The Evolution of the Trigger and Data Acquisition System in the ATLAS Experiment

N Garelli¹, on behalf of the ATLAS Collaboration

¹ SLAC National Accelerator Laboratory, 2575 Sand Hill Rd, Menlo Park, CA 94025, United States
E-mail: nicoletta.garelli@cern.ch

Abstract. The ATLAS experiment, which records the results of LHC proton-proton collisions, is upgrading its Trigger and Data Acquisition (TDAQ) system during the current LHC first long shutdown. The purpose of this upgrade is to add robustness and flexibility to the selection and the conveyance of the physics data, simplify the maintenance of the infrastructure, exploit new technologies and, overall, make ATLAS data-taking capable of dealing with increasing event rates.

While the TDAQ system successfully operated well beyond the original design goals, the accumulated experience stimulated interest to explore possible evolutions. With higher luminosities, the required number and complexity of Level-1 triggers will increase in order to satisfy the physics goals of ATLAS, while keeping the total Level-1 rates at or below 100 kHz. The Central Trigger Processor will be upgraded to increase the number of manageable inputs and accommodate additional hardware for improved performance, and a new Topological Processor will be included. A single homogeneous high level trigger system will be deployed. The current second and third trigger levels will be executed together on a unique hardware node. This design has many advantages: the radical simplification of the architecture, the flexible and automatically balanced distribution of the computing resources, the sharing of code and services on nodes.

In this paper, we report on the design and the development status of the upgraded TDAQ system, with particular attention to the tests currently on-going to identify the required performance and to spot its possible limitations.

1. Introduction

The Large Hadron Collider (LHC) at CERN produced proton-proton collisions with increased performance between 2010 and 2012, reaching a center of mass energy of 8 TeV and instantaneous luminosity of $7.73 \cdot 10^{33} \text{cm}^{-2}\text{s}^{-1}$ during the Run 1.

ATLAS (A Toroidal LHC ApparatuS) [1], one of the two general purpose experiments at the LHC, successfully recorded more than 21 fb$^{-1}$ of data, corresponding to more than 93% of the total delivered luminosity. The trigger and data-acquisition system of ATLAS (referred to as TDAQ) coped extremely well with the changing requirements, operating beyond the design. The amount and quality of the recorded data allowed the achievement of extraordinary results, such as the discovery of a Higgs Boson [2].

On February 2013 the Long Shutdown 1 (LS1) phase started, and it will last for about two years. The accelerator complex will be upgraded to reach a center of mass energy of 13 TeV and instantaneous luminosity of $2 \cdot 10^{34} \text{cm}^{-2}\text{s}^{-1}$ during Run 2. Table 1 shows some major LHC
parameters: the original design ones, the running conditions in 2012, and three out of the many possible scenarios for Run 2.

Table 1. Some operation parameters of the LHC: the center of mass energy ($E_{CM}$), the number of bunches per beam (K), the instantaneous luminosity and the number of additional interactions (pile-up). For 2012 the maximum instantaneous luminosity reached in ATLAS is shown.

|                | $E_{CM}$ (TeV) | K   | Luminosity ($cm^{-2}s^{-1}$) | Pile-Up |
|----------------|----------------|-----|-------------------------------|---------|
| Nominal (25 ns)| 14             | 2808| $10^{34}$                      | 23      |
| 2012 (50 ns)   | 8              | 1380| $7.73 \cdot 10^{33}$          | 35      |
| Run 2 (25 ns)  | 13             | 2760| $0.85 \cdot 10^{34}$          | 23      |
| Run 2 (25 ns) low emittance | 13             | 2600| $1.5 \cdot 10^{34}$          | 42      |
| Run 2 (50 ns) low emittance | 13             | 1260| $2.0 \cdot 10^{34}$          | 110     |

2. TDAQ Design
The TDAQ system has been designed to select interesting events from the nominal LHC collision rate of 40 MHz down to 200 Hz of permanently stored data [3].

The trigger system is composed of three levels. The Level-1, based on custom hardware, reduces the rate down to 75 kHz with decision latency of less than 2.5 $\mu$s. The two subsequent software-based trigger levels, collectively called High Level Trigger (HLT), are the Level-2 and the Event Filter (EF).

The Level-1 is composed of the calorimeter trigger, the muon trigger and the Central Trigger Processor (CTP), which serves the Level-1 results to the detectors. The Level-1 systems provide co-ordinates of centers of areas in the $\eta\phi$ space where the Level-1 selected events occurred, referred to as Regions of Interest (RoI). The Region of Interest Builder (RoIB) collects and merges this information for each event, and forward it to the Level-2 farm. The Level-2 trigger algorithms operate on data from within these regions, reducing the rate down to about 4 kHz. The ReadOut System (ROS) receives event fragments from the detector readout via $\sim 1600$ optical links upon Level-1 accepts. The ROS buffers the event fragments during the Level-2 decision and forwards them on request to the Level-2 or to the event builder farm, where the events are fully built. The ROS was designed to sustain a Level-2 request rate of 16 kHz and an event builder rate of 3.5 kHz. The EF operates the last data selection accessing the fully assembled events, and sends the accepted ones to the local mass storage.

The TDAQ architecture is structured around two distinct data networks: the Data Collection (DC) and the Back-End (BE), for different data traffics. The Level-2 is connected to the DC network, together with the ROS and the event builder farm, which couples the two networks. The event builder farm, the EF and the local mass storage are connected to the BE network.

Figure 1 shows a simplified schema of the TDAQ system.

2.1. TDAQ performance in Run 1
By the end of Run 1, most of the components of the TDAQ system were operating beyond design. A challenge for the TDAQ system was posed by the increase of the pile-up, which affects the event size, thus the bandwidth, and the HLT processing time, thus the computing power.

The ROS had to cope with higher request rate and was upgraded in the course of Run 1 to enable the Level-2 farm to access all data from the innermost detectors for full tracking reconstruction and to have special values pre-computed by the front-end boards of all the
calorimeter detectors to evaluate the missing transverse energy. In addition, the typical RoI size corresponded to $\sim 5-10\%$ of the total event size, and not only $\sim 2\%$ as per design. In 2012 the ROS coped with 15 kHz Level-2 request rate plus 10 kHz missing transverse energy request rate, and an event builder rate of 5 kHz.

The HLT computing farm accounted for $\sim 1600$ machines organized in 35 so-called XPU racks and 14 dedicated EF racks. The XPU racks were connected to both data networks and they could execute both HLT workloads. Depending on the trigger requirements, the HLT resources were balanced across the two farms, as described in [4]. The output of the EF also exceeded the design values, reaching an average bandwidth of 1 GB/s in 2012.

3. TDAQ Evolution during LS1

The TDAQ system is evolving to satisfy the Run 2 physics requirements, whilst handling the expected pile-up increase.

The Fast Tracker (FTK) will be partially available during Run 2, to provide full track reconstruction as input to the HLT. That is discussed elsewhere [5].

3.1. New Topological Trigger and Central Trigger Processor Upgrade

In order to achieve additional background rejection, a Topological Processor (TP) will be introduced to the Level-1 trigger and the Central Trigger Processor (CTP) will be upgraded. That will allow to keep the trigger rate within the limit of 100 kHz, avoiding to reject events due to the saturation of trigger pipelines.

The TP will merge detailed information, including geometrical, from the trigger detectors to determine complex observables, such as $\phi$ separation between objects, invariant masses, etc. The TP has to receive a total bandwidth of about 1 Tbps, process the data in FPGAs within...
less than 100 ns, and send the results to the CTP. Changes in the CTP, the calorimeter trigger merger modules, and the interface between the muon trigger and the CTP are required.

The CTP is composed of several custom-built VME boards. The CTP-Machine-Interface board receives the signals from the LHC, while three input boards receive a total of 372 input signals from the trigger detectors. Amongst those 372 signals, 160 are selected by a switch matrix and sent to the CTP-CORE module via the back-plane, the Pattern In Time (PIT) bus. The Level-1 trigger results and the timing signals are transmitted to the sub-detectors via four output modules. The CTP-CORE will receive 192 new inputs from the Topological Processor. The PIT bus will operate at double data rate, 80 MHz, providing double the number of the inputs (320). The upgraded version of the CTP-CORE will thus form 512 trigger items instead of 256. An additional output board is planned. Details about the TP and the CTP upgrade can be found in [6], [7].

3.2. Network Evolution
The data networks have merged into a single one at the beginning of the LS1, as described in [8]. The functionality of the Back-End (BE) network is provided by an enhanced Data Collection (DC) network. The core routers of the DC network, according to the rolling replacement policy, have been refurbished, and two overlapping class B sub-networks installed.

During Run 1, a layer of concentrator switches was used between the ROS PCs and the core routers: every group of 10 ROS copper Gbit Ethernet interfaces was aggregated into two 10 Gbps links connecting to the core. In Run 2, each ROS PC will connect to the two sub-networks via two 10 Gbps links, to cope with the increased request rate.

The HLT farm accounts for 49 racks with 31 to 40 nodes each at the time of the article. The nodes are connected to a switch in each rack, connected to the DC core via a 10 Gbps link. The HLT farm could be expanded up to 80 racks, depending on the trigger needs. The data network routers are scalable, and will be able to accommodate all these ports.

3.3. ReadOut System Upgrade
To sustain a Level-1 rate of 100 kHz and an increased request rate, a new ROS is foreseen for Run 2 [8]. In addition to the two 10 Gbps links, a new purpose-built PCI board (the ROBIN card) which processes the detector inputs will be deployed. The new ROBIN is referred to as ROBIN-NP, where NP stands for No Processor, since the CPU of the host PC will do the job of the on-board processor of the current ROBIN. It is a PCI-Express card supporting twelve optical input links. The ROBIN-NP is based on the Combined Read-Out Receiver Card (C-RORC) developed by the ALICE collaboration [9], but the firmware has been redesigned.

3.4. New TDAQ Software Architecture
The new architecture of the TDAQ system is depicted in Figure 2, while the new TDAQ software applications are shown in Figure 3.

The Level-2, the event builder and the EF farms become a single HLT farm, matching the new network topology. The predetermination of the sharing of HLT resources between the processing levels will no longer be needed. The unique HLT farm reduces the event processing latency, since there is no need to pack and transport the detailed decisions for the Level-2 accepted events to the EF. The amount of ROS requests will be optimized, since a second request by the event builder in case of Level-2 accepted events will not occur anymore.

The RoIB sends at a rate of 100 kHz the RoIs via S-links to the HLT SuperVisor (HLTSV), which buffers them. The HLTSV is a single application which replaces the previous Level-2 Supervisor farm. The HLTSV serves the RoIs to the available HLT nodes and sends the clear buffer messages to the ROS.
A Data Collector Manager (DCM) and a variable number of HLT Processing Units (HLTPU) runs in each HLT node. The DCM, upon reception of a Level-1 result from the HLTSV, requests and caches the event fragments from the ROS. Then, it assigns the fragments to the available HLTPU. Once the HLT algorithms complete the selection process, fully built events are sent to the local mass storage by the DCM. The DCM builds the events depending on the trigger algorithms sequence, allowing for incremental event building, previously not possible.

In order to improve the maintainability, a new common data communication library has been deployed, based on an asynchronous I/O protocol (Boost.Asio).

3.5. New HLT core software
The main goal of the new HLT core software is to leverage the Copy-on-Write (CoW) mechanism of the Linux kernel to maximize the sharing of memory pages between processes. A dedicated process, referred to as HLTPU Mother, at the start of the data taking session creates copies of itself, the HLT workers. The HLT workers communicate to the DCM to get the data fragments and send the HLT algorithms results.

The new HLT core software is under deployment at the time of this paper.

4. First Test Results
The new TDAQ software applications have been tested during the summer of 2013. Figure 4 shows the rate of the HLTSV during a test session in the TDAQ test stand. The HLTSV could operate at 175 kHz, corresponding to 80% of the total available bandwidth. In the final design the HLTSV application will run on a node equipped with two 10 Gbps links (was two 1 Gbps links in the test stand). Moreover it will be connected to the RoIB, thus some extra CPU consumption due to reading the S-links has to be expected. This test shows that the HLTSV can sustain a rate of 100 kHz.
**Figure 3.** Schematic of the new TDAQ software applications.

**Figure 4.** Screenshot of a monitoring tool which shows the HLTSV rate.

Figure 5 shows the network scalability of the new HLT farm. On the x-axis is reported the number of HLT racks in use and on the y-axis the measured aggregated bandwidth in MB/s. Since the ROS has not been upgraded at the time of the test, the bandwidth limitation is given by the available ROS bandwidth, which corresponds to 251 Gbps. The bandwidth is saturated deploying 24 HLT racks, as expected, considering that each one has a 10 Gbps link network connection. Once the ROS will be upgraded, the available HLT input will set the maximum bandwidth.

**5. Conclusion**

The TDAQ system successfully operated beyond design during Run 1. In preparation to the enhanced LHC performance in Run 2, the TDAQ system is evolving to satisfy the requirements of the ATLAS physics program.

A Topological Processor will be added to the Level-1 trigger and the Central Trigger Processor.
Figure 5. The aggregated bandwidth with respect to the number of HLT racks.

will be upgraded accordingly. The TDAQ system will use a single data network with increased ReadOut System bandwidth. The new TDAQ software architecture merges the second and third trigger level, running them on the same nodes to ensure an optimal sharing of the resources.

The first tests of the new TDAQ software show the scalability and reliability of the new architecture.

References
[1] ATLAS Collaboration, “The ATLAS experiment at the CERN Large Hadron Collider”, JINST 3 S08003 (2008).
[2] ATLAS Collaboration, “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”, Physics Letters B 716 (2012).
[3] ATLAS Collaboration, “ATLAS, High-Level Trigger, Data Acquisition and Controls: Technical Design Report”, CERN/LHCC/2003-022, Geneva, CERN, 2003.
[4] N. Garelli et al., “Balancing the resources of the High Level Trigger farm of the ATLAS experiment”, J. Phys.: Conf. Ser. 396 (2012).
[5] M. Sochet et al., “Fast TracKer (FTK) Technical Design Report”, CERN-LHCC-2013-007 ATLAS-TDR-021 (2013).
[6] C. Ohm et al., “An upgraded ATLAS Central Trigger for 2015 LHC luminosities”, ATL-DAQ-PROC-2013-012 (2013), http://cds.cern.ch/record/1600809/.
[7] B. Bauss et al., “An FPGA based Topological Processor prototype for the ATLAS Level-1 Trigger upgrade”, JINST 7 C12007 (2012).
[8] R. Bartoldus et al., “TDAQ Phase-I Upgrade Technical Design Report”, CERN-LHCC-2013-018 ATLAS-TDR-023 (2013).
[9] H. Engel et al, “ALICE C-RORC as CBM FLES Interface Board Prototype”, online www-alt.gsi.de/information/wti/library/scientificreport2011/PAPER/PHN-NQM-CBM-39.pdf.