The ATLAS ReadOut System improved performance for the switchbased setup

N. Schroer[6], G. Crone[2], D. Della Volpe[3], B. Gorini[4], B. Green[1], M. Joos[4], G. Kieft[5], K. Kordas[8], A.Kugel[6], A. Misiejuk[1], P. TeixeiraDias[1], L. Tremblet[4], J. Vermeulen[5], P. Werner[4], F. Wickens[7]

[1]Royal Holloway University of London, [2]University College London, [3]Universita & INFN Napoli, [4]CERN, [5]Nikhef Amsterdam, [6]Ruprecht-Karls-Universität Heidelberg, [7]Rutherford Appleton Laboratory, [8]University Bern

Introduction

About 120 GB/s of data, generated by the detectors of the ATLAS experiment at CERN, are buffered in custom made circuit boards (ROBIN) plugged into the PCI slots of the ReadOut System (ROS) PCs. These PCs are connected via Gigabit Ethernet to the Data Collection (DC) network, which manages the selection and storage of events for later analysis. This baseline *bus-based scenario* has its main bottleneck in the network interface, as the CPU of the ROS PC which handles the network protocol can only manage the load of two GbE links. For use cases where many channels are requested at once this setup can not deliver sufficient performance.

The ROBINs have the potential to be directly connected to the DC network with their built-in GbE ports in the so called *switch-based scenario,* allowing to offload the message handling to their PowerPC (PPC). For this so far unused approach the firmware of the ROBIN needed to be modified in order to improve the performance of the network interface and to adapt the message handling to the demands of the DC network.

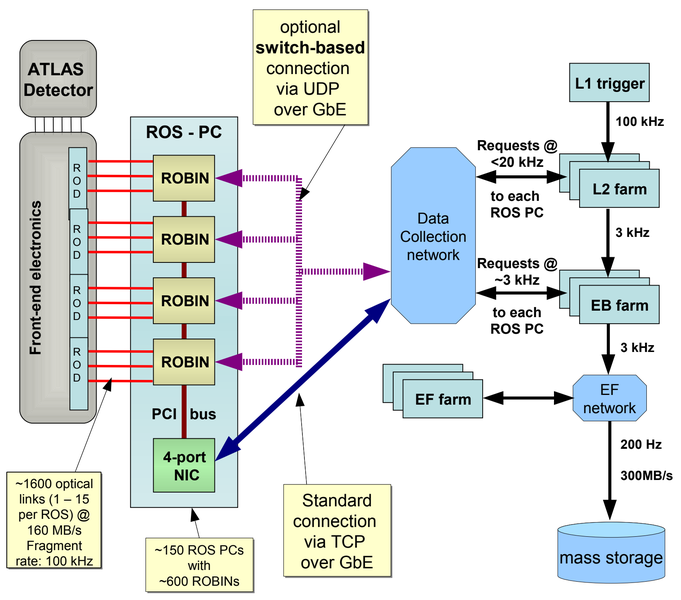


Figure 1: Overview of the ATLAS readout system

Modifications & test setup

On a ROBIN the two main components are the FPGA and the PPC. Their original firmware is fully functional, but the built-in network interface is not optimized for the communication with the DC network.

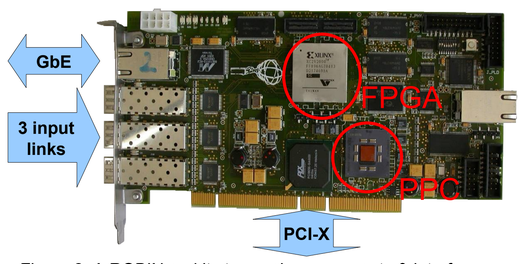


Figure 2: A ROBIN and its two main components & interfaces

The firmware of the PPC has been modified to respond to messages at the network interface in the same way as a ROS PC to allow the integration into the DC network. Data from the three input links can be bundled in one message to minimize overhead in the transmissions. Furthermore the possibility to use DMA for internal data handling is operational.

The latest FPGA firmware was modified by adding a second buffer to the transmission part of the network interface to speed up the data transfer. Performance for bigger fragment sizes is improved by adding the support of jumbo Ethernet frames of up to 6kB.

For the test setup the ROBIN is housed in a ROS PC and another PC is used to run a test program to simulate the DC network. This test program requests data fragments and sends delete messages to free the ROBIN buffers.

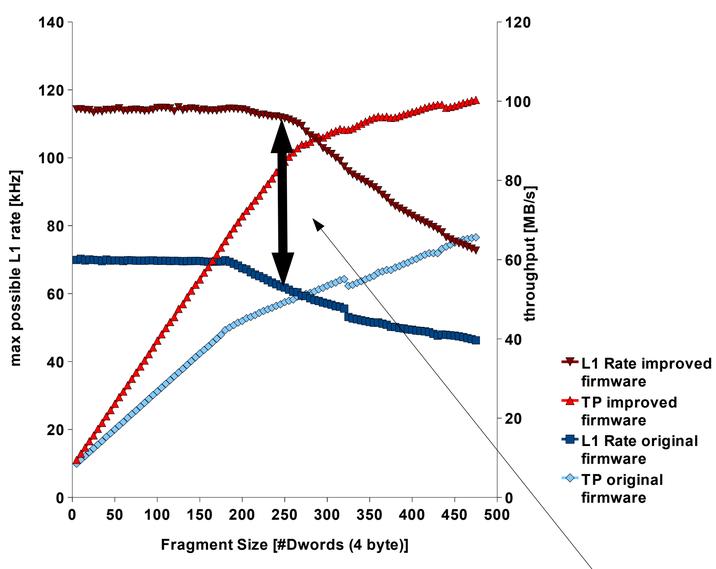
The goal is to be able to request fragments of about 1kB at a rate of 23 kHz via the network interface while the input rate (L1 rate) is 100 kHz, which corresponds to requesting 23% of the data.

Results

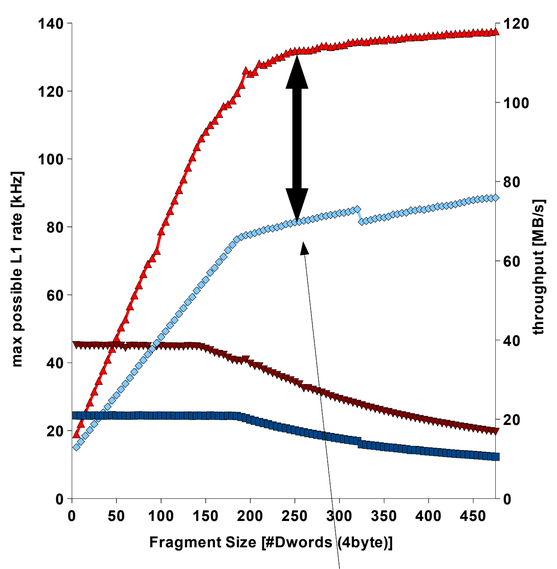
In the measurements fix percentages of fragments are requested for different fragment sizes and the throughput and the request frequency is measured. Hence the maximal possible L1 frequency is determined. Finally the request rate at a fixed L1 rate of 100 kHz is calculated from the data of many measurements. These are the most significant figures as this is a condition of the ATLAS experiment.

In the following graphs the canonical fragment size is marked with a vertical arrow, denoting the difference between the original firmware and the one after the modifications, both in *switch-based mode*.

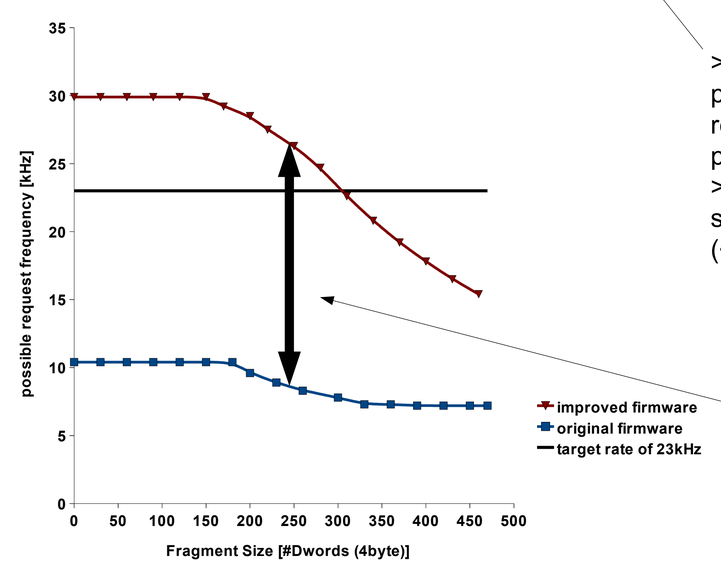
Graph 1 shows the increase in possible L1 rate at the desired request ratio, while in graph 2 the maximal throughput is compared. The last graph is the estimation of the maximum request rate for a fixed L1 rate of 100 kHz.



Graph 1: Measurement to determine possible L1 rate at target request ratio



Graph 2: Measurement to determine maximum throughput with 100% of fragment data requested



Graph 3: Calculated request rate for a fixed L1 rate of 100 kHz

>75% increase in possible L1 rate at fixed request ratio of 23%, possible L1 rate >100kHz for fragment sizes of up to ~1.25kB (~300 Dwords)

>50% increase in throughput, now reaching the limit of GigabitEthernet (~120 MB/s)

~3 times more request rate possible at a fixed L1 rate of 100 kHz

Conclusions

The modifications of the ROBIN firmware result in a significant performance increase of the network interface, making it possible to request event data of up to 1.25kB (~300 Dwords) per fragment with more than 23 kHz at a fixed L1 rate of 100 kHz, therefore the goal to be able to do that for fragments of 1 kB has been reached. Used in *switch-based mode* each ROBIN can provide more than half of the output data rate of a *bus-based ROS*, thus a typical ROS with 4 ROBINs can provide over twice the output of a *bus-based ROS*. This yields the potential to be applied for use cases with high L2 request rates or for trigger types which have additional bandwidth demands such as Inner Detector or Calorimeter full scans.

With the modification of the message handling of the network interface to the standard format used in the DC network, an integration into the system is fairly straightforward, although additional cabling is required as each ROBIN needs to be connected to a switch. This setup would be used only in those parts of the readout system with high demands, thus the amount of extra cabling and switches is modest. Small tradeoffs are that the switches need to be able to handle jumbo frames and that only UDP can be used to communicate directly with the ROBIN.

Even though the already almost complete utilization of the resources of the ROBIN prevents the implementation of the complex but more convenient TCP, the H/W design of our board together with reconfigurable components could be used to optimize the performance and implement alternative data transfer solutions.

For further information

Please contact *nicolai.schroer@ziti.uni-heidelberg.de*
or visit *www.ziti.uni-heidelberg.de*
More information on the ATLAS project can be obtained at *www.atlas.ch*
or more general at *www.cern.ch*