Journal of Physics: Conference Series

Operating the ATLAS data-flow system with the first LHC collisions

Operating the ATLAS data-flow system with the first LHC collisions.

Nicoletta Garelli, on behalf of the ATLAS TDAQ Collaboration [1]
CERN

E-mail: nicoletta.garelli@cern.ch

Abstract. In this paper we will report on the operation and the performance of the ATLAS data-flow system during the 2010 physics run of the Large Hadron Collider (LHC) at 7 TeV. The data-flow system is responsible for reading out, formatting and conveying the event data, eventually saving the selected events into the mass storage. By the second quarter of 2010, for the first time, the system will be capable of the full event building capacity and improved data-logging throughput.

We will in particular detail the tools put in place to predict and track the system working point, with the aim of optimizing the bandwidth and the computing resource sharing, and anticipate possible limits. Naturally, the LHC duty cycle, the trigger performance, and the detector configuration influence the system working point. Therefore, numerical studies of the data-flow system capabilities have been performed considering different scenarios. This is crucial for the first phase of the LHC operations where variable running conditions are anticipated due to the ongoing trigger commissioning and the detector and physics performance studies. The exploitation of these results requires to know and track the system working point, as defined by a set of many different operational parameters, e.g. rates, throughput, event size. Dedicated tools fulfill this mandate, providing integrated storage and visualization of the data-flow and network operational parameters.

1. Introduction

During 2010 the Large Hadron Collider (LHC) at CERN provided proton-proton collisions with increasing intensity at the center-of-mass energy of $\sqrt{s} = 7\,\mathrm{TeV}$.

ATLAS (A Toroidal LHC ApparatuS) [2], one of the two general purpose experiments at the LHC, successfully recorded $45.0\,\mathrm{pb}^{-1}$ of data, corresponding to 92% of the total delivered luminosity, as shown in Figure 1. At the time of this conference, a peak instantaneous luminosity of $10^{32}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ in ATLAS was reached. The ATLAS experiment reaps the results of the two previous years of cosmic data-taking and detector commissioning: an incredibly high data-taking efficiency has been achieved and the trigger and data-acquisition system (referred to as TDAQ) coped extremely well with the evolving requirements, operating even beyond the design.

This paper focuses on the ATLAS TDAQ performance during the first year of proton-proton collisions, describing the status of the hardware farms, the tools put in place to predict and track the system working point, and the mechanism dedicated to the maximization of the data-taking efficiency.

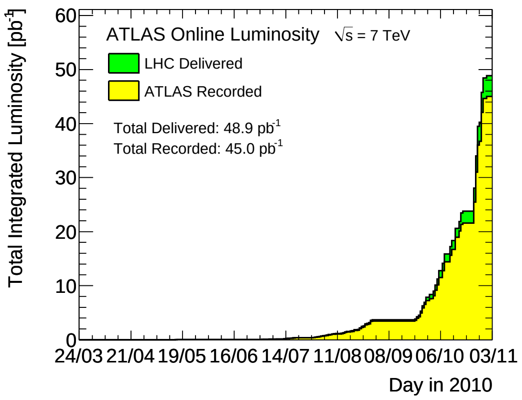

Figure 1. ATLAS total integrated luminosity. The plot shows the total integrated luminosity at the center-of-mass energy of $\sqrt{s}$ = 7 TeV delivered to (LHC delivered) and recorded by (ATLAS Recorded) ATLAS in 2010.

2. The TDAQ system

The TDAQ system has been designed to convey data from the detector front-end to the mass-storage, reducing the event rate from the LHC nominal collision rate of 40 MHz down to 200 Hz [3]. This is achieved with a three-level trigger system. The first one, the Level-1, is implemented in custom built electronics which analyzes the information coming from the muon chambers and the calorimeters to produce a coarse event selection. The other two levels, collectively called High-Level Trigger (HLT), are software-based and have access to the detector data at the full granularity. The second trigger level (Level-2) has tight timing constraints and thus it accesses only a subset of event data in the so-called Regions of Interest (RoIs). The RoIs are limited areas in the $\eta\phi$ plane [1] defined by the Level-1. Normally, a RoI corresponds to $\sim 2\%$ of the total event data. The last trigger level, called Event Filter (EF), analyzes the events selected by the Level-2 and sends the accepted ones to the Sub-Farm Output (SFO). In the SFO the events are streamed into local data files, which are asynchronously moved to the mass storage.

The backbone of the TDAQ system is composed of two Gigabit Ethernet networks. In the so-called Data Collection area the data movement is organized around a push-pull architecture: the Read-Out System (ROS) receives via $\sim$ 1600 optical fibers the event fragments from the detector and provides them on request to the Level-2 and the Event Builder (EB). The EB, which decouples the two network domains, is composed of Sub-Farm Input (SFI) applications. They merge all the data fragments to form ATLAS events and send them to the EF, via the second network, the EF network. The EB handles a wide range of event sizes, also up to $O(10)$ MB. The event size is defined by the number of active detector channels and by the front-end zero suppression configuration: during the 2010 run the typical event size was $\sim$ 1.5 MB.

A schematic overview of the TDAQ system is depicted in Figure 2. As shown, in 2010 the TDAQ system was working beyond design conditions, with an LHC collision rate of only $\sim$ 1 MHz. The loose trigger rejection factors and the consequent higher data throughput allowed to commission the trigger physics configuration.

2.1. Calibration Event Data

Unlike physics events, calibration events require only part of the full event data information, normally resulting in a size lower than 500 kB. Two mechanisms have been developed to allow the recording of calibration data minimizing the TDAQ resource usage.

Dedicated Level-2 algorithms select events for the calibration of the different ATLAS components. Those events are built based on a list of detector identifiers: this mechanism

[1] The pseudo-rapidity η is defined as $\eta = -\ln[\tan(\theta/2)]$, where the polar angle θ is the angle between the particle momentum and the beam axis. ϕ is the azimuthal angle, measured around the beam axis.

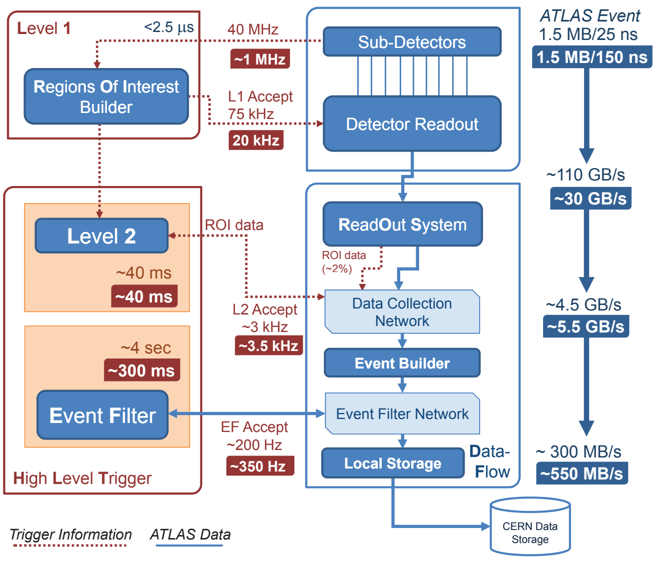

Figure 2. ATLAS TDAQ architecture. The trigger path is sketched on the left, while the data one on the right. The design parameters are reported for each component. The highlighted numbers refer to trigger decision time, rate and bandwidth at the time of the conference.

is called Partial Event Building (PEB). An event selected by the Level-2 both for physics and for calibration purposes has to be fully built. In this case, the small calibration event can be recovered after the EF decision via the so-called stripping mechanism.

Collectively, these functionalities allowed to reduce the calibration bandwidth to $\sim 2.5\%$ of the total bandwidth.

3. The TDAQ Sub-Farms Status

All of the TDAQ sub-farms are fully deployed with respect to the design, with the exception of the HLT one. The HLT computing resources are in fact progressively installed following the ATLAS needs, which depend on many factors, as the expected instantaneous luminosity. In 2010 the HLT farm reached 50% of the final size. It accounts for 27 XPU racks, each one with 30 process units. The name XPU indicates that they can be configured on a run by run basis to work either as Level-2 or EF processing unit. The possibility to move computing power between the two HLT sub-farms allows high flexibility to meet the trigger needs, especially during the commissioning phase. The maximum average decision time of the trigger algorithms depends in fact on the number of processing units in use.

On the EF network the XPU racks have a 2 Gbps link per rack, resulting in a maximum EF installed network of ~ 6 GB/s, which is smaller than the EB one of ~ 10 GB/s. With the announced increase of luminosity expected at the end of September 2010, it was clear that the EF farm was in need for further computing power and bandwidth. 9 out of the 50 dedicated EF racks foreseen by the TDAQ design have been installed in the course of the summer, and the XPU resources have been redistributed amongst the Level-2 and the EF farms, as shown in Table 1. These new racks have a 10 Gbps link each, resulting in a throughput of ~ 11 GB/s. This allowed to fully exploit the installed EB resources.

HLT Resources	$\mathcal{L} < O(10^{31})\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$	$\mathcal{L} \sim 5 \cdot 10^{31}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$
Max EF Bandwidth	$\sim 4.5\,\mathrm{GB/s}$	$\sim 15\,\mathrm{GB/s}$
L2 racks	9 XPU	12 XPU
EF racks	18 XPU	15 XPU & 9 EF

Table 1. HLT sub-farms details. The maximum EF bandwidth and the Level-2 and EF rack sharing before and after the increase of LHC luminosity are reported.

3.1. The Sub-Farm Output Farm

In the summer of 2010 the SFO farm has been renewed with 6 nodes, each one equipped with 3 RAID array cards, sporting for a total capacity of $\sim$ 50 TB, which corresponds to 2 days of disk buffer in case of mass storage failure, while recording data at the design bandwidth.

By design, the SFO farm should sustain a rate of 300 MB/s using only 4 out of the 6 available nodes. In 2010 it was regularly used beyond design specifications, operating at $\sim$ 550 MB/s. In some special runs, a sustained input throughput of $\sim$ 1 GB/s has been achieved.

4. Monitoring and predicting the TDAQ working point

Monitoring the TDAQ working point is mandatory to guarantee a successful and efficient data-taking. In particular, it is essential to dynamically reconfigure the system in response to the changes of the detector conditions, the beam position and the LHC luminosity.

Three tools have been developed to study the evolution of the TDAQ performance, anticipating possible limits. These tools were extremely useful for taking decisions concerning, for example, the sharing of the computing power between the Level-2 and the EF, and the activation of specific trigger algorithms.

4.1. Network Monitoring Tool

The network monitoring tool, detailed in [4], is an integrated system for collecting and displaying with the same look and feel network and computer operational information, environmental conditions, and data-taking parameters. It has been tuned for monitoring the network resources, but assures a transparent access to data collected by different monitoring tools, allowing to easily correlate the network utilization with the TDAQ performance.

4.2. HLT Resources Monitoring Tool

A dedicated calibration trigger collects data for evaluating the HLT performance [5], exploiting the PEB and the stripping features. The recorded information are elaborated offline to report trigger rates, the execution order, the processing time, and the Level-2 access pattern and latency. The information coming from rejected events are stored as well for studying the HLT execution costs. The HLT resource monitoring tool is fundamental to spot trigger inefficiencies and to predict the rejection factor of each algorithm.

4.3. Simulation of the TDAQ operational capabilities

A simple simulation of the TDAQ operational capabilities has been developed to enhance the network and the HLT resource monitoring tools information. The simulation gives an overview of the TDAQ resource usage, helping in decisions like the XPU sharing or the deployment of additional computing power.

The operation envelope is evaluated as a function of the XPU sharing between Level-2 and EF, the number of installed specific EF racks, the processing time and the rejection factor of the algorithms. For example, Figure 3 shows the simulation obtained considering a Level-1 rate of 75 kHz, an ATLAS event size of 1.5 MB and 10 installed dedicated EF racks. The x-axis reports the Level-2 rejection power, while the y-axis the number of XPU racks configured as Level-2 (on the left), which has a direct correspondence to the maximum average Level-2 processing time (on the right). The colored curves represent the maximum EF processing time with respect to the number of XPU racks in use in the EF farm. The vertical dashed lines delimit regions in which the system operates at the indicated EB bandwidth and EF rejection factor to sustain an SFO bandwidth of 500 MB/s. In these conditions, to sustain an event building rate of 3.5 kHz the maximum EF processing time is 1.8 s, maintaining a fair XPU sharing between the Level-2 and EF. To allow for a longer EF processing time, as the design, it is necessary to either decrease the event building rate or to install more EF computing power.

At the time of the conference, the Level-2 and EF processing time, rejection factor and number of racks in use were still far from the boundary working condition, leaving a good margin for running the system during the commissioning phase.

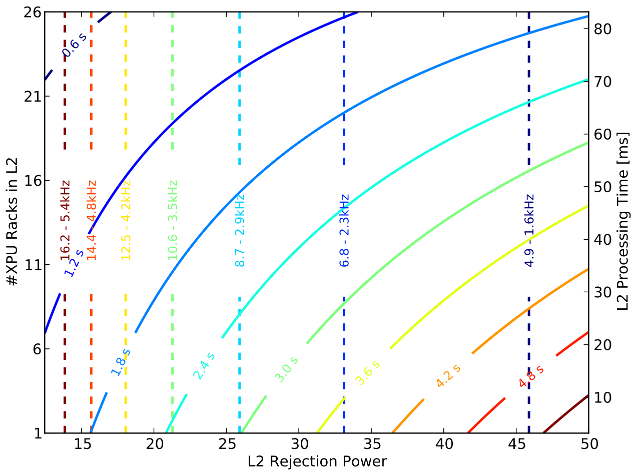

Figure 3. Simulation of the TDAQ operational capabilities. The x-axis reports the Level-2 rejection factor, the y-axis the number of XPU racks configured as Level-2 (left), and the maximum average Level-2 processing time (right). The colored curves represent the maximum EF processing time. The vertical dashed lines delimit the region in which the system operates with the indicated EB bandwidth and EF rejection factor, needed to sustain an SFO throughput of 500 MB/s. A Level-1 rate of 75 kHz and an event size of 1.5 MB have been considered.

5. The ATLAS Run Efficiency

A main challenge for the 2010 run was to maximize the data-taking efficiency. Therefore, automated procedures handled by the Detector Control System (DCS) and by a dedicated control framework called Expert System have been developed.

To ensure the safety of the silicon sensors, the nominal voltage can be set only when the LHC declares the circulating beams as stable. It is however possible to start a run well-before this. As soon as the physics data-taking conditions are met, an automatic procedure takes care of switching-on the inner detector and setting the proper trigger configuration. A similar procedure is used as soon as the beams become unstable, towards the end of the LHC fill. In this way, the up-time of the run during stable beams condition is maximized.

During a run some of the about 90 millions of the readout channels could start malfunctioning. These faulty components can be automatically disabled from the read-out system configuration, preventing dead-time and inefficiencies. The same components can be automatically re-enabled at run time after the needed reconfiguration.

The ATLAS data-taking efficiency, shown in Figure 4, is evaluated in two ways: as the fraction of time in which ATLAS is recording data and in which ATLAS is recording physics data with all the detectors at nominal voltages. Thanks to the described automatic procedures ATLAS achieved an high run efficiency of 96.4% (92.9% with physics configuration).

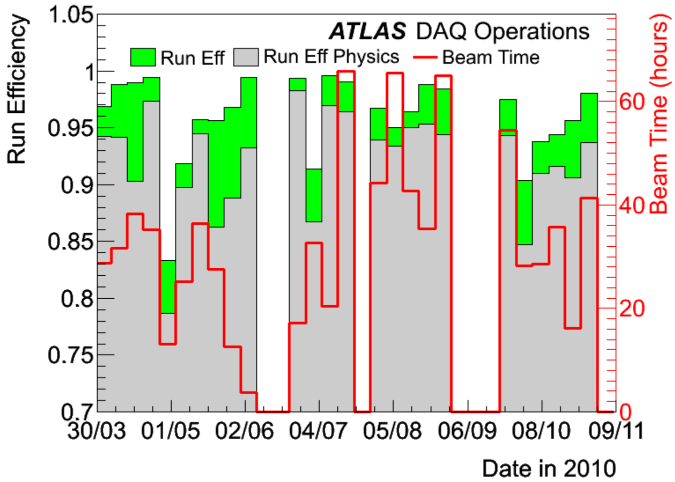

Figure 4. ATLAS weekly data-taking efficiency in 2010. In 2010 the LHC delivered stable beams at the center-of-mass energy of $\sqrt{s} = 7\,\text{TeV}$ for 487 hours. The "Run Eff" area shows the data-taking efficiency, while the "Run Eff Physics" the efficiency with physics configuration.

6. Conclusion

In 2010 the ATLAS TDAQ system has been operated beyond the design requirements to meet the evolving conditions of the accelerator and the detector, while commissioning the trigger and providing valuable physics data.

The PEB and the stripping allowed for continuously tuning the detector and commissioning the trigger, maximizing the rate at which calibration events can be collected. Dedicated monitoring tools have been put in place to monitor the TDAQ working point, establishing the optimal HLT computing resource balancing.

2010 was a successful year for the ATLAS experiment which achieved an incredibly high run efficiency of 96.4% (92.9% for physics).

References

[1] ATLAS Collaboration, "The ATLAS Trigger/DAQ Authorlist, version 4.0", ATL-DAQ-PUB-2010-002, CERN, Geneva, 2010, http://cdsweb.cern.ch/record/1265604.

[2] ATLAS Collaboration and G. Aad et al., "The ATLAS experiment at the CERN Large Hadron Collider", J.Instrum.3 S08003 (2008).

[3] ATLAS Collaboration, "ATLAS, High-Level Trigger, Data Acquisition and Controls: Technical Design Report", CERN/LHCC/2003-022, Geneva, CERN, 2003.

[4] Dan Savu et al., "Integrated System for Performance Monitoring of ATLAS TDAQ Network", in Proc. CHEP 2010, Taipei, Taiwan.

[5] Antonio Sidoti, Wojtek Fedorko, Rustem Ospanov and Martin zur Nedden, "Diagnostic Systems and Resource Utilization of the ATLAS High Level Trigger", ATL-DAQ-PROC-2010-047. It will appear in Nuclear Science Symposium Conference Record (NSS/MIC) 2010 IEEE.