The ATLAS ReadOut System - performance with first data and perspective for the future

G. Crone[a], D. Della Volpe[b], B. Gorini[c], B. Green[d], M. Joos*[,c], G. Kieft[e], K. Kordas[f], A. Kugel[g], A. Misiejuk[d], N. Schroer[g], P. Teixeira-Dias[a], L. Tremblet[c], J. Vermeulen[e], F. Wickens[h], P. Werner[c]

[a] *University College London*
[b] *Universita and INFN, Napoli*
[c] *CERN*
[d] *Royal Holloway University of London*
[e] *Nikhef, Amsterdam*
[f] *University Bern*
[g] *Ruprecht-Karls-Universitaet Heidelberg*
[h] *Rutherford Appleton Laboratory*

Abstract

The ATLAS ReadOut System (ROS) receives data fragments from ∼1600 detector readout links, buffers them and provides them on demand to the second-level trigger or to the event building system. The ROS is implemented with ∼150 PCs. Each PC houses a few, typically 4, custom-built PCI boards (ROBIN) and a 4-port PCIe Gigabit Ethernet NIC. The PCs run a multi-threaded object-oriented application managing the requests for data retrieval and for data deletion coming through the NIC, and the collection and output of data from the ROBINs. At a nominal event fragment arrival rate of 75 kHz the ROS has to concurrently service up to approximately 20 kHz of data requests from the second-level trigger and up to 3.5 kHz of requests from event building nodes. The full system has been commissioned in 2007. Performance of the system in terms of stability and reliability, results of laboratory rate capability measurements and upgrade scenarios are discussed in this paper.

Key words:
ATLAS, Data Acquisition, ReadOut System

1. Introduction

The trigger of the ATLAS experiment [1] at the CERN Large Hadron Collider (LHC) consists of three stages. The first level, implemented with special-purpose hardware, reduces the average event rate from 40 MHz (the frequency at which proton bunches collide inside the ATLAS detector) to at maximum 75 kHz (upgradable to 100 kHz). Accepts from the first-level trigger (L1) cause transfer of the event data to the ReadOut System (ROS). L1 also provides pointers to Regions of Interest (RoIs) to the software based second-level trigger (L2). Event data buffered in the ROS are selectively read out by L2 on the basis of the RoI information. The expected rate of accepted events (including special events that are accepted by L2 without processing, e.g. for calibration) is 3 kHz at the startup of LHC, increasing later to 3.5 kHz. Accepted events are built by the Event Builder, after retrieving the full event data from the ROS, and forwarded to the Event Filter processor farm. There the last stage of event processing takes place. Events are accepted at a rate of 200 Hz and stored permanently for later offline analysis.

*Corresponding author. Tel.: +41 22 7672364
Email address: markus.joos@cern.ch (M. Joos)

The ROS receives event data from sub-detector specific special-purpose hardware, the ReadOut Drivers (RODs), via ∼1600 fibre optical links and communicates via Gigabit Ethernet (GBE) links with L2 and the Event Builder. For each event the data buffered in the ROS are deleted either after successful building or after an L2 reject has been issued, i.e. with an average rate equal to the L1 accept rate of up to 100 kHz. To reduce communication overheads discard commands are sent in groups of 100 to the ROS.

2. Implementation and commissioning of the ROS

Conceptually the ROS consists of two major building blocks: ROBIN cards [2] and ROS PCs. A ROBIN card is a custom-built PCI card with three fibre optic interfaces for receiving event data, utilizing the CERN S-Link protocol [3]. A paged 64 MB dual-ported buffer memory is associated with each input. An FPGA and a PowerPC processor manage the three buffer memories as well as I/O via the PCI interface or via the on-board GBE network port.

The ROS PC is a 4U high, rack-mountable PC. Its software controls the ROBIN cards and interfaces to other sub-systems of ATLAS such as L2 and the Event Builder,

the run control system and the operational and physics monitoring systems. The PC is based on the SuperMicro X6DHE-XB motherboard [4] equipped with a single 3.4 GHz Intel Xeon (Irwindale) CPU. The motherboard was chosen for its mix of PCI slots (six 64-bit PCI and one x4 PCIe), which allows for the installation of several ROBIN cards and of a 4-port GBE PCIe NIC (Silicom PEG-4 [5]). Other features of the PC are an IPMI 2.0 based Board Management Controller (BMC) and a triple redundant power supply.

A health monitoring system based on the Nagios monitoring application [6] has been implemented. It retrieves software parameters such as the status of the RAM disk or system memory usage and, via the IPMI interface, hardware parameters like temperatures, voltages and fan speeds. Abnormal values are reported to ROS experts by automatic e-mails so that corrective actions are possible before failure of the PC causes irrecoverable data loss.

3. Performance of the ROS in 2008 (in terms of stability and reliability)

In total 150 ROS PCs are installed in the ATLAS USA15 underground counting room since 2006/2007. The average age of the hardware is 2.7 years (April 2009). Since their installation the PCs and therefore also the ROBIN cards were powered for most of the time. The record of all hardware related problems observed since the installation of the first PC, shows that the annual failure rate of any given component is very low (typically below 1%). In addition the failure rate in 2008 was below that of 2006/2007, indicating that the entire system has reached a plateau of high reliability. Even though the system is used in a controlled environment (stable temperature, filtered air) aging effects will sooner or later lead to an increase of the failure rate. We have addressed this by the procurement of a relative large number of spare components and by regularly evaluating more modern PC components for their suitability as potential replacements.

In 2008 the ROS has been used predominantly for the acquisition of data at low rates on the basis of a cosmic-ray trigger. During special periods, however, simulated physics data were downloaded into the ROS PCs and replayed at full speed. And last but not least: the first data produced by a beam in the LHC were successfully collected. In 2008 the ROS system demonstrated its stability in three ways:

1. during an uninterrupted data taking run of more than 4 days not a single PC failed,
2. from August to October about 900 TB of data has been fed through the ROS without major problems,
3. high rate tests, with data pre-loaded in the ROS PCs, have indicated the ability of the ROS to meet most of the ATLAS performance requirements, and no short comings have been identified.

4. Performance of the ROS (in terms of event data handling)

The main requirements in terms of event data handling for a ROS PC with 4 ROBIN cards, as defined in [7], are:

1. reception, via 12 optical links in parallel, of event fragments of up to 400 32-bit words at a rate of 75 kHz, upgradable to 100 kHz. Freeing of buffer memory needs also to take place at this rate,
2. provision of data from 2 - 3 optical links (per ROS PC) at a rate of approximately 20 kHz to L2,
3. provision of full event data (of all 12 optical links) at a rate of 3 kHz (initially) to the Event Builder.

The installed ROS PCs were booted with an SMP kernel and with hyper-threading enabled in the BIOS, as this seemed most natural for the multi-threaded design of the application running on the PCs. Results obtained with two test setups under conditions comparable to those of the deployed system have shown that the target performance could only be reached for event fragment sizes of 100 32-bit words or smaller (Fig. 1) and that the necessary processing by the PC determines the performance. For these tests event data were generated in the ROBIN cards whenever space became free in the buffers. The measurement procedure consisted of varying the L2 request rate until a delete rate of 100 kHz was observed, while the Event Building rate was a constant fraction of the delete rate. Subsequent tests have shown that a much higher performance can be reached if a uni-processor kernel is used, hyper-threading is disabled in the BIOS, interrupt coalescence is optimized for the network driver, and an optimized configuration of SELinux [8] is used. These measures contribute roughly equally and result in an improvement of the performance that now meets all requirements (Fig. 1).

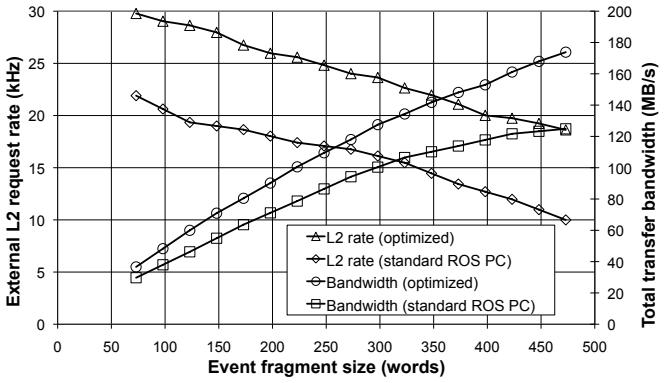

Figure 1: Maximum L2 request rate and total transfer bandwidth as a function of fragment size for a standard ROS PC with 4 ROBIN cards and two GBE connections, without and with optimizations as described in the text. The results are for a L1 rate (fragment arrival rate) of 100 kHz, an Event Building rate of 3 kHz and each L2 request retrieving data from 3 input links (chosen at random). Discard commands were received by the ROS PC and forwarded to the ROBIN cards in groups of 100. The TCP network protocol was used exclusively in this set of tests.

Nevertheless further improvements of the performance of the ROS may become necessary if ROS PCs need to handle higher L2 request rates and/or L2 requests for data from on average more than 3 input links, and to allow for additional bandwidth-demanding types of triggers, e.g. a b-physics trigger requesting all data from the inner detector or a missing E_T trigger requesting all data from the calorimeters.

The increased luminosity of LHC after the Phase 1 upgrade [9], foreseen for 2013, will result in increased data rates, but we believe that the current ROS architecture with upgraded rate capability can handle the expected rates. The next upgrade of the LHC (Phase 2, from 2018) will bring a further increase in luminosity, resulting in a data rate which is likely to be beyond the reach of the current ROS architecture.

5. Upgrade options for the ROS

The data handling performance of the ROS is currently limited by processing in the PC. Two possible approaches have been identified for increasing this performance:

1. replace the motherboard, CPU and memory by faster components. The third port of the NIC has to be connected, as two GBE links may no longer provide enough bandwidth for the data output by the PC,
2. use the on-board GBE interface of each ROBIN card for communication with L2 (using UDP), to reduce the load on the CPU of the ROS PC.

We have studied the effect of upgrading the PC hardware of the ROS PC, using a SuperMicro X7DB8-X motherboard [4] with two 2.66 GHz quad core Xeon processors and DDR2 667 MHz RAM, and observed a much higher performance in terms of throughput (Fig. 2). The maximum performance is obtained only if the ROS application runs exclusively on two cores connected to the same L2 cache. The synchronization between the various threads seems to cause too much overhead, resulting in a reduction of the rate capability, if the application runs on two cores connecting to different L2 caches or on more than two cores. The results also suggest that the increase in memory bandwidth removes the dependence of the maximum L2 rate on the fragment size observed for the standard ROS PC.

In the second approach the CPU of the PC has to deal only with the requests for data from the Event Builder and with delete requests, but the processors of the ROBIN cards are more heavily loaded, as a request-response cycle using the network port needs more processing time than a cycle using the PCI bus. With a test setup the validity of this approach has been demonstrated. The performance boundaries need still to be explored, but the first results indeed indicate, for the standard ROS PC, an increased throughput with respect to handling all input to and output from the ROBIN cards by the ROS PC.

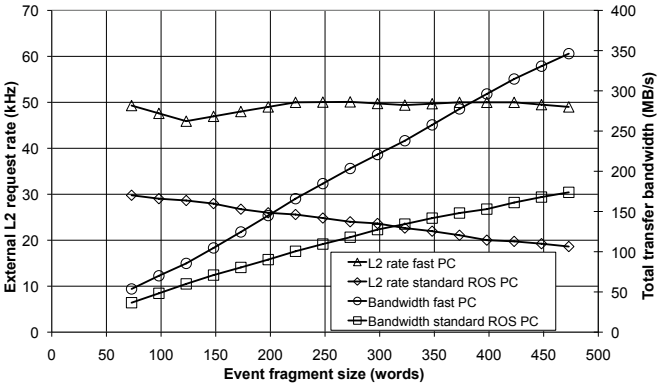

Figure 2: Maximum L2 request rate and total transfer bandwidth as a function of fragment size for a standard ROS PC and an upgraded ROS PC (see text) with 4 ROBIN cards and with two respectively three GBE connections. The measurement conditions were the same as for the results shown in Fig. 1.

Another concern for the operation of the ROS in its current form over the next decade is the availability of suitable motherboards with a sufficiently large number of PCI slots. The trend in industry is clearly away from parallel PCI towards PCIe. In order to address this, the development of a PCIe based ROBIN card has been started. This opportunity is also used to replace the PowerPC processor by a newer version with a higher clock speed. Even though it is not planned to replace all ROBINs, the PCIe based card promises a further increase of performance for those ROS PCs that have to handle a high L2 request rate.

6. Summary

The ROS does not only meet the performance requirements but also works very reliably. The motivations for a further increase in rate capability, several upgrade scenarios and results of first studies have been presented. These results show that the ROS has a clear upgrade path, making it possible to meet the requirements of ATLAS up to the LHC Phase 2 upgrade.

References

[1] The ATLAS Collaboration, G. Aad et al., The ATLAS Experiment at the CERN Large Hadron Collider, JINST 3 (2008) S08003.
[2] R, Cranfield et al., The ATLAS ROBIN, JINST 3 (2008) T01002.
[3] http://hsi.web.cern.ch/HSI/s-link/
[4] http://www.supermicro.com/
[5] http://www.silicom-usa.com/default.asp?contentID=609
[6] http://www.nagios.org/
[7] The ATLAS HLT, DAQ & DCS Technical Design Report, http://atlas-proj-hltdaqdcs-tdr.web.cern.ch/atlas-proj-hltdaqdcs-tdr/
[8] http://www.nsa.gov/research/selinux/
[9] F. Zimmerman, Presentation during February 2009 ATLAS Upgrade week, http://indico.cern.ch/materialDisplay.py?contribId=48& sessionId=27&materialId=slides&confId=45460