THE ATLAS DATA ACQUISITION AND HIGH-LEVEL TRIGGER: CONCEPT, DESIGN AND STATUS

B. Gorini, CERN, Geneva, Switzerland,
M. Abolins, G. Comune, Y. Ermoline, R. Hauser, B. Pope, Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan, US
I. Alexandrov, V. Kotov, M. Mineev, JINR, Dubna, Russia
A. Amorim, N. Barros, L. Vaz Gil Lopes, Laboratório de Instrumentação e Fisica Experimental de Particulas, Universities of: Lisbon, Coimbra, Católica-Figueira-da-Foz, Porto & Nova-de-Lisboa
I. Aracena, H.P. Beck, S. Gadomski, C. Haeberli, S. Kabana, V. Perez-Reale, K. Pretzl, E. Thomas, Laboratory for High Energy Physics, University of Bern, Switzerland
S. Armstrong, K. Cranmer, D. Damazio, M. LeVine, T. Maeno, Brookhaven National Laboratory (BNL), Upton, New York, US
E. Badescu, M. Caprini, C. Caramarcu, National Institute for Physics and Nuclear Engineering, Bucharest, Romania
J. Baines, D. Emeliyanov, J. Kirk, V. Perera, F. Wickens, M. Wielers, Rutherford Appleton Laboratory, Chilton, Didcot, UK
C. Bee, C. Meessen, F. Touchard, Centre de Physique des Particules de Marseille, IN2P3, CNRS et Université d'Aix-Marseille 2, France
M. Bellomo, R. Ferrari, G. Gaudio, A. Negri, D. Scannicchio, W. Vandelli, V. Vercesi, Dipartimento di Fisica Nucleare e Teorica, Università di Pavia e I.N.F.N., Pavia, Italy
M. Biglietti, G. Carlino, F. Conventi, M. Della Pietra, Dipartimento di Fisica, Università degli studi di Napoli `Federico II' e I.N.F.N., Napoli, Italy
R. Blair, J. Dawson, G. Drake, W. Haberichter, J. Schlereth, Argonne National Laboratory, Argonne, Illinois, US
J.A. Bogaerts, D. Burckhart-Chromek, M. Ciobotaru, P. Conde-Muino, A. Corso-Radu, R. Dobinson, M. Dobson, N. Ellis, E. Ertorer, D. Francis, S. Gameiro, S. Haas, J. Haller, A. Hoecker, M. Joos, A. Kazarov, L. Leahu, M. Leahu, G. Lehmann Miotto, L. Mapelli, B. Martin, J. Masik, R. McLaren, C. Meirosu, G. Mornacchi, R. Garcia Murillo, C. Padilla, T. Pauly, J. Petersen, M. de Albuquerqu Portes, D. Prigent, C. Santamarina, J. Sloper, I. Soloviev, R. Spiwoks, S. Stancu, Z. Tarem, L. Tremblet, N.G. Unel, P. Werner, M. Wiesmann, CERN, Geneva, Switzerland
T. Bold, Faculty of Physics & Nuclear Techniques AGH-University of Science & Technology, Cracaw, Poland
M. Bosman, P. Casado, H. Garitaonandia, C. Osuna, E. Sole Segura, S. Sushkov, Institut de Física d’Altes Energies (IFAE), Universidad Autónoma de Barcelona, Barcelona, Spain
B. Caron, R. Moore, J. Pinfold, R. Soluk, University of Alberta, Edmonton, Canada
G.Cataldi, E. Gorini, M. Primavera, S. Spagnolo, A. Ventura, Università degli Studi di Lecce e I.N.F.N., Lecce, Italy
R. Cranfield , G. Crone, N. Konstantinidis, M. Sutton, Department of Physics and Astronomy, University College London, London, UK
A. De Santo, S. George, R. Goncalo, B. Green, G. Kilvington, A. Lowe, T. McMahon, A. Misiejuk, J. Strong, P. Teixeira-Dias, Department of Physics, Royal Holloway and Bedford New College, University of London, Egham, UK
T. Del Prete, A. Dotti, C. Roda, G. Usai, Dipartimento di Fisica, Università di Pisa e I.N.F.N., Pisa, Italy
M. Diaz-Gomaz, O. Gaumer, X. Wu, Section de Physique, Université de Genève, Switzerland
A. Di Mattia, S. Falciano, L. Luminari, F. Marzano, A. Nisati, E. Pasqualucci, A. Sidoti,

Dipartimento di Fisica, Università di Roma I 'La Sapienza' e I.N.F.N., Roma, Italy
A. Dos Anjos, W. Wiedenmann, H. Zobernig, Department of Physics, University of Wisconsin, Madison, Wisconsin, US
M.L. Ferrer, K. Kordas, W. Liu, LNF, Frascati, Italy
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
S. Hillier, A. Watson, E. Woehrling, School of Physics and Astronomy, University of Birmingham, Birmingham, UK
R. Hughes-Jones, T. Wengler, Department of Physics and Astronomy, University of Manchester, Manchester, UK
A. Khomich, A. Kugel, R. Männer, M. Müller, M. Yu, Lehrstuhl für Informatik V, Universität Mannheim, Mannheim, Germany
G. Kieft, S. Klous, J. Vermeulen, NIKHEF, Amsterdam, Netherlands
T. Kohno, Physics department, Keeble Road, Oxford, UK (Now CERN, Geneva, Switzerland)
S. Kolos, A. Lankford, S. Wheeler, University of California, Irvine, US
A.Kootz, Fachbereich Physik, Bergische Universität Wuppertal, Germany
K. Korcyl, T. Szymocha, The Henryk Niewodniczanski Institute of Nuclear Physics, Polish Academy of Sciences, Cracow, Poland
M. Landon, Physics Department, Queen Mary, University of London, London, UK
P. Morettini, F. Parodi, C. Schiavi, Dipartimento di Fisica, Università di Genova e I.N.F.N., Genova, Italy
Y. Nagasaka, Hiroshima Institute of Technology, Hiroshima, Japan
N. Panikashvili, S. Tarem, Technion Israel Institute of Technology, Israel
C. Potter, P. Rheaum, S. Robertson, B. Vachon, A. Warburton, McGill University, Montreal, Canada
Y. Ryabov, Petersburg Nuclear Physics Institute, Petersburg, Russia
D. Salvatore, F. Zema, Dipartimento di Fisica, Università della Calabria e I.N.F.N., Cosenza, Italy
I. Scholtes, University of Trier, Germany
S. Sivoklokov, Institute of Nuclear Physics, Moscow State University, Moscow, Russia
R. Stamen, S. Tapprogge, J. Van Wasen, Institut für Physik, University of Mainz, Mainz, Germany
E. Stefanidis, University of Athens, Athens, Greece
H. von der Schmitt, Max Planck Institut für Physik, Germany
Y. Yasu, High Energy Accelerator Research Organization (KEK), Tsukuba, Japan

Abstract

The Trigger and Data Acquisition system (TDAQ) of the ATLAS experiment at the CERN Large Hadron Collider is based on a multi-level selection process and a hierarchical acquisition tree. The system, consisting of a combination of custom electronics and commercial products from the computing and telecommunication industry, is required to provide an online selection power of 10^5 and a total throughput in the range of Terabit/sec.

This paper introduces the basic system requirements and concepts, describes the architecture of the system, discusses the basic measurements supporting the validity of the design and reports on the actual status of construction and installation.

INTRODUCTION

The ATLAS experiment [1] is one of the four experiments aimed at studying high-energy particle interactions at the Large Hadron Collider (LHC), that is under construction at CERN in Geneva and is scheduled to start to operate in the year 2007. At present the different components of the ATLAS detector are being installed in the underground cavern and the commissioning process has started.

The ATLAS TDAQ has been designed to take maximum advantage of the physics nature of very high-energy hadron interactions. In particular, the Region-of-Interest (RoI) mechanism is used to minimise the amount of data needed to calculate the trigger decisions thus reducing the overall network data traffic considerably.

The selection and data acquisition software has been designed in-house, based on industrial technologies (such as CORBA, CLIPS and Oracle). Software releases are produced on a regular basis and exploited on a number of test beds as well as for detector data taking in test labs and test beams.

The final system will consist of a few thousands processors, interconnected by multi-layer Gbit Ethernet networks.

CONCEPTS AND DESIGN

The ATLAS TDAQ is based on three levels of online event selection. Figure 1 shows the different functional elements of the system and the expected event rate at each stage. One can see that the TDAQ system is logically divided into a fast first level trigger (Level 1), a High Level Trigger system (the next two selection stages) and a Dataflow system that comprises all the elements responsible for the temporary data storage and the movement of the data between the different processing nodes.

The first level trigger (L1) [2] provides an initial reduction of a factor ~10^3 of the event rate starting from the 40 MHz nominal bunch crossing rate of the LHC, based on information from the muon trigger chambers and on reduced-granularity information from the calorimeters.

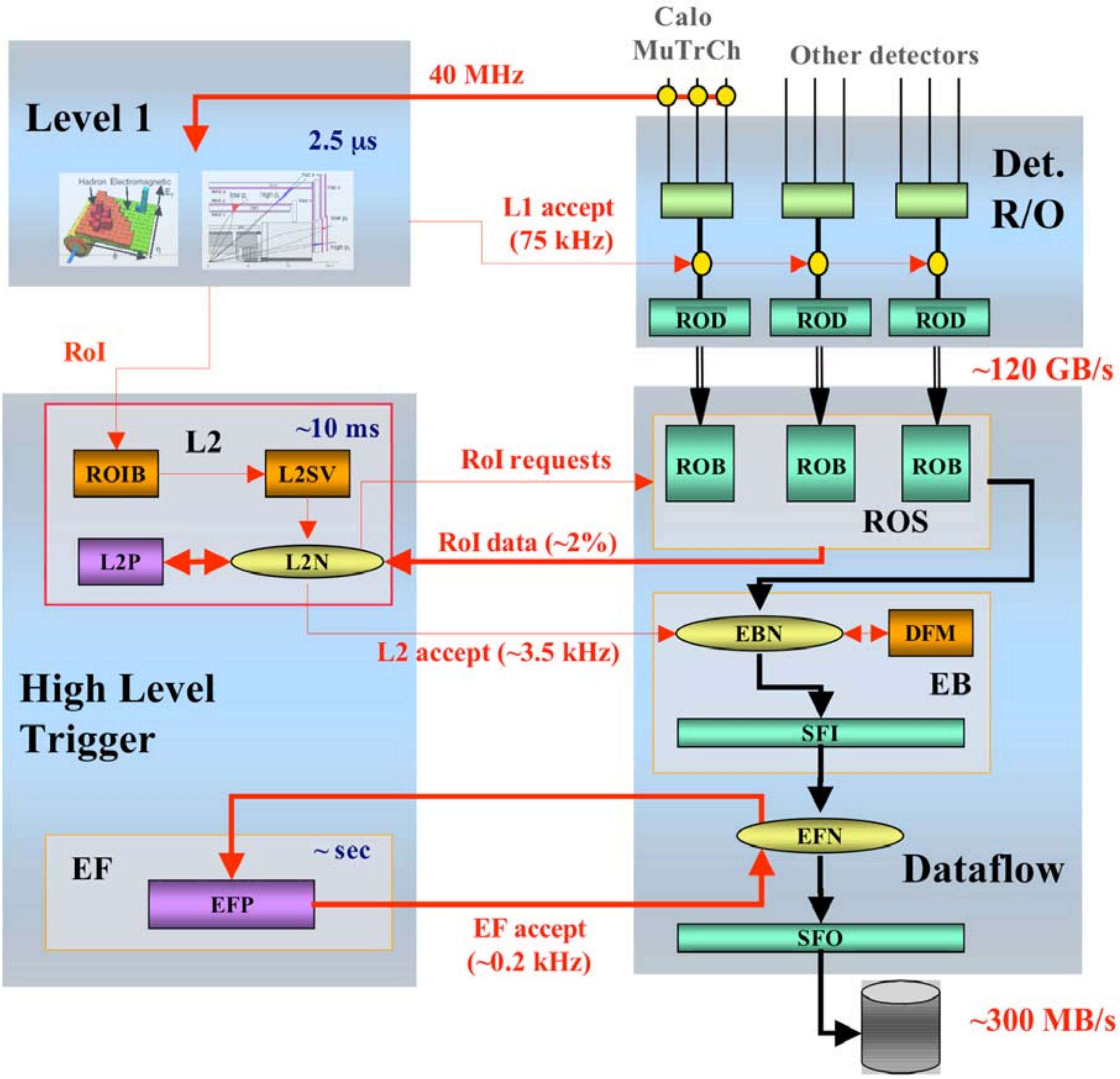

Figure 1: ATLAS TDAQ architecture. Thinner arrows indicate the flow of control messages, thicker ones indicate the flow of data fragments. The black arrows show the main data path.

During the latency of the L1 trigger selection algorithms (up to 2.5 μs), the complete event data is kept in the pipeline memories of the detector front-end electronics. Only the data for the events selected by the L1 trigger is then transferred from these front-end memories into the readout buffers (ROBs) contained in the readout system units (ROSs), where it is temporarily stored and provided on request to the following stages of event selection.

The data from the large number of detector readout channels is combined into ~1600 data fragments by the detector-specific readout drivers (RODs) and each of these fragments is sent for storage to an individual ROB.

The maximum rate of events accepted by the L1 trigger that can be handled by the ATLAS front-end systems is limited to 75kHz, but an upgrade to 100kHz is considered for a later phase. Trigger studies estimates the L1 rate required to meet the ATLAS physics program needs, to be about a factor two lower than this limit.

For every accepted event, the L1 system produces the “Region of Interest” (RoI) information, which includes the positions of all the identified interesting objects in units of pseudo-rapidity (η) and azimuthal angle (φ). This information is sent by the different elements of the L1 trigger system to the RoI builder (RoIB), which assembles it into a unique data fragment and sends it to a Level 2 supervisor (L2SV).

The L2SVs are the steering elements of the second trigger level (L2), which is designed to provide an additional factor 20-30 in reduction power with a latency of ~10 ms. The L2SVs receive the RoI information, assign the events to one of the processing units (L2PUs), and handle the results of the selection algorithms. The number of requested L2SVs scales with the L1 rate at which one runs the ATLAS experiment.

To provide the requested reduction power the L2PUs need to access detailed information from all the ATLAS detector elements (muon system, calorimeters and inner detector). To minimise the data transfers required at this early stage, the L2PUs retrieve only the few data fragments related to the geographical addresses of the interesting objects identified by the L1 (1-2 % of the total data volume). To do so it uses the RoI information received by the L2SV to identify and access only the few ROBs containing the relevant data fragments. A fast identification of the relevant ROBs is made possible by the fact that there is simple and fixed correspondence between the RoI regions and the ROBs, as each of them always receive data fragments from the same specific detector front-end modules.

The L2 system is really the most characteristic element of the ATLAS architecture, and provides detailed selection power before the full event-building and consequently reduces the overall dataflow power needs.

The results of the L2 algorithms are sent by the L2SVs to the dataflow manager (DFM), which assigns the accepted events to the event building nodes (SFIs) according to load-balancing criteria. The SFIs collect the data fragments related to any assigned event from all the ROBs and assemble them in a unique event fragment. The expected rate of events at this stage is ~3.5 kHz, that given a mean ATLAS event size of 1.6 Mbyte, corresponds to a total throughput of about 6 GByte/s out of the event building system.

The resulting complete event fragments are then sent to the event filter processors (EFPs) for the last selection stage, and the accepted events are finally sent to the output nodes (SFOs) to be permanently saved on mass storage. At this stage the rate of events is expected to be ~0.2 kHz i.e. more than a factor 10^5 lower than the original LHC bunch-crossing rate.

The DFM also manages the list of events that can be removed from the dataflow system, as they have either been rejected by the L2 or received by an EFP, and periodically sends to the ROBs the list of data fragments to be released.

SYSTEM IMPLEMENTATION

The ROBs are implemented into custom PCI cards (ROBINs) each housing 3 independent buffers. The ROBINs are itself installed into PCs each one corresponding to a ROS.

The connection between the RODs (detector specific) and the ROBs is implemented with point-to-point optical readout links (ROLs) conforming to the S-LINK specification [3] and providing individual data throughput of up to 160 MByte/s.

A ROS typically houses 4 ROBINs, for a total of 12 ROBs, and handles the data requests (from L2PUs and SFIs) for all of them through its network interfaces. Upon reception of a data request, the ROS application collects the relevant data fragments from the ROBIN modules through the PCI bus (from few selected ROBs for L2PU requests or from all of them for SFI requests), combines them into a unique ROS data fragment and sends it back to the requester. The total number of ROSs will be of ~150.

The existing RoIB prototype is implemented as a custom VMEbus system, receiving the individual RoI information fragments and sending the combined result to the L2SVs with the same point-to-point link technology as the one used for the ROLs. The performance of the communication protocol between the RoIB and the L2SVs and hence the maximum L1 rate that can be handled has been measured to scale linearly with the number of L2SVs.

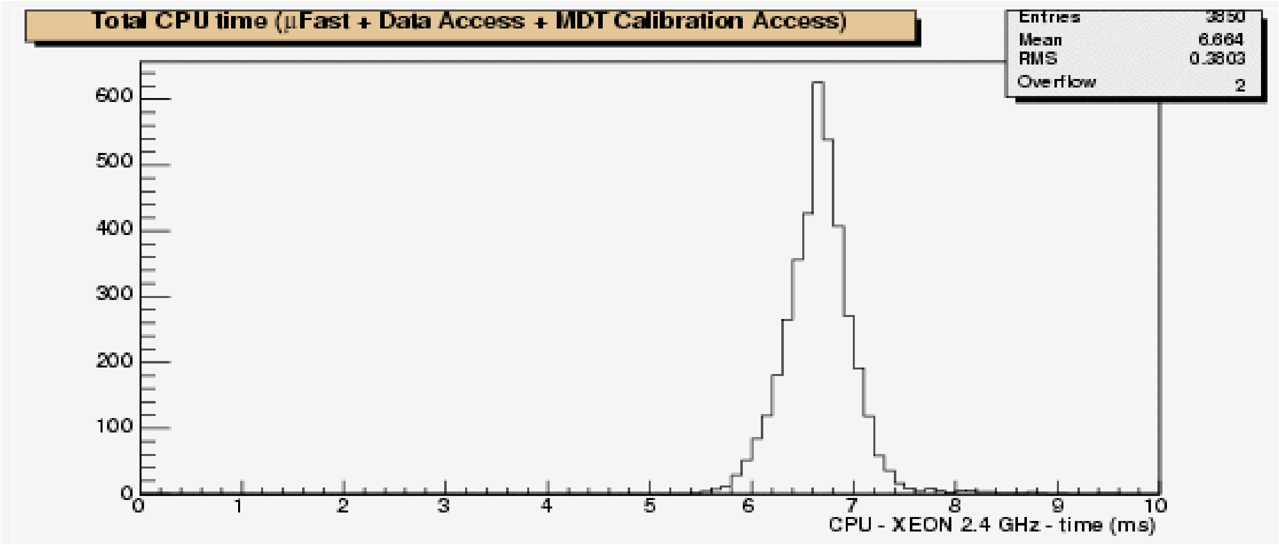

Figure 2: Muon reconstruction time at L2, on a 2.4 GHz XEON L2PU..

All the HLT and Dataflow nodes are implemented as multi-threaded C++ applications running on Linux PCs.

The various nodes are interconnected by multi-layer Gbit Ethernet networks and a custom message passing protocol has been developed to manage the data movements.

The size of the final system will be largely dominated by the number of processing nodes (L2PUs and EFPs).

The number of L2PUs is determined by the latency of the algorithms. For the nominal algorithm latency of 10 ms and a maximum L1 rate of 100 kHz, 1000 independent computing units will be needed. This required computing power will be provided by ~500 dual-CPU machines.

For the Event Filter stage we estimate a need of ~1600 EFP nodes. The Event Filter system is designed to be scalable. In the initial running phases few nodes will be deployed and the system will grow afterwards to cope with the increasing requirements. Some Event Filter clusters may even be deployed in remote institutes sites.

The number of SFI nodes is instead entirely determined by the event building throughput requirements, and by the design choice of never using any data network line to more than 75% of its capacity. A simple calculation shows that the final number of SFIs shall be ~100.

Table 1: Number TDAQ nodes required to handle the maximum ATLAS L1 rate

Application type	Number of nodes
ROS	~150
SFI	~100
L2PU	~500
EFP	~1600

A complex online software infrastructure has been developed to configure[4], control[5], and monitor[6] such a large system. Details on specific aspects can be found in papers presented to this conference [7], [8].

Coherent software releases containing both the dataflow applications and the infrastructure components, are produced several times per year.

Details on the network design and management, and on the global system management can be found in [9] and [10].

SYSTEM VALIDATION

All the components of the system have been implemented and tested in various testbeds and in a pre-series system installed in the ATLAS experimental cavern (see papers [11] and [12] presented to this conference) and results have been used to calibrate a detailed simulation of the final system. System prototypes have also been deployed as the main DAQ systems for the ATLAS test beams over the past few years.

The principal critical parameters have been studied in detail.

Figure 2 shows results from the measurements of basic L2 algorithm latencies (namely the muon reconstruction) with today's standard CPUs. Measurements on different algorithms show similar performances and indicate that reaching the required global L2 latency of ~10 ms requires a reasonable computing power increase. Studies have indicated that even if CPU clock speed doesn't seem to increase as quickly as originally expected, the required CPU performance will be provided on the proper time scale by multi-core machines (see [13]).

Another critical element of the architecture is the ability of the ROSs to handle the high rate of data requests from L2PUs and SFIs. It is important to point out here that the various ROSs will receive a much different rate of requests from the L2PUs depending on the detector and the extension of the η,φ region covered by the front-end modules to which they are connected. Hence very few ROSs (2-4) may become limiting factors for the system while the others will be largely under-utilized: in case of performance limitation one could hence provide few extra units to offload the few critical ones. Figure 3 shows the

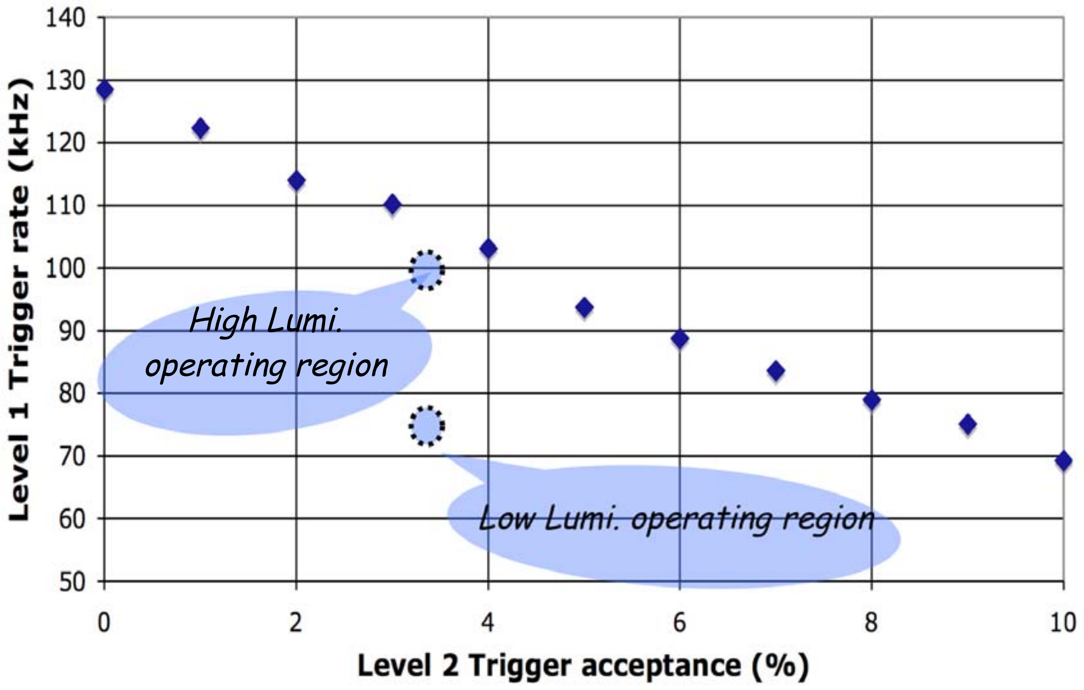

Figure 3: maximum performance achieved by the "worst-case" ROS as a function of the L2 event selection power, and compared with the ATLAS operating conditions for both high and low beam luminosity.

.

measurements of the maximum L1 rate sustainable by the worst-case ROS for different values of the L2 event selection power. The results are compared with the expected ATLAS operating conditions.

CONCLUSIONS

ATLAS has designed a trigger and data acquisition system with a three-level trigger hierarchy, based on the Region-of-Interest mechanism that provides an important reduction of data movement.

The architecture has been validated by deploying the system on different testbeds. and on ATLAS test beams and using a detailed modelling software to extrapolate the reduced scale results to the system full size.

Dataflow applications and protocols have been tested and perform according to specifications.

HLT software performance studies based on complete algorithms and realistic raw-data input indicate that our target processing times are realistic

The system design is complete and the installation has started. Part of the ROS system is already used by some ATLAS sub-detectors for their commissioning.

ACKNOWLEDGEMENTS

B. Gorini would also like to thank the local organising committee for the excellent organisation of the conference.

REFERENCES

[1] http://atlas.web.cern.ch/Atlas/index.html

[2]http://atlas.web.cern.ch/Atlas/GROUPS/DAQTRIG/LEVEL1/level1.html

[3] http://hsi.web.cern.ch/HSI/s-link/

[4]http://atlas.web.cern.ch/Atlas/GROUPS/DAQTRIG/CWG/index.html

[5]http://atlas.web.cern.ch/Atlas/GROUPS/DAQTRIG/ControlWG/Controls.html

[6] http://atlas-tdaq-monitoring.web.cern.ch/

[7] A. Kazarov et al., “A rule-based control and verification framework for ATLAS Trigger-DAQ”, CHEP06, Mumbai, India

[8] R. Murillo Garcia, G. Lehmann Miotto, “A log service package for the ATLAS TDAQ/DCS group”, CHEP06, Mumbai

[9] C. Meirosu et al. “Planning for predictable network performance in the ATLAS TDAQ”, CHEP06, Mumbai, India

[10] M. Dobson et al., “The architecture and administration of the ATLAS online computing system”, CHEP06, Mumbai, India

[11] D. Burckhart-Chromek et al., “Testing on a large scale: Running the Atlas Data Acquisition and High Level Trigger software on 700 pc nodes”, CHEP06, Mumbai, India

[12] N.G. Unel et al., “Studies with the ATLAS Trigger and Data Acquisition "pre-series" setup”, CHEP06, Mumbai, India

[13] K. Kostas et al., “ATLAS High Level Trigger Infrastructure, RoI Collection and EventBuilding”, CHEP06, Mumbai, India