The ATLAS Data Acquisition and Trigger: concept, design and status

K. Kordas [a], M. Abolins, I. Alexandrov, A. Amorim, I. Aracena, S. Armstrong, E. Badescu, J. T. M. Baines, N. Barros, H. P. Beck, C. Bee, M. Bellomo, M. Biglietti, R. Blair, J. A. C. Bogaerts, T. Bold, M. Bosman, D. Burckhart-Chromek, M. Caprini, C. Caramarcu, G. Carlino, B. Caron, M. P. Casado, G. Cataldi, M. Ciobotaru, G. Comune, P. Conde-Muino, F. Conventi, A. Corso-Radu, R. Cranfield, K. Cranmer, G. Crone, D. Damazio, J. Dawson, A. De Santo, T. Del Prete, M. Della Pietra, A. Di Mattia, M. Diaz-Gomaz, R. W. Dobinson, M. Dobson, A. Dos Anjos, A. Dotti, G. Drake, N. Ellis, D. Emeliyanov, Y. Ermoline, E. Ertorer, S. Falciano, R. Ferrari, M. L. Ferrer, D. Francis, S. Gadomski, S. Gameiro, H. Garitaonandia, G. Gaudio, O. Gaumer, S. George, A. Gesualdi-Mello, R. Goncalo, B. Gorini, E. Gorini, B. Green, S. Haas, W. N. Haberichter, H. Hadavand, C. Haeberli, J. Haller, J. Hansen, R. Hauser, S. J. Hillier, A. Höcker, R. E. Hughes-Jones, M. Joos, S. Kabana, A. Kazarov, A. Khomich, G. Kieft, G. Kilvington, J. Kirk, S. Klous, T. Kohno, S. Kolos, N. Konstantinidis, A. Kootz, K. Korcyl, V. Kotov, A. Kugel, M. Landon, A. Lankford, L. Leahu, M. Leahu, G. Lehmann-Miotto, M. J. Le Vine, W. Liu, C. Lowe, L. Luminari, T. Maeno, R. Männer, L. Mapelli, B. Martin, F. Marzano, J. Masik, R. McLaren, T. McMahon, C. Meessen, C. Meirosu, M. Mineev, A. Misiejuk, R. Moore, P. Morettini, G. Mornacchi, M. Müller, R. Murillo-García, Y. Nagasaka, A. Negri, A. Nisati, C. Osuna, C. Padilla, N. Panikashvili, F. Parodi, E. Pasqualucci, T. Pauly, V. Perera, V. Pérez-Réale, J. Petersen, J. L. Pinfold, B. Pope, M. Portes de Albuquerqu, C. Potter, K. Pretzl, D. Prigent, M. Primavera, P. Rheaum, S. Robertson, C. Roda, Y. Ryabov, D. Salvatore, C. Santamarina-Rios, D. A. Scannicchio, C. Schiavi, J. L. Schlereth, I. Scholtes, M. Seixas, A. Sidoti, S. Sivoklokov, J. Sloper, E. Sole-Segura, I. Soloviev, R. Soluk, S. Spagnolo, R. Spiwoks, R. Stamen, S. Stancu, E. Stefanidis, J. Strong, S. Sushkov, M. Sutton, T. Szymocha, S. Tapprogge, S. Tarem, Z. Tarem, P. Teixeira-Dias, E. Thomas, R. Torres, F. Touchard, L. Tremblet, N. G. Unel, G. Usai, B. Vachon, J. Van Wasen, W. Vandelli, L. Vaz Gil Lopes, A. Ventura, V. Vercesi, J. Vermeulen, H. von der Schmitt, A. Warburton, A. Watson, T. Wengler, P. Werner, S. Wheeler, F. Wickens, W. Wiedenmann, M. Wielers, M. Wiesmann, E. E. Woehrling, X. Wu, Y. Yasu, M. Yu, F. Zema, H. Zobernig

[a]INFN-Laboratori Nazionali di Frascati, via E. Fermi 40, I-00044, Frascati (RM), Italy

ATL-DAQ-CONF-2007-022
23 August 2007

This article presents the base-line design and implementation of the ATLAS Trigger and Data Acquisition system, in particular the Data Flow and High Level Trigger components. The status of the installation and commissioning of the system is also presented.

1. INTRODUCTION

The ATLAS experiment is designed to observe collisions between protons of 7 TeV. These will be the highest energy collisions in a controlled environment to-date and they are going to be provided by the Large Hadron Collider (LHC) at CERN, by mid-2008. At nominal operation conditions, bunches of 10^{11} protons will cross each other at 40 MHz, resulting in ~ 25 proton-proton interactions per bunch crossing at the centre of ATLAS. Nevertheless, only a small fraction of this $\sim$ 1 GHz event rate results in interesting physics processes. The Trigger and Data Acquisition (TDAQ) system of ATLAS has to select a manageable rate of such events for permanent storage and further analysis. The amount of information to be recorded is about 1.6 MB per event and we aim in keeping ~ 200 events/s.

We are in the process of assembling the system. The functionality and performance is verified in a number of test beds and the system as-

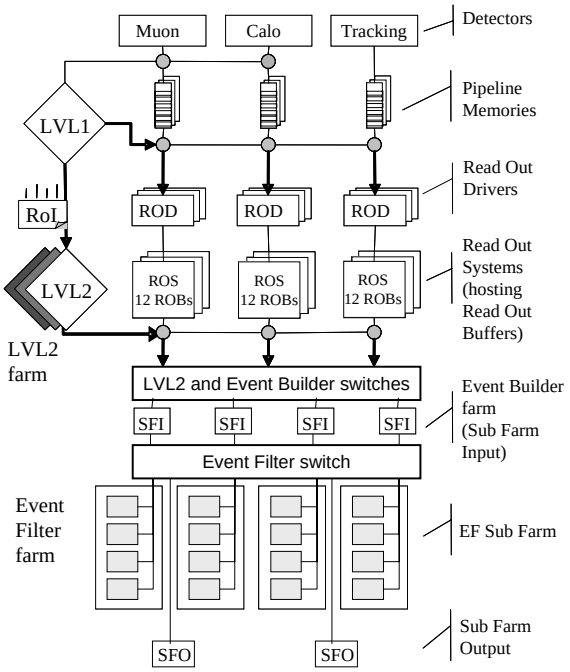

Figure 1. The ATLAS Trigger and Data Acquisition system.

assembled is used in commissioning data-taking with the various ATLAS sub-detectors as the experiment gets completed.

2. CONCEPT AND DESIGN

The ATLAS TDAQ is designed in an economical way which takes advantage of the characteristics of high energy collider physics; it uses the concept of the "Regions of Interest" in order to use the minimum amount of data needed to reach a trigger decision. Thus, the ATLAS TDAQ has moderate requirements on the network which transfers the data between its' components. In order to also minimize the required CPU resources at the trigger level, the trigger uses sequences of "feature extraction" and "hypothesis testing" algorithms. The execution order of these algorithmic pairs is arranged such that an event is rejected as early as possible.

The event selection is performed in three levels. An overview of the system is seen in Fig. 1 and can be viewed as made of three parts: the first level trigger (LVL1); the High Level Trigger (HLT), which is composed of the next two levels of triggering; and the Data Flow system which deals with temporary buffering of the data, serving them to the HLT in the form needed, and eventually to storage.

LVL1 uses coarse calorimeter and muon detector information to select events at 75 KHz (upgradeable to 100 kHz), reaching its' decision within 100 bunch crossings (2.5 μs). During this time, the front-end electronics of the various subdetectors keep the complete event data in pipeline memory buffers. Data for rejected events are discarded, while the data for selected events (up to 160 GB/s) are passed via the Readout Drivers (RODs) into the Readout Buffers (ROBs), hosted in the Readout Subsystem (ROS) PCs. Event data remain there and are pulled by the second level trigger (LVL2) and by the Event Builder (EB) nodes on demand.

LVL2 provides an additional rejection factor of 20-30, bringing the rate to ~ 3 kHz with an average latency of $\mathcal{O}(10)$ ms. For each event accepted by LVL1, a list of the Regions of Interest (RoIs) is given to LVL2. The list contains the positions (in η, ϕ coordinates) of all "interesting" objects found by LVL1 and it is assembled and communicated to LVL2 by the Region of Interest Builder (RoIB). LVL2 then accesses the appropriate ROSs to pull and analyze data from the ROBs corresponding to the Regions of Interest; thus, the LVL2 uses only $\sim 2\%$ of the full event information to take a decision and needs only 3 GB/s to be extracted from the ROSs.

Subsequently, the EB nodes, also known as the "Sub-Farm Input" (SFI) nodes, collect all data from the ROSs at the LVL2 accept rate of about 6 GB/s. Upon request, the SFIs provide fully assembled events to the Event Filter (EF) farm, which is the third level trigger. The EF analyzes the entirety of each event data to achieve a further rate reduction to ~ 200 Hz, with a latency of $\mathcal{O}(1)$ second per event. Accepted events are sent for local TDAQ storage to the "Sub-Farm Output"

(SFO) nodes. From there, the data are pulled to the central mass storage facility at CERN.

3. IMPLEMENTATION OF THE DATA FLOW AND HIGH LEVEL TRIGGER

The only custom-built hardware in TDAQ are the LVL1 trigger, the RoI Builder and the custom-built cards hosting the ROBs inside the ROS PCs. The complete Data Flow and HLT system will consist of $\mathcal{O}(3000)$ PCs running multi-threaded C++ software on a Linux platform, interconnected via Gigabit Ethernet in a multi-layer network topology. The flow of data out of the ROSs and onwards complies to a pull scenario: data are requested according to needs from subsequent clients.

3.1. Data Flow for LVL2 and the Event Builder

After a LVL1 accept, each detector pushes its' data from the front-end electronics to its' RODs and subsequently into the ROBs. The connection between the 1600 (detector specific) RODs and the (generic) ROBs is a point-to-point optical link, conforming to the S-LINK protocol [1], which provides a data throughput up to 160 MB/s. The ROBs are implemented as buffers hosted in custom-made PCI cards (ROBINs), with each card hosting three ROBs. The ROBINs are in turn hosted in the ROS PCs; each ROS PC hosts a maximum of four ROBINs, for a total of 12 ROBs. The 1600 ROBs in the system are thus hosted in about 150 ROS PCs. For each data request, the ROS PC fetches the data from the needed ROBINs via the PCI bus and groups them into a data fragment which is sent back to the requester. In addition to the PCI interface, each ROBIN card has a NIC which can be connected to the Gbit Ethernet network of the LVL2 and event building systems.

While the data are buffered into the ROSs, the RoIB collects RoI information from the LVL1 calorimeter and muon triggers and from the LVL1 Central Trigger Processor. The RoIB is implemented as a custom VMEbus system. This information is put in the "L1Result" message and is forwarded to one of the LVL2 Supervisors (L2SVs) in a round-robin fashion. Each L2SV serves a sub-farm of Processing Units (L2PUs) and assigns one of them to process the event. The L2PU figures out the ROBs corresponding to the given RoI and requests data only from the involved ROSs. Since each ROB is connected to a specific ROD the identification of the involved ROBs is fast, because the mapping of $\eta-\phi$ regions and ROBs is fixed.

For accepted events only, the L2PU puts the decision details in the pseudo-ROS (pROS). In any case, the L2PU produces an accept/reject decision which is passed back to the L2SV, which forwards it to the only Data Flow Manager (DFM) in the system, which has the following role: if the decision is to reject the event, the DFM sends clear messages to the ROSs to free the involved buffer space; if the event is to be kept, the DFM assigns an SFI to assemble the event by requesting data from all participating ROSs and the pROS. Events are buffered in the SFI and made available to the EF upon request. The Event Building needs $\mathcal{O}(100)$ PCs, driven by the throughput requirements and by the fact that in steady-state conditions we want to use only 60-70% of the Gbit/s link into each SFI node.

3.2. Data Flow for the Event Filter

The ATLAS EF system is organized as a set of independent processor farms (sub-farms), connected to Sub-Farm Input (SFI) and Sub-Farm Output (SFO) elements via the Event Filter Network (see Fig. 1). Unlike the LVL2 system which involves many components to deal with the dataflow and the trigger aspects of the work, each EF node hosts both functionalities. Dataflow functionalities are provided by the Event Filter Dataflow process (EFD), while the processing tasks (PTs) are in charge of data processing and event selection. The EFD manages the communication with the SFI and SFO elements and makes the events available to the PTs via a memory mapped file, called the SharedHeap, which is used for local event storage and provides a safe event recovery mechanism in case of EFD crash. The PT cannot corrupt the event because it access the SharedHeap in read-only mode. PT problems are handled by the EFD which can identify PT

crashes and dead locks. In both cases, the EFD, which owns the event, can assign it to another PT or send it directly to the SFO. Inside the PT, the event selection algorithms produce a filtering decision and a selection object (used to classify the event and guide the off-line analysis steps) which are communicated back to the EFD. The filtering decision steers the EF dataflow and thus decides the fate of the event. Accepted events are sent to the SFO nodes, where a data-logger application streams and indexes the events into different files, according to each event's trigger path.

3.3. High Level Trigger

Both LVL2 and EF use online software for the control and data collection aspects, but the event processing and trigger algorithms are developed and tested in the ATLAS offline software environment ("Athena"). A common approach of the L2PU and the PT for the steering, seeding and sequential processing of the selection algorithms has been developed.

The dataflow and selection software are interfaced with a another software layer, common to LVL2 and the EF. This way, event selection algorithms developed offline can be "plugged" easily into the online HLT framework. For the EF this task is easier, because the less stringent requirements on the decision-latency ($\mathcal{O}(1)$ s compared to $\mathcal{O}(10)$ ms per event at LVL2), allow the straight-forward reuse of offline algorithms as configurable plug-ins. For LVL2 algorithms, the relatively long idle time between requests and arrival of RoI data ($\sim$ 10 to 20% of total [2]) allow resource sharing in multiple-CPU nodes. An L2PU could then deal with multiple "worker threads", each handling one event. Thus, algorithms which run at LVL2 should be designed to be thread-safe and efficient. For most algorithms this is not yet the case. The alternative (and our baseline solution) is to run multiple applications on each node.

4. STATUS AND OUTLOOK

We are in the process of assembling and commissioning the system described above, implemented with about 3000 commercial PCs hosted in 100 racks. All 153 ROS PCs (including spares) are installed and commissioned stand-alone, while 44 of them are also connected to the RODs of most of the barrel calorimeter and the LVL1 Central Trigger Processor. The system has been used to take cosmic data in the summer of 2006, with events built at the ROS level. A series of commissioning runs is envisaged, with a progressively larger fraction of ATLAS detectors and TDAQ components integrated.

For the event building, the first 32 SFIs, corresponding to 30% of the final system, have been installed and are in the commissioning phase. Along with them, 2 L2SVs, 1 pROS and 12 DFMs have also being installed. Having more DFMs operational will allow to run multiple, mutually-exclusive, partitions of ATLAS in parallel. This feature of the system will be used extensively during commissioning.

For the network, the complete cabling infrastructure is layed down now and is close to completion. The addition of switches takes place incrementally, according to needs.

The HLT will be completed last in order to take advantage of the ever increasing computational abilities of PCs. The rack space available to HLT corresponds to 2100 PCs. At the time of the TDR, we've assumed 16 GHz processing capacity out of each PC [3]. Since then, the industry has shifted from high clock rate processors to multi-core processors. Nevertheless, we have demonstrated in a dual-core dual-CPU PC, that the LVL2 decision rate scales linearly with the number of identical L2PU applications running on the node, till the resources (four cores in this case) are exhausted; additional L2PU applications do not increase the LVL2 rate any further [2]. Therefore, we believe that the multi-core multi-CPU technology will provide the necessary performance per PC, at the cost of higher memory needs and latency. Early in 2007 we will purchase and install the first 4 racks of HLT nodes ($\sim$ 120 PCs), which will be multi-core machines.

The functionality and performance of the system is verified in a number of test beds [4,5], with the most realistic of them having about 10% of the DataFlow system and 2% of the High Level Trigger system connected via the the final

network and infrastructure. Extensive tests are scheduled regularly to exercise the full chain as if it was an ATLAS run. Preloading the ROS with physics data-sets, we have recently integrated for the first time online electron/γ, muon, τ and jet algorithms.

Given the LHC schedule, the system-size described above should be able to largely satisfy the ATLAS needs in 2007. We are working on delivering a system which shows the proper scaling and matches the performance and robustness requirements of ATLAS towards the 14 TeV centre-of-mass collisions between protons by mid-2008.

REFERENCES

1. CERN S-LINK Homepage: http://hsi.web.cern.ch/HSI/s-link/
2. K. Kordas et al., ATLAS High Level Trigger Infrasturcture, RoI Collection and Event Building, CHEP06 conference, Mumbai, India, 13-17 Feb. 2006, ATL-DAQ-CONF-2007-002, 2007. Sumbitted to IEEE Trans. Nuclear Sci., (Available: http://cdsweb.cern.ch/record/934208)
3. The ATLAS TDAQ Collaboration, ATLAS High-Level Trigger Data Acquisition and Controls Technical Design Report, CERN/LHCC/2003-022, 2003 (Available: http://cern.ch/atlas-proj-hltdaqdcs-tdr/).
4. G. Unel et al., Studies with the ATLAS Trigger and Data Acquisition pre-series setup, CHEP06 conference, Mumbai, India, 13-17 Feb. 2006, ATL-DAQ-CONF-2006-019, 2006 (Available: http://cdsweb.cern.ch/record/934462)
5. D. Burchart-Chromek et al, Testing on a large scale: Running the ATLAS Data Acquisition and High Level Trigger software on 700 PC nodes, CHEP06 conference, ibid., ATL-DAQ-CONF-2006-002, 2006 (Available: http://cdsweb.cern.ch/record/941077)