# Studies with the ATLAS Trigger and Data Acquisition "pre-series" Setup


N. G. Unel*, University of California at Irvine, US and CERN, Geneva, Switzerland

M. Abolins, G. Comune, Y. Ermoline, R. Hauser, B. Pope, Michigan State University, Dept. of Phys. & Astro., Michigan, US

P. Adragna, A. Dotti, C. Roda, Dipartimento di Fisica dell'Universita di Pisa e INFN Pisa, Italy

I. Alexandrov, V. Kotov, M. Mineev, JINR, Dubna, Russia

A. Amorim, N. Barros, LIFEP, Coimbra, Catolica-Figueira-da-Foz, Porto & Nova-de-Lisboa, Portugal

S. Armstrong, T. Maeno, M. LeVine, Brookhaven National Laboratory (BNL), Upton, New York, US

E. Badescu, M. Caprini, National Institute for Physics and Nuclear Engineering, Bucharest, Romania

J. Baines, V. Perera, F. Wickens, Rutherford Appleton Laboratory, Chilton, Didcot, UK

H.P. Beck, S. Gadomski, C. Haeberli, K. Pretzl, Laboratory for High Energy Physics, University of Bern, Switzerland

C. Bee, C. Meessen, F. Touchard, CPPM, IN2P3-CNRS-Universite d'Aix-Marseille 2, France

R. Blair, J. Dawson, G. Drake, W. Haberichter, J. Schlereth, Argonne National Laboratory, Argonne, Illinois, US

J.A. Bogaerts, D. Burckhart-Chromek, M. Ciobotaru, A. Corso-Radu, M. Dobson, N. Ellis, D. Francis, S. Gameiro,

B. Gorini, S. Haas, J. Haller, S. Hillier, A. Hoecker, M. Joos, A. Kazarov, L. Leahu, M. Leahu, G. Lehmann Miotto, L. Mapelli,

B. Martin, J. Masik, R. McLaren, G. Mornacchi, R. Garcia Murillo, C. Meirosu, T. Pauly, C. Padilla, J. Petersen, M. de Albuquerqu Portes,

D. Prigent, J. Sloper, I. Soloviev, R. Spiwoks, L. Tremblet, P. Werner, M. Wiesmann, CERN, Geneva, Switzerland

T. Bold, Faculty of Physics & Nuclear Techniques AGH-University of Science & Technology, Cracaw, Poland

M. Bosman, H. Garitaonandia, E. Sole Segura, S. Sushkov, IFAE, Universidad Autonoma de Barcelona, Barcelona, Spain

C. Caramarcu, National Institute for Physics and Nuclear Engineering "Horia Hulubei", Bucarest, Romania

R. Cranfield, G. Crone, Department of Physics and Astronomy, University College London, London, UK

M. Della Pietra, Dipartimento di Fisica dell'Universita degli studi di Napoli 'Federico II' e INFN, Napoli, Italy

A. Di Mattia, S. Falciano, E. Pasqualucci, Dipartmento di Fisica dell'Universita di Roma I 'La Sapienza' e INFN, Roma, Italy

A. Dos Anjos, W. Wiedenmann, H. Zobernig, Department of Physics, University of Wisconsin, Madison, Wisconsin, US

E. Ertorer, Dept. of Phys., Ankara University, Ankara, Turkey & CERN, Geneva, Switzerland

R. Ferrari, G. Gaudio, W. Vandelli, Departimento di Fisica Nucleare e Teorica dell'Universita di Pavia e INFN, Pavia, Italy

M.L. Ferrer, K. Kordas, W. Liu, INFN, Frascati, Italy

S. George, A. Misiejuk, B. Green, J. Strong, P. Teixeira-Dias, Dept. of Phys., Royal Holloway & Bedford New College, University of London, UK

A. Gesualdi Mello, M. Seixas, R. Torres, Universidade Federal do Rio de Janeiro, COPPE/EE, Rio de Janeiro, Brazil

H. Hadavand, Department of Physics, Southern Methodist University, Dallas, Texas, US

J. Hansen, Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark

R. Hughes-Jones, T. Wengler, Department of Physics and Astronomy, University of Manchester, Manchester, UK

S. Klous, G. Kieft, J. Vermeulen, NIKHEF, Amsterdam, Netherlands

T. Kohno, Physics department, Keeble Road, Oxford, UK (Now CERN, Geneva, Switzerland)

S. Kolos, A. Lankford, A. Negri, S. Stancu, S. Wheeler, University of California, Irvine, US

K. Korcyl, T. Szymocha, The Henryk Niewodniczanski Institute of Nuclear Physics, Polish Academy of Sciences, Cracow, Poland

A. Kugel, R. Manner, M. Muller, M. Yu, Lehrstuhl fur Informatik V, Universitut Mannheim, Mannheim, Germany

M. Landon, Physics Department, Queen Mary, University of London, London, UK

P. Morettini, C. Schiavi, Dipartimento di Fisica dell'Universita di Genova e INFN, Genoa, Italy

Y. Nagasaka, Hiroshima Institute of Technology, Hiroshima, Japan

Y. Ryabov, Petersburg Nuclear Physics Institute, Petersburg, Russia

D. Salvatore, F. Zema, Dipartimento di Fisica, Univ. della Calabria & INFN Cosenza, Italy

I. Scholtes, University of Trier, Germany

S. Tapprogge, R. Stamen, J. Van Wasen, Institut fur Physik, University of Mainz, Mainz, Germany

H. von der Schmitt, Max Planck Institut fur Physik, Germany

X. Wu, Section de Physique, Universite de Geneve, Switzerland

Y. Yasu, High Energy Accelerator Research Organization (KEK), Tsukuba, Japan

*Abstract*

The pre-series test bed is used to validate the technology and implementation choices by comparing the final ATLAS readout requirements, to the results of performance, functionality and stability studies. We show that all the components which are not running reconstruction algorithms match the final ATLAS requirements. For the others, we calculate the amount of time per event that could be allocated to run these not-yet-finalized algorithms. We also report on the experience gained during these studies while interfacing with a sub-detector for the first time at the experimental area.

## INTRODUCTION

The ATLAS experiment [1] at LHC will start taking data in 2007. As preparative work, a full vertical slice of the final higher level trigger and data acquisition (TDAQ) chain, "the pre-series", has been installed in the ATLAS experimental zone. In the pre-series setup, detector data are received by the readout system (ROS) and partially analyzed by the second level trigger (LVL2). On acceptance by LVL2, all data are passed through the Event Building (EB) to the Event Filter (EF) farms. Finally the selected events are written to mass storage. The details of the TDAQ design can be found elsewhere [2]. The layout of the ATLAS TDAQ architecture is illustrated in Fig. 1 where units from of all Trigger and DAQ applications are being put together in the experimental area. This article summarizes the performance and functionality studies which investigate the required number of application instances, the expected performance of various applications, network switch loads and alike in such a large and realistic test-bed.

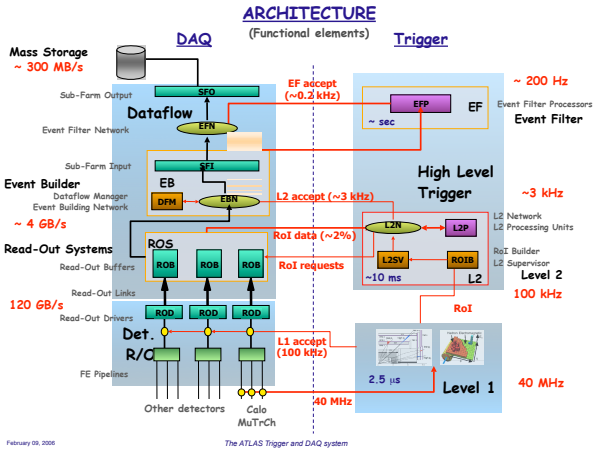


Figure 1: The ATLAS Trigger DAQ architecture.

*Pre-series Layout*

The pre-series setup is a complete vertical slice of the ATLAS TDAQ system to exercise the full functionality in a 10 % scale of the final system. The composition of the pre-series test bed and the estimations for the final system originating from the ATLAS TDR are given in Table 1. Although the final hardware specifications will be defined and the nodes will be bought in due time in accordance with the readout needs of the ATLAS detector in construction, the test bed components were selected as rack mountable, 1U high end PCs. Each node has at least two gigabit network connections: one for the control and monitoring operations and other for data transfer to EB, LVL2 or EF networks (see [3] for the details of network topology). The ROS nodes which need connection to both EB and LVL2 systems were equipped with a single 4-port network interface card on PCI bus giving 2 times redundancy for data transfer.

Table 1: Pre-series layout.

| Node | Preseries setup | Final setup, estimation |
|---|---|---|
| ROS | 12 | 150 |
| L2PU | 30 | 500 |
| L2SV | 2 | 10 |
| SFI | 6 | 100 |
| EF | 12 | 1500 |
| SFO | 2 | 30 |
| MON | 6 | n/a |

The ROS nodes, which receive up to 12 event fragments from different sections of ATLAS detector are also equipped with the custom made PCI cards (ROBINs) that will be used in the final system to receive and buffer these fragments. The event fragment input necessary for the studies can be preloaded into the ROBIN or internally generated by the ROBIN itself. Additionally, the ROS PC can emulate the ROBIN behavior which was shown to match the actual hardware performance within few percent of the LVL1 rate for any LVL2 accept ratio.

## THE EXPLOITATION

The requirements for the TDAQ system is that it maintains up to 100 kHz of LVL1 rate on its 1600 input links for event fragments with an average size of 1kB, and with subsequent reduction by LVL2 and EF systems, deliver full event data of about 1.5 MB to mass storage at a rate of about 200 Hz. This high reduction rate is achieved via a LVL2 farm in which each processing node has 10 ms to decide on an event based on a few percent of the full event. The design accept rate of 3.5 % imposes 3.5 kHz rate to the EF farm. Therefore the ROS nodes should be able to match the requests originating from both EB and LVL2, and also clear the event fragments which are either rejected or successfully assembled by the EB system. The rest of the section aims to show that the readout requirements are fulfilled, to estimate the number of nodes in various farms and to summarize various other studies such as run time stability and detector integration.

* gokhan.unel@cern.ch

### *EB and LVL2 Studies*

Initially the effect of the EB and LVL2 requests on the ROS were studied separately. Various parameters such as number of ROS nodes and event fragment size were varied to study the effect of the total event size on the EB throughput. The latter was found to be linearly dependent only on the number of SFIs (provided the gigabit link speed is not the limiting factor). The linear behavior allows the prediction of the number of SFIs required for the final ATLAS system, based on the estimated maximum EB throughput. As shown in Figure 2, when 70% of the gigabit bandwidth is utilized, the number of required SFIs is 80 nodes.

A similar configuration of the pre-series involving the ROS and LVL2 nodes allowed the measurement of the time necessary to retrieve Region of Interest (ROI) information from the ROS nodes. The average number of Read out links (ROL) per ROI request determined by modelling is 2 channels out of 12 in each ROS. The data retrieval time for this case is measured to be about 35 $\mu$s. For 6 channels the retrieval time increases to about 70 $\mu$s. Even in this case, which is expected to occur with a probability of 5 % on only a few ROS nodes out of the 132 contributing to ROI mechanism, the data retrieval time is less than 1 % of the allocated 10 ms per event. Other studies involving the ROI builder hardware were also performed showing the egalitarian trigger assignment amongst LVL2 supervisor nodes. Each supervisor is shown to sustain up to 35 kHz of LVL1 rate, confirming the initial estimation of 10 such nodes for the final system as being adequate to deal with the maximum 100 kHz LVL1 rate.

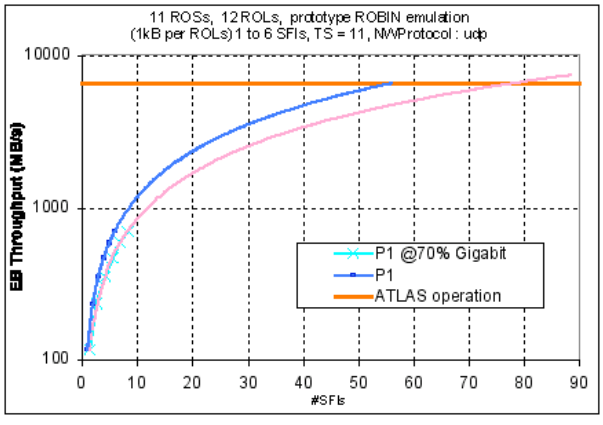


Figure 2: SFI estimate from test bed.

### *Combined Studies and Modeling*

In a configuration of 8 ROSs, 8 SFIs and up to 20 L2PUs, the triggers were generated as fast as possible to stress test the readout chain. Each L2PU only requested event fragments, i.e. no selection software was executed, from a randomly selected ROS. In this configuration each L2PU requests event fragments many times faster than the final ATLAS system, therefore making each L2PU resemble a multi-node L2PU farm behind a concentrator switch, e.g. at 3.5% each of the 14 L2PU requests event fragments approximately 44 times faster than the final system. Figure 3 shows the EB rate for different LVL2 accept ratios. It should be noted that even in this configuration of an overdriven system, the performance reaches a plateau and remains stable. The stability is ensured by respecting equation 1 which correlates the number of various TDAQ configuration parameters and the number of TDAQ applications:

$$\frac{TS \times N_{\mathrm{SFI}}}{WT \times N_{\mathrm{L2PU}}} = a \times N_{\mathrm{ROS}} \ , \qquad (1)$$

where $TS$ is the number of requests for event fragments issued by an SFI node at any single instance, $WT$ is the number of events being processed in parallel by a LVL2 farm node and $a$ is the LVL2 accept ratio. The solid line

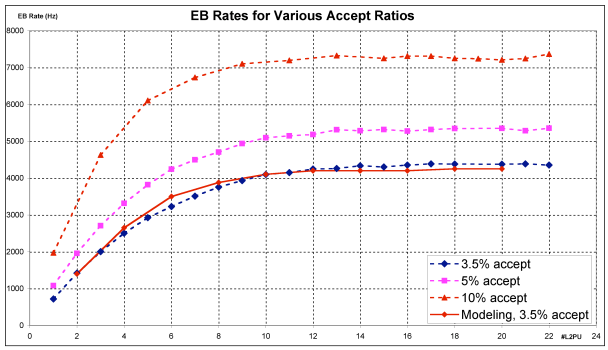


Figure 3: combined measurement results

in Figure 3 shows the results of a discrete event simulation of this pre-series configuration and for 3.5% LVL2 accept ratio. As can be seen there is good agreement between the results of the modeling and the measurements made on the pre-series. This agreement and other comparisons between the results of measurements on the pre-series and the discrete event simulation are used to validate the simulations of the components and subsequently simulate the full size ATLAS TDAQ system. The modeling results have confirmed of the ability of the foreseen final system to meet the ATLAS TDAQ requirements and allowed, for example, the event building rate, latencies of the various stages, buffer occupancies of the network switches to be studied. For example, on the central network switch ports to the SFIs, which are the busiest ones, the model finds that 60% of the time the queue was empty and only 1.4% of the time it is 1/3 of its maximum length and therefore proves that the final ATLAS network will be able to cope with the requirements. The lack of event selection algorithms in the modeling makes it a worst case study since they will slow down the LVL2 farms and lighten the overall network load.

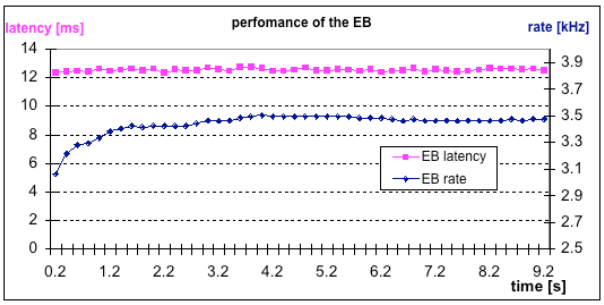


Figure 4: modeling results

### *Addition of Event Filter*

In the results presented in the previous sections, the SFI application has not performed any output to the EF. When the output to EF is performed, the maximum input bandwidth per SFI decreases by approximately 20% giving an efficiency figure of about 80%. Therefore the previous estimation of 80 nodes has to be corrected for the output efficiency yielding a total of 100 node estimation for the final ATLAS EB system.

The events (approximately 1.5 MB event size) selected by the HLT are stored at a rate of 200 Hz and for a maximum of 24 hours to accommodate possible failures of the connection to the CERN computer center. This functionality is performed by 30 SFO nodes of 1U height, each equipped with approximately 1 TB of disk space. The growing disk capacity in the market allowed consideration of raid options which provide protection against disk failures while keeping the total available disk capacity about the same. Table 2 shows the results of studying different Raid options. The 0.5 to 1.5 MB events were generated on the fly and written to different Raid partitions. The filesystem of choice was Linux ext2; the Raid1 exercise was performed with 2 disks and the Raid5 tests were performed with 3 disks. Results show that 6 nodes of 3U height, each with approximately 5.2 TB of Raided disk would match the requirements. Such an implementation would reduce both the space and cooling power required by the SFO farm and also reduce the number of nodes to be maintained.

Table 2: Writing speed in MB/s for SFO nodes.

| **RaidType** | **1U sw** | **1U hw** | **3U hw** |
|---|---|---|---|
| Raid1 | 44 | 48 | 51 |
| Raid5 | 53 | 73 | 73* |

* with 6 disks, speed increases to 93 MB/s

### *Integration with a Detector*

It has been possible to integrate the pre-series with the barrel of the ATLAS Hadronic calorimeter. This integration has allowed configuration and control issues of a detector by the TDAQ to be performed in addition exercising a complete ATLAS slice: the reception of the triggers by the ROIB hardware; readout of the detector electronics; reception of event fragments by 2 ROSs (16 ROLs in total); ROI collection by the LVL2 trigger; event building; EF selection; central data recording in CERN computer center. Although the HLT algorithms were not used during this exercise, it helped understanding the compatibility issues around the event format, trigger system and state transition synchronization.

### *Stability and Monitoring Issues*

Systematic failures have been forced in both hardware and software to check the system recovery possibilities on the pre-series test bed. In particular studies of the Control applications have shown that failures can be recovered at run time. In addition, all applications (except the ROS) associated to the movement of the data can be recovered and reintegrated into a running system. As the ROS requires hardware handshake with the ATLAS detector, it requires a reconfiguration for reintegration into a run. For hardware failures, the current implementation does not allow a reintegration after reboot, however as the Process Manager will be part of Unix services in the next implementation, it will be possible to reintegrate a dead node into the TDAQ chain.

The monitoring possibilities were also investigated in a configuration in which only EB was performed by running monitoring samplers on the ROSs and SFIs simultaneously and separately, and sending monitoring data over the control network to the monitoring nodes. For both ROS and SFI applications, measurements show that up to 3 % of operation rate can be used for sampling events for the purpose of monitoring without having any impact on the performance.

### *Replaying Physics Data*

Simulated physics data were preloaded into both ROS emulators and ROIB to be processed through the TDAQ chain up to mass storage, using the current implementations of the HLT selection algorithms. In addition to testing the functionality of the whole readout, this exercise allows the measurement of the ROI collection and event selection algorithm times for different trigger objects in a realistic environment. The details of these studies can be found in [4].

## CONCLUSIONS

Results from the pre-series, a complete functional slice of the final ATLAS TDAQ system, have been presented. The results confirm the ability of the final system to meet the ATLAS requirements and have allowed further studies of various functions and parameters, for example: event building rate, latencies of the various stages, buffer occupancies of the network switches . Functionality studies such as error recovery, the determination of allowed monitoring rates were also performed to further understand the

behavior of the TDAQ under a realistic load. The first full readout chain from cosmic rays up to sending the full event to mass storage was also achieved using all the TDAQ components. The pre-series is the workbench for the deployment of ATLAS TDAQ up to the commissioning of the final TDAQ system. As such continuous studies of software releases and configurations will take place during the installation and commissioning of the ATLAS TDAQ system. On completion of the final ATLAS TDAQ system, the pre-series will become the software and hardware validation platform prior to installation in the ATLAS TDAQ system.

# REFERENCES


[1] ATLAS Collaboration, ATLAS Technical Proposal, CERN/ LHHCC/94-43, LHCC/P2, CERN, Geneva, Switzerland, 1994.

[2] B. Gorini *et al.*, The ATLAS Data acquisition and High-Level Trigger: concept, design and status, CHEP06, Mumbai, India, 2006.

[3] S. Stancu *et al.*, "Networks for ATLAS Trigger and Data Acquisition", CHEP06, Mumbai, India, 2006.

[4] K. Kordas *et al.*, "ATLAS High Level Trigger Infrastructure, RoI Collection and EventBuilding" CHEP06, Mumbai, India, 2006.