Data Logger nodes has been expanded to keep the capacity of locally buffering two days of data.

During the 2011 the HLT computing resources were incremented by \( \sim 50\% \) following the evolution of the instantaneous LHC luminosity (Fig. 2). The HLT computers are hosted in racks containing 30 or 40 computing nodes (i.e.: motherboards with two CPU sockets). There are two types of racks: the EF racks are connected to BE network only, while the XPU ones (eXchangeable Processing Units) see both data networks and can be configured, on a run by run basis, as either L2 or EF resources. Since the installation of worker nodes and relative networking equipment was staged over the course of the last three years, the farm is not homogeneous. Furthermore the network infrastructure is heterogeneous since the older XPU racks have only 2 Gb/s links to the BE network, while the other racks have a 10 Gb/s link. This heterogeneity led to load balance problems. For example since the EB and EF systems were configured in slices, a big effort was needed to fine tune the allocation of resources amongst them. The issue has been addressed by moving to a flat structure of the EB-EF system: the slice concept is removed and each node of the EF farm can connect to any application in the EB farm. However, in order to reduce the number of sockets to be managed, the number of connections by node is limited via a configuration parameter. The connections are allocated using a fair share algorithm which assures that each EB application serves the same number of EF nodes.

Another critical issue related to the heterogeneity of the system is the allocation of processing resources between L2 and EF. The sharing of XPU racks depends not only on the trigger menu, but has also constraints from EB network bandwidth and processing power of each rack. These constraints are included in a model that correlates the main parameters of the DAQ/HLT system. Its output is shown in Fig. 3. The colored lines in the plot represent the maximum allowed L2
Figure 4. Average L2 (up) and EF (down) processing times as function of the pileup measured in 2012. The two series refer to the two different kinds of XPU hardware.

Figure 5. Sub-detector fragment size as function of the pileup. The points below $\langle \mu \rangle = 12$ were taken in 2011, while the others were measured this year.

rate (indicated in the lines labels) as a function of the L2 and EF average latencies. The lines take the form of a staircase function because the nodes are allocated with a granularity defined by the physical installation: from right to left each step represents a group of nodes being moved from the L2 to the EF. The varying sizes of the steps reflect the heterogeneous hardware: with three different hardware configurations being represented in the three vertical sections of the plot. The yellow area marks the zone exceeding the maximum EB bandwidth. The blue area marks the zone where the L2 processing capacity is saturated. In both cases the system would not be able to cope with the L1 rate, thus leading to dead-time. A tool for managing the HLT resources has been developed and it is described in [9].

During the 2011 the full EF farm and the motherboards of half of the ROS PCs were subject to the rolling replacement policy. The change entailed the major operational problem of the year: frequent network card failures in the upgraded ROS nodes. Despite the concentrated efforts to understand and solve the problem, its source is still unknown. A workaround was found via the installation of different network cards that do not interfere with the ROS motherboard.

4. Operational performance 2012

During the winter shutdown an additional core router was installed to provide redundancy to the BE network and sixteen of the XPU racks were replaced under the rolling replacement scheme. The current composition of the HLT farm is listed in Table 1.

The functionality was further extended via the implementation of new functionalities in the data collection area. For example, the data collection software has been updated to allow $E_{\text{miss}}^T$ trigger selection at L2, a functionality that was missing in the original system. Indeed, since the L2 was designed to access only a subset of the detector data, the vector and scalar calorimeter sums are not computable at this level. But summary information was available in the front-end electronics of all calorimeters (including the sums over the $x$, $y$ components of the transverse energy and the scalar sum over all connected channels to each front-end board). So the ROS
Table 1. HLT processing farm composition at May 2012.

| Type | Racks | Nodes | Model                  | Cores | Memory   | DC link | BE link |
|------|-------|-------|------------------------|-------|----------|---------|---------|
| EF   | 14    | 448   | E5540@2.53GHz          | 8     | 16 GB    | 10 Gb/s |         |
| XPU  | 11    | 341   | E5420@2.50GHz          | 8     | 16 GB    | 10 Gb/s | 2 Gb/s  |
| XPU  | 25    | 904   | X5650@2.67GHz          | 12    | 24 GB    | 10 Gb/s | 10 Gb/s |

Figure 6. Schematic representation of the DAQ evolution design. The numbers represent the multiplicity of each component, \( n \) is the number of available cores in the node (at the moment 8 or 12).

Figure 7. CPU used by L2 and EF algorithm and by the DCM application varying runtime the L2 trigger acceptances and processing times. The node has 8 cores, so the maximum CPU usage is 800%.

software was modified to allow the gathering of this information from the front-end electronics of all calorimeters and make it available to the L2 that can now apply \( E_T^{miss} \) trigger selection.

Furthermore a lot of effort was dedicated to consolidate monitoring tools, to predict possible bottlenecks and to improve automatic recovery procedures as described in [10].

The major issue for 2012 is the pileup. As shown in Fig. 4 and 5, both the HLT processing time and the event size show a pileup dependency. The 2012 measurements show that they scale linearly up to \( \langle \mu \rangle \sim 23 \). These trends have been extrapolated up to the maximum expected pileup of \( \langle \mu \rangle \sim 35 \). At the maximum luminosity we expect a 25% CPU margin shared across L2 and EF. But, given the extrapolation uncertainty, the CPU usage evolution is being monitored. As regards to the event size, it is expected to increase up to about 1.8 MB, a value that will leave limited operational margins at peak luminosity. Therefore spare nodes have been put in production to increase the EB and data logger performance. The high pileup also affects the ROS performance, not only through fragment size but also via the potentially higher number of RoI per bunch. The updated ROS do not suffer anymore from CPU saturation, but some PCs are facing access rate and bandwidth limitations, that need to be monitored in the following months.

5. Data flow evolution
The 2013–2014 shutdown is the first occasion for a major update of the data flow software. As shown above, the past years of operation of the TDAQ system has confirmed the success of the current design but also stimulated interest to explore possible evolutions. Building on the
operational experience gained during 2010–2012, it is planned to introduce further scalability and flexibility into the design of the data flow system. It was decided to evolve the data flow infrastructure toward a more homogeneous architecture, in which the L2, EB and EF components are merged in a single system. Every node of the HLT farm executes all the steps required by the trigger selection process. Each L1 event is assigned to an available HLT node which executes the L2 algorithms using a subset of the event data and, upon positive selection, builds the event and processes the EF algorithms. This solution preserves the RoI concept while providing a simplified system and a more efficient use of network and processing resources. For example via the automated load balancing between L2 and EF selection steps and the caching of fragments already collected for L2 processing. Furthermore, the full treatment of the HLT selection on a single node allows more flexible approaches, which are under evaluation. For example “incremental event building” in which trigger algorithms progressively enlarge the size of the analyzed region of interest, before requiring the building of the complete event.

The new architecture is shown in Fig. 6. The RoI builder sends the L1 results to one of the HLT Supervisors (HLTSV), which in turn assigns the event to one of the available HLT nodes. The HLT nodes run several HLT Processing Units (HLTPUs) and a Data Collection Manager (DCM), which is in charge of collecting and locally storing the event fragments. The HLTPU requests from the DCM first only the fragments that correspond to the Region of Interest (RoI) passed by the L1 system, executes the L2 trigger algorithms and in doing so can request data beyond the RoI, and decides to discard the event or to ask the DCM to fully build it. At this stage, full event reconstruction algorithms can be executed. If the event is accepted for storage, the HLT result is transferred to the DCM, which packs all information together and ships the full event to one of the Data Logger.

In order to test ideas and better understand some technical issues, a prototype was developed and tested on the ATLAS HLT farm. It was developed in 2011 adapting current DAQ applications to the duties of the new components. Since the HLT steering part of the prototype is under development, HLT processing is emulated by a dummy implementation, which provides realistic CPU burning algorithms.

The Fig. 7 shows the node CPU usage in a run in which the L2 trigger acceptance and processing time were changed at runtime. At any configuration change the L2 and EF relative weights vary, but the CPU is steadily fully exploited. The plot demonstrates how the system is always capable of sharing the CPU resources between the L2 and EF algorithms for a wide
range of operating conditions.

The Fig. 8 shows the maximum HLTSV rate as a function of the number of HLTPUs (equal to the number of computing cores) for running condition close to the nominal values of the current system. The system scales linearly up to the number of CPU cores used in these tests (more than 12000 on about 1100 nodes). Furthermore the measured rates are consistent with the ones calculated from the configured burning time and therefore there is no contribution from network latencies. Indeed excessive queueing in the network switches are prevented via a traffic shaping strategy similar to the one used by the current EB system. As shown in Fig. 9 the best performance can be obtained limiting in each DCM the number of outstanding requests to less than about 30.

6. Conclusions

The ATLAS TDAQ system coped well with the 2011 data taking providing a very high efficiency of 94.9%. The flexibility inherent in the design of the system has been exploited to meet the changing needs of ATLAS data taking and in some cases push performance beyond the original design performance specification. During the 2011 run and the winter shutdown part of the system was subject to the rolling replacement scheme. The TDAQ system maintained the high operation efficiency also in the first months of 2012. The high pileup has implications of the systems which are being addressed and will be surveyed during the year.

Despite the success of the current design, during the 2013–2014 shutdown the system will evolve toward a more homogeneous architecture. To spot possible limitations of the new approach a prototype has been implemented. It demonstrated that the merging of L2, EB and EF components into a single system is a realistic possibility. The performance scales as expected, no limitations were observed so far and the appealing aspects of the new design were confirmed. Therefore the design phase of the final system has started and first implementation is scheduled for the end of the year.

References

[1] The ATLAS Collaboration, The ATLAS Trigger/DAQ Authorlist, version 11.09, ATL-DAQ-PUB-2011-002, CERN, Geneva, (2011), https://cdsweb.cern.ch/record/1386334.
[2] L. Evans and Ph. Bryant, LHC Machine, JINST 3 S08001, (2008).
[3] The ATLAS Collaboration, The ATLAS experiment at the CERN Large Hadron Collider, JINST 3 S08003, (2008).
[4] The ATLAS collaboration, ATLAS high-level trigger, data acquisition and controls technical design report, CERN-LHCC-2003-022, http://cdsweb.cern.ch/record/616089, (2003).
[5] R. Blair et al., The ATLAS high level trigger region of interest builder, JINST 3 P04001, (2008).
[6] S. Armstrong et al., Studies for a Common Selection Software Environment in ATLAS: From Level-2 Trigger to the Offline Reconstruction, IEEE Trans. Nucl. Sci. 51 no. 3, 915920, (2004)
[7] H.C. van der Bij et al., S-LINK, a data link interface specification for the LHC era, IEEE Trans. Nucl. Sci. 44 398, (1997)
[8] R. Cranfield et al., The ATLAS ROBIN, JINST 3 T01002, (2008).
[9] N. Garelli et al., Balancing the resources of the High Level Trigger farm of the ATLAS experiment, in the proceedings of this conference, (2012).
[10] G. Avolio et al., Applications of advanced data analysis and expert system technologies in the ATLAS Trigger-DAQ Controls framework, in the proceedings of this conference, (2012).