PROCEEDINGS OF SCIENCE

The ATLAS Data Acquisition and High Level Trigger Systems: Experience and Upgrade Plans

Reiner Hauser*
On behalf of the ATLAS Collaboration
Michigan State University
E-mail: Reiner.Hauser@cern.ch

The ATLAS Data Acquisition (DAQ) and High Level Trigger (HLT) system are designed to reduce the Level 1 rate of 75 kHz (upgradable to 100 khz) to a few kHz event build rate after Level 2 and a few hundred Hz output rate to disk. It has operated with an average data taking efficiency of about 94% during the recent years. The performance has far exceeded the initial requirements, with about 5.5 kHz event building rate and 400 Hz of average output rate in 2012, only considering data for prompt physics analysis. Several improvements and upgrades are foreseen in the upcoming long shutdowns, both to simplify the existing architecture and improve the performance. On the network side new core switches will be deployed and possible use of 10 Gbps Ethernet links for critical areas is foreseen. An improved read-out system to replace the existing solution based on PCI is under development. A major evolution of the HLT system foresees a merging of the Level 2 and Event Filter functionalities on a single node, including the event building. This will represent a big simplification of the existing system, while still maintaining the flexibility of the Region of Interest based approach. It will furthermore open up new optimizations and simplifications in the existing HLT code.

36th International Conference on High Energy Physics,
July 4-11, 2012
Melbourne, Australia

*Speaker.

PoS(ICHEP2012)517

1. Introduction

The ATLAS detector [1] is one of two general purpose particle detectors at the Large Hadron Collider. It consists of several sub-detectors with about 10^8 read-out channels in total. The trigger and data acquisition (TDAQ) system [2] is responsible for reducing the initial collision rate (20 MHz in 2009 to 2012) to something that can be stored off-line. We present our experience with the current system during this first running period and discuss plans on how it will evolve during the upcoming long shutdown in 2013-2014.

2. The ATLAS Trigger/DAQ System

The ATLAS trigger [3] consists of three levels. The first level is implemented in hardware and reduces the data rate from the initial bunch crossing rate of 40 MHz (20 MHz in 2012) to about 75 kHz (about 65 kHz in 2012). The second and third (called Event Filter) trigger level are purely in software and executed on a farm of commercial PCs.

The first level trigger is able to identify potentially interesting regions in the detector and to provide this information to a special hardware called Region of Interest Builder (RoI Builder). The RoI Builder combines all the information related to an event and forward it to a set of supervisor nodes, which are responsible for scheduling the event processing on the LVL2 farm.

The second level trigger (LVL2) processes only a fraction of event data, requesting data fragments pertaining to the Regions of Interest to the ReadOut System (ROS). On average, only 5-10% of the full event data is needed to take a decision, thus reducing the required aggregated network throughput. During 2012 data taking, the LVL2 output peak rate was about 5.5 kHz.

During LVL2 processing, the event data fragments are stored in ReadOut Buffers (ROB), distributed over the ReadOut System. A set of event building nodes collect fragments of accepted events and forward them to the Event Filter (EF) farm. Here offline-like algorithms analyze the full data, reducing the average rate to about 400 Hz of events ready for prompt analysis. On top of that, up to 200 Hz of events are stored for delayed analysis, and an additional rate is provided by data generated by calibration triggers. Events accepted by the EF are sent to a small set of data loggers, which store them temporarily on disk. Data are then forwarded to the central CERN storage system.

The actual implementation of the HLT/DAQ network can be seen in Fig.1. Multiple central core switches are deployed, both for the data collection (DC) network for level 2 and event building and for the back-end network (BE) for the event filter and data logger. In addition there are concentrator switches inside each rack of HLT nodes and for the read-out system machines.

HLT nodes come in two flavors, depending on how they are connected to the core switches. The XPU nodes can be configured both as level 2 processing units (L2PU) and event filter processing units (EFPU), while EF nodes can only be used as the latter. XPU nodes therefore provide flexibility in balancing the needed processing power between level 2 and event filter. However, this decision has to be made before a run is started and cannot be changed later.

3. Experience in 2009-2012

During the 2009 to 2012 data taking period the DAQ/HLT system operated at an average

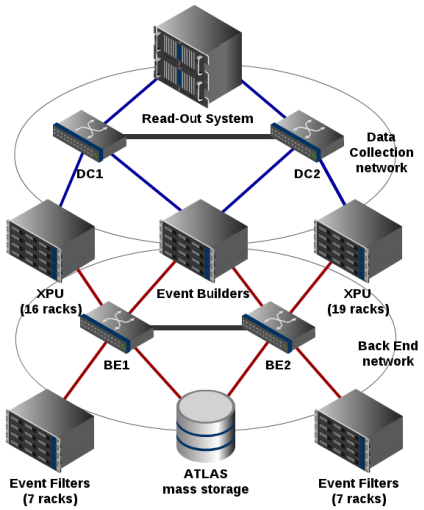

Figure 1: Data Network Structure

	Design Rate	in 2012	Design CPU Time	in 2012	Design Bandwidth (avg)	in 2012
LVL1	75 kHz	65 kHz	-	-	100 GB/s	70 GB/s
LVL2	3 kHz	5.5 kHz	40 ms	60 ms	7.5 GB/s	5 GB/s
EF	200 Hz	400 Hz	4 s	1 s	300 MB/s	550 MB/s

Table 1: Rates and Bandwidths during 2012 vs. Initial Design

efficiency of 94%. During this time it had to cope with a large range of different conditions, including a peak luminosity that spans several orders of magnitude (see Fig. 2).

The typical event rates and bandwidth in 2012 can be seen in Table 1, together with the initial design values. As can be seen in several cases the actual performance is higher than initially planned.

During this time the HLT machines went through several rolling replacement cycles. At the end of 2012 three different generations of motherboards were in use. In addition, all other machines (event builder, read-out system, data logger) have also gone through at least one replacement cycle. The data logger system could be rather easily up-scaled with more nodes and disks to accommodate the additional requests that emerged during running (from 200 Hz to 400 Hz). In addition to the normal accepted events the system also writes out so-called delayed data that will not be immediately processed. This adds another 100-200 Hz to the output rate.

The heterogeneity of the HLT system proved to become an issue for the various supervising tasks who employed a too simple scheduling algorithms. Operational problems like loosing a rack of machines left the system in an unbalanced state. Changes to the logical organization of the system and the schedulers solved these problems.

Many system and kernel parameters were tuned over the past years for optimal performance. In some cases the software had to work in tandem with the network switch configuration to provide an adequate quality of service and avoid noticable latency drops for important packets.

With the initial low start-up luminosity the LVL2 system could run full scan algorithms e.g.

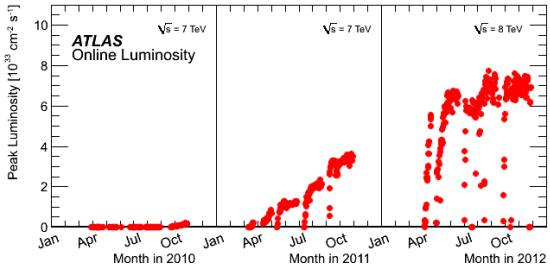

Figure 2: Peak luminosities in 2010-2012

for tracking. This had to be turned off at higher initial peak luminosity, however. The missing transverse energy (MET) triggers proved to be a bigger issue. At LVL2 only the MET information from LVL1 was available with a much lower resolution than desired. That had the effect that many trigger items including MET at LVL2 were simply accepted and forwarded to the event filter, adding a large load on the event building system. Since the calorimeters in ATLAS make up about half of the total event size, the performance of the current read-out system would not allow to request all this data at the required rate. A solution was finally found by the calorimeter front-end boards providing an energy sum for each board in the data. A special LVL2 request can ask only for these few words, thereby reducing the data volume and still providing an accurate MET value with a quality comparable to the EF.

When ATLAS decided to increase the average output rate of 200 Hz to 400 Hz, the size of the raw event data also became an issue. This was solved with transparent compression of the raw events. For performance reasons the compression could not be done on the current data logger nodes, so the actual compression is now done off-line during a merging step. The plan is to do this inside the HLT in the new architecture and distribute this task among the thousands of HLT nodes.

Overall the DAQ/HLT system has been exceptionally reliable and performant during the last years of data taking. Most challenges could be met with reasonable effort, and the system itself scales in most areas by simply adding more components.

4. Evolution of the ATLAS Trigger/DAQ System

The current DAQ/HLT system has evolved historically and some of the initial assumptions are either no longer correct, or proved to be wrong from the outset. Together with the still remaining issues of the current system these drive the current plans for an evolution during the 2013-2014 shutdown.

Originally the LVL2 environment was thought to be much different from the event filter. While the latter was always meant to run adapted off-line code, the assumption was that LVL2 would use specially written algorithms running in a constrained (possibly real-time) enviroment. This turned out to be not true. By 2005 it became clear that the same software environment could be used for both LVL2 and EF. In the ATLAS case this is an adapted Athena/Gaudi [4] framework. As a follow-on effect the two systems actually literally share the same code for configuration and menu steering. The only difference between a LVL2 and an EF algorithm is that the former starts processing using the regions of interest and uses a slightly different way of requesting its data.

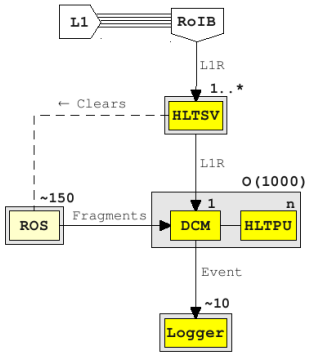

Figure 3: New DAQ/HLT architecture

The initial LVL2 system assumed a single application per node, using multi-threading to process several events in parallel while hiding the latency of network accesses. In practice this turned out to be unmaintainable since the large amount of re-used off-line software did not allow enough control to keep the system thread-safe. The advent of multi-core CPUs instead of even faster single-core CPUs helped alleviate this problem. Currently a LVL2 node is running one LVL2 application per core. The disadvantage of this scheme is the larger number of network connections required, and the large amount of memory used which has a lot of overlap between processes.

As already mentioned the balancing of the LVL2 and EF farm is static right now. In some realistic use cases this is not desirable. At the beginning of a fill with high luminosity the processing power is needed at LVL2 for rejection. Once the luminosity has dropped one wants to enable e.g. B physics triggers which typically require a full scan of the tracking detectors. Therefore the latter will typically run on the event filter nodes. A more dynamic approach seems desirable.

The evolution of the DAQ/HLT systems aims to address all these issues. Based on the above observations the idea is to merge the LVL2 and EF processing steps on a single node. Event processing will start with level 2 like algorithms, requesting data based RoIs as before. At some point the steering will decide to switch to an EF like processing mode. For this the event building will have to be done on the same HLT node, thereby avoiding to request the data twice from the read-out system.

A single application per node will be responsible for all data transfers. On the other hand there will be multiple HLT applications per node doing the actual processing. To reduce the memory usage these applications will all fork from a common mother process after configuration only, taking advantage of the memory managment of a typical Unix/Linux system.

A simple depiction of the new system can be seen in Fig.3. The number of data flow applications is almost reduced by a factor of two. The HLT nodes are connected in a much more uniform way to the underlying network. While this system can emulate the current one in terms of processing and data access, it is more interesting to consider the additional flexibility. E.g. the distinction between LVL2 and EF is no longer a hard decision. Any data that is requested by LVL2 stays on the same node and won't be requested again for the event building. This allows to run algorithms that need a full sub-detector but not the full event at some intermediate point. The strategy on when to request full event building is also under study. Rather than waiting for the full LVL2 like

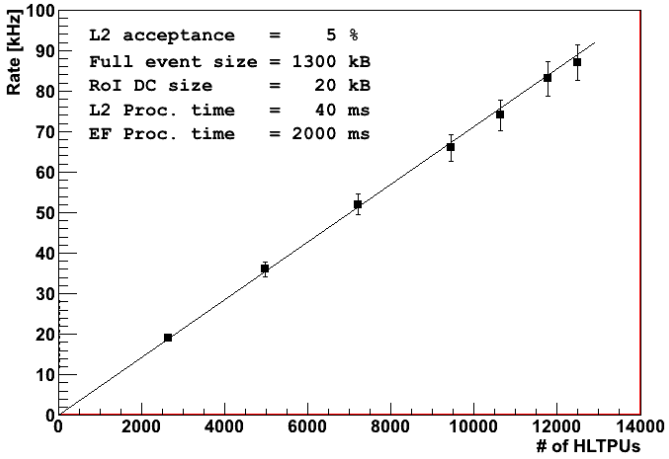

Figure 4: Scalability results with new prototype: Total rate vs. number of HLT nodes.

processing to be done, it could be initiated immediately when the first trigger chain accepts the event.

Initial scalability results with a prototype for the new architecture can be seen in Fig. 4. for a given point in parameter space that reflects our current estimations for the future requirements.

The architectural change implies a different connectivity at the network level. At this point in time the central core switches used for the data collection and back-end networks have reached their end of warranty. Their replacements will provide a significantly higher bandwidth and this intervention will also be used to set up the new network infrastructure. Details regarding the redundancy and actual number of central core switches are still under discussion.

In addition to these changes to the overall architecture there will be also upgrades to the components which have a custom hardware part. The Region of Interest builder is currently a fully custom hardware solution. Tests are on-going to see if it can be replaced with a software solution running on a state of the art PC.

For the read-out system a redesign of the custom input buffers is foreseen. This will allow the transition from PCI-X to PCI Express technology, a denser packing of links, and a higher overall performance. A 10 Gbps link into the data flow network will improve the maximum possible output rate.

This combination of many improvements of individual components and an overall simpler and yet more flexible architecture will make the ATLAS TDAQ system well suited for the increased demands during the next period of LHC running.

References

[1] The ATLAS Experiment at the CERN Large Hadron Collider, JINST 3 S08003, 2008.

[2] High-Level Trigger, Data Acquisition and Controls TDR, CERN/LHCC/2003-022, 2003.

[3] Performance of the ATLAS Trigger System in 2010, Eur. Phys. J. C 72:1849, 2012

[4] ATLAS Computing TDR, CERN/LHCC/2005-022, 2005.