PERFORMANCE OF THE ATLAS DAQ DATAFLOW SYSTEM

G. Unel[#], J.A. Bogaerts, M. Ciobotaru, B. DiGirolamo, R. Dobinson, E. Palencia Cortezon, D. Francis, S. Gameiro, P. Golonka, B. Gorini, M. Gruwé, S. Haas, M. Joos, G. Lehmann, T. Maeno, L. Mapelli, B. Martin, R. McLaren, C. Meirosu, G. Mornacchi, J. Petersen, D. Prigent, A. Corso-Radu, P. de Matos Lopes Pinto, R. Spiwoks, S. Stancu, L. Tremblet, P. Werner, CERN, Geneva, Switzerland
R. Blair, J. Dawson, J. Schlereth, Argonne National Laboratory, Argonne, Illinois, USA
M. LeVine, Brookhaven National Laboratory (BNL), Upton, New York, USA
E. Pasqualucci[+], Dept. di Fisica dell'Università di Roma I 'La Sapienza' e I.N.F.N., Roma, Italy
M. Shimojima, Dept. of Electrical Engineering, Nagasaki Institute of Applied Science, Nagasaki, Japan
H. Zobernig, Department of Physics, University of Wisconsin, Madison, Wisconsin, USA
R. Cranfield, G. Crone, Dept. of Physics and Astronomy, University College London, London, UK
B. Green, A. Misiejuk, J. Strong, Dept. of Physics, Royal Holloway and Bedford New College, University of London, Egham, UK
R. Ferrari, W. Vandelli, Dept. Fisica Nucleare e Teorica dell'Università di Pavia e INFN, Pavia,Italy
R. Hughes-Jones, Dept. of Physics and Astronomy, University of Manchester, Manchester, UK
Y. Hasegawa, Department of Physics, Faculty of Science, Shinshu University, Matsumoto, Japan
Y. Nagasaka, Hiroshima Institute of Technology, Hiroshima, Japan
K. Nakayoshi, Y. Yasu, KEK, High Energy Accelerator Research Organisation, Tsukuba, Japan
M. Beretta, M.L. Ferrer, Laboratori Nazionali di Frascati dell' I.N.FN., Fracasti, Italy
H.P. Beck, C. Haeberli, S. Gadomski[*], V. Perez Reale, Laboratory for High Energy Physics, University of Bern, Switzerland
A. Kugel, M. Müller, C. Hinkelbein, M. Yu, Lehr stuhl für Informatik V, Universität Mannheim, Mannheim, Germany
M. Abolins, Y. Ermoline, R. Hauser, Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan, USA
G. Kieft, J. Vermeulen, NIKHEF, Amsterdam, The Netherlands
D. Botterill, F. Wickens, Rutherford Appleton Laboratory, Chilton, Didcot, UK
A. Kaczmarska, K. Korcyl, M. Zurek, The Henryk Niewodniczanski Institute of Nuclear Physics, Polish Academy of Sciences, Cracow, Poland
A. Lankford, R. Mommsen, University of California, Irvine, California, USA
A. Dos Anjos, M. Losada Maia, Universidade Federal do Rio de Janeiro, COPPE/EE, Rio de Janeiro, Brazil

Abstract

The baseline DAQ architecture of the ATLAS Experiment at LHC is introduced and its present implementation and the performance of the DAQ components as measured in a laboratory environment are summarized. It will be shown that the discrete event simulation model of the DAQ system, tuned using these measurements, does predict the behaviour of the prototype configurations well, after which, predictions for the final ATLAS system are presented. With the currently available hardware and software, a system using ~140 ROSs with 3GHz single cpu, ~100 SFIs with dual 2.4 GHz cpu and ~500 L2PUs with dual 3.06 GHz cpu can achieve the dataflow for 100 kHz Level 1 rate, with 97% reduction at Level 2 and 3 kHz event building rate.

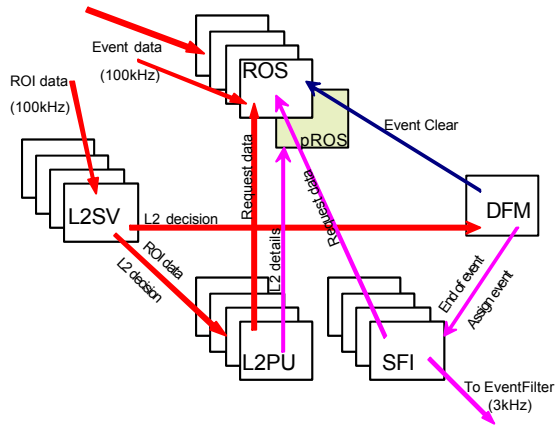
ATLAS DATAFLOW SYSTEM

The 40 MHz collision rate at the LHC produces about 25 interactions per bunch crossing, resulting in terabytes of data per second, which has to be handled by the detector electronics and the trigger and DAQ system [1]. A Level1 (L1) trigger system based on custom electronics will reduce the event rate to 75 kHz (upgradeable to 100 kHz – this paper uses the more demanding 100 kHz). The

#. Also affiliated with University of California at Irvine, Irvine, USA
*. On leave from Henryk Niewodniczanski Institute of Nucl. Physics, Cracow
+. Presently at CERN, Geneva, Switzerland

DAQ system is responsible for: the readout of the detector specific electronics via 1630 point to point read-out links (ROL) hosted by Readout Subsystems (ROS), the collection and provision of "Region of Interest data" (ROI) to the Level2 (L2) trigger, the building of events accepted by the L2 trigger and their subsequent input to the Event Filter (EF) system where they are subject to further selection criteria. The DAQ also provides the functionality for the configuration, control, information exchange and monitoring of the whole ATLAS detector readout [2]. The applications in the DAQ software dealing with the flow of event and monitoring data as well as the trigger information are called "DataFlow" applications. The DataFlow applications up to the EF input and their interactions are shown in Figure 1.

Figure 1 ATLAS DAQ-DataFlow applications and their interactions (up to the EventFilter)

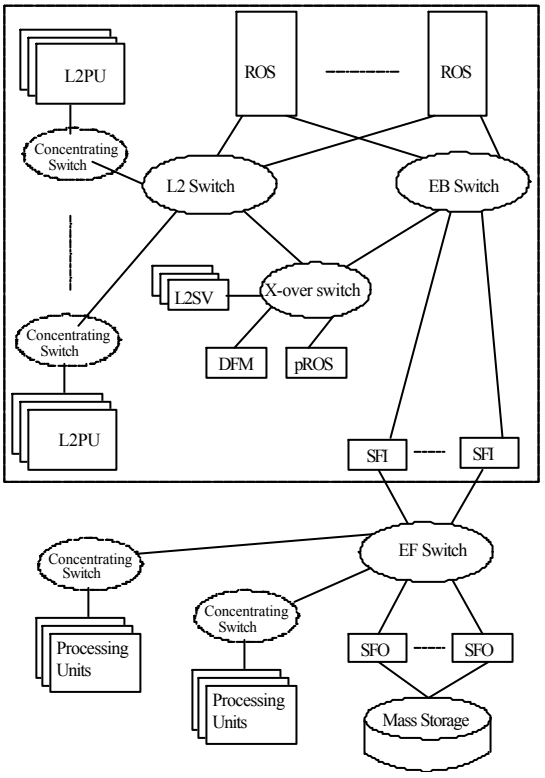


Once the L1 accept signal (L1A) is received, the event fragments from the detectors are sent from the detector specific electronics into the readout buffers (ROB) hosted in the ROS units. Meanwhile, the trigger data that caused the L1A is sent to a L2 supervisor (L2SV) in the form of ROI information. The L2SV assigns a L2 processing unit (L2PU) the task of processing the event and returning with an 'accept' or 'reject' decision which is forwarded to the Dataflow Manager (DFM). In the final system, the 100kHz of L1 rate has to be shared between all the L2SVs. If one assumes having about 10 supervisors, the requirement on a L2SV is to operate at a rate of 10kHz. The L2PU requests the related event fragments from the implicated ROS(s), runs a selection algorithm on the acquired event fragments and replies to L2SV with a decision message for that particular event. For accepted events, the L2PU sends meta-data about the decision process details to a dedicated ROS (pROS). The requirements for L2PUs are dictated by the physics requirements and the complexity of the L2 selection algorithms that are not yet finalized. Therefore, the goal is to minimize the data collection contribution to the total time budget, therefore to allow more complicated L2 selection algorithms during the same time.

The DFM receives the L2 decision and broadcasts the list of rejected events to the ROS units and assigns the accepted events to a Subfarm Input (SFI) for full event building. It is foreseen to have only one DFM in the final system given the non-stringent requirements on its performance. The SFI receives the identifier of the accepted event from the DFM and collects all the fragments from all the ROS nodes. After merging all the fragments, it sends the event to the Event Filter (EF) network for further rate reduction. It is required to have enough SFIs to do full size event building at a rate of 3-3.5kHz using the gigabit line throughput at an acceptable level of 60-70 % to avoid congestion.

The ROS hosts the input channels receiving data from various subdetectors. It responds both to L2 and EB requests and also allows local monitoring of the event data. The baseline selection for the start of the ATLAS experiment uses a custom made PCI card to host 3 input channels and an auxiliary Gigabit output channel for possible future upgrades. Using PCs hosting a maximum of 4 such cards, the 1630 ROLs coming from the detectors imply a minimum of 140 ROS nodes. The requirements for a ROS at the full LHC luminosity are to sustain event fragment input (and clear) at L1 rate, to deploy ROI fragments to the L2 network at a rate up to 20 kHz and ROS fragments from all its input channels to the EB network at a rate of 3-3.5 kHz in addition to local monitoring at a low rate. The requirement on the pROS is to sustain only 6-7 kHz (both EB and L2) of few 100 Byte, since it is only involved for accepted events.

Figure 2 The prototype test setup (Switches shown in dashed lines are foreseen for the final setup.)



Layout of the Prototype setup

A prototype test setup, representing about 20% of the final ATLAS DAQ system has been put together for both functionality and performance studies to evaluate the needs for each component to match the readout requirements. Figure 2 shows a schematic view of the prototype setup used for the measurements presented in this note. The setup was built around three gigabit switches and 63 PCs with Intel Xeon processors [3] and 64-bit/66MHz PCI busses, running Atlas DataFlow applications with Linux as the Operating System [4]. Table 1 summarizes the PC hardware and DAQ applications they were running. The Foundry FastIron 800, 64 port Gigabit switch [5] hosted both the EB and the EF network (ROS, SFI, SFO), whereas the Batm T6, 31 port Gigabit switch [6] hosted the L2 network (ROS and L2PU). A major difference from the final system was the direct connection of L2PUs to the L2 network, without the concentrating switches. The Foundry EdgeIron switch hosted the L2SVs, DFM and pROS and acted as a cross over connection between the L2 and EB networks. UDP was the network protocol of choice throughout this note, although TCP and RAW sockets were also implemented and tested. For the final system, the event fragments will be sent using UDP for performance, scalability and fine-tuning considerations, however the control messages (such as an event assignment) will be distributed using TCP. Since the custom hardware to receive data from the subdetectors into the ROBs is still in the development stage, the ROSs were configured to perform as though they were equipped with 12 ROLs with 1 KByte event fragments each.

To assess scalability issues, discrete-event models of all DataFlow components were developed. Individual components (DataFlow application software, nodes and switches) were calibrated in dedicated setups [7]. Calibration data were obtained either by putting components in question under heavy load and recording maximal rate or by instrumenting the application software with time stamps giving insight on time spent in various places of the application. Calibrated models were then compared to measurements from the prototype setup resulting in good agreement within few percent.

Table 1: PC Hardware and DAQ applications

amount x type	**DAQ Application**	**NIC Manufacturer**	**FSB** (MHz)
16x2.4GHz SMP,rackmount	SFI	Intel-e1000	533
14x3.06GHz SMP,desktop	L2PU, ROS*	Intel-e1000	533
33x2.00GHz UP,desktop	L2SV, pROS,ROS	Intel-e1000 Broadcom	400

** Only 5 were used as ROS for small-scale performance studies*

MEASUREMENTS ON INDIVIDUAL COMPONENTS

Out of the DataFlow system components presented in section 1, the performances of the DFM, L2SV, and pROS have been investigated and presented elsewhere [8]. For the rest, the performances of the EB and L2 systems were studied both separately and combined together. The goal of these studies is to prove that no component will be a bottleneck in the final ATLAS system. The various readout rates and their relations are defined via,

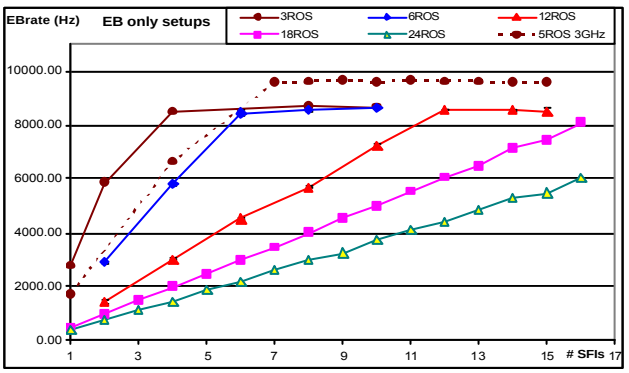
$$R_{EB} = R_{L1} \times ac; \quad R_{L2} = R_{L1} \times n_{ROI}/n_{ROS} \qquad (1)$$

where R_{L1} is the Level1 rate, R_{L2} is the ROI request rate per ROS and REB is the Event building rate; ac is the L2 acceptance fraction, n_{ROS} is the number of ROS units in the system and n_{ROI} is the number of requested ROI fragments per event.

The EB system: SFI performance measurement

In an EB only system, all triggers are internally generated by the DFM and events are built by the SFIs; therefore the EB rate is governed by either the ROS or the SFI performance depending on the configuration. Figure 3 shows the measured EB rate in Hz for systems sending 12.4 KBytes/ROS versus the number of SFIs. The solid curves are for 2 GHz ROS systems used to study scalability and functionality whereas the dashed curve is for a 3GHz ROS system used to study the performance. The same maximum rate reached by all 2 GHz ROS systems of different sizes shows that there is no extra latency introduced with growing event size. The number of SFIs limits the EB rate in the linear region. The EB rate in the plateau region is limited by the ROS performance. The ROS limitation for the solid curves is dictated by the cpu speed, since the throughput of 108 MBytes/s is less than the Gigabit bandwidth which can be reached by a 3GHz ROS (dashed curve). For these fast nodes, a special test program, benefiting from two separate Gigabit NICs, was utilized to obtain the maximum EB rate of about 14.5 kHz.

Figure 3 (a) EB rate for different event sizes (b) Scaling of the EB throughput



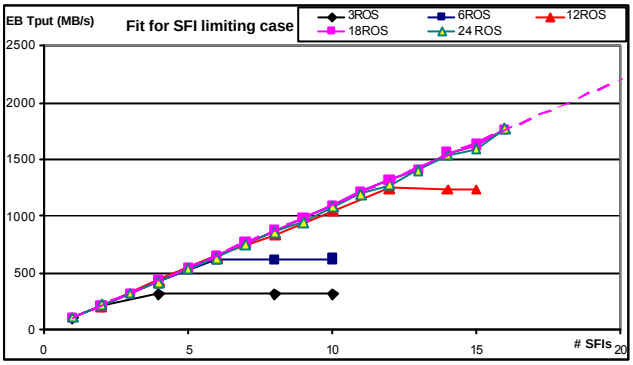
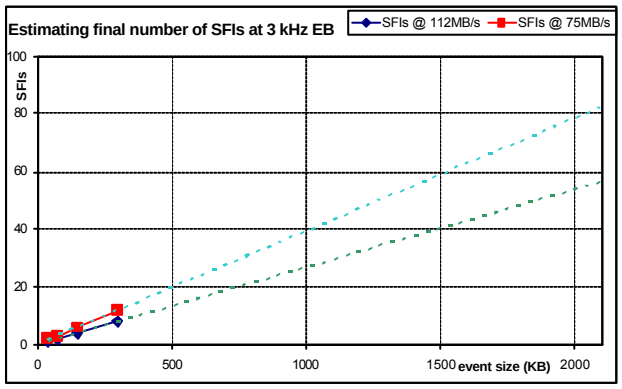


Figure 3-b contains the rates converted into throughput showing the linear scaling of the EB system. Therefore when the ROS is not the limiting factor one can predict the total throughput of a setup as represented by the fitted curve (dotted line) of Figure 3-b. This linear fit can be used to estimate the expected EB rate in a system with N ROSs and M SFIs provided that it is not limited by the ROS. The final system throughput of ~6GB/s thus requires a minimum of 60 SFIs. Figure 4 shows a similar value using the requirement of 3 kHz of EB for 2 MByte events. Data received by the SFIs was not output to the EF for the measurements discussed in this note. Taking into account some contingency and the worst-case performance decrease of 40%, (measured with very small event size) in case of output to EF network, the number of SFIs matching the final system requirements was predicted to be about 100, using 2.4 GHz SMP PCs.

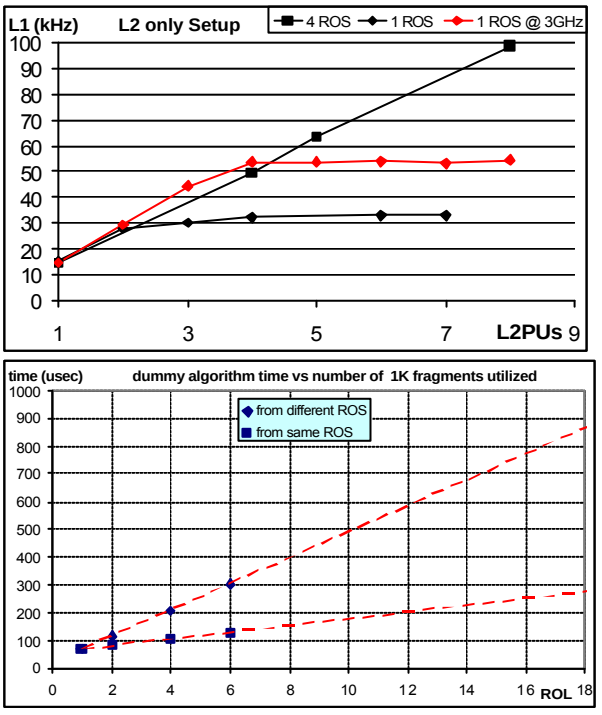
Figure 4 Estimation of required number of SFIs for final ATLAS events of 1.5 -2 MByte



L2 setup: L2PU performance measurement

In the final ATLAS system each L2PU will run complex physics algorithms using ROI data to accept only about 3% of the events. Since these selection algorithms are not yet finalized, the L2PUs run without any algorithms, each L2PU processing 6 events concurrently. In this way, with limited resources in the prototype setup, it becomes possible to find the limitations of the ROSs and to exploit the larger systems. It can also be argued that such an overloaded L2PU would in fact represent a cluster of L2PUs connected to the L2 network via a concentrating switch foreseen for the final system but not available in the prototype setup. In the final configuration of an L2PU, it is foreseen to process only 2 events concurrently and to have about 6 such nodes connected to the L2 network via concentrating switches as shown in Figure 2 by dashed ellipses. Figure 5a shows the L1 rate obtained for n_{rol}=? versus number of L2PUs. The diamond shaped points are data corresponding to a single ROS showing its cpu limitations in the plateau region, in case of a 2 GHz node (black curve) and a 3.06 GHz node (red curve). For the case of 4 ROSs (black squares), when ROS is not the limiting factor the linear increase proves the good scaling behaviour of the L2 system. Figure 5b shows the time necessary to fetch ROI fragments for a single L2PU. The diamond shaped data points are for the case each fragment was requested in series from different ROS units whereas the squares are for the same ROS. The two linear extrapolations (dashed lines) are to study two extreme cases for higher number of ROI fragments. Thus the worst-case scenario of 16 fragments of 1 KByte, all from different ROS units requires 800 µs. This amounts to 8 % of 10ms, the L2 time budget per event on a 3.06 GHz SMP L2PU, if the final system is constructed with 500 such nodes, processing only two events concurrently.

Figure 5 a) Maximum L1 rates for different setups b) latency of the L2 system



Combined setup: ROS performance measurement

The plateau region that is common to both EB and L2 only systems comes from the limitations on the ROS cpu utilization which can be expressed as:

$$CPU = R_{EB} \times CPU^{EB} + R_{L2} \times CPU^{L2} + R_{L1} \times CPU^{Cl} \quad (2)$$

CPU^{EB}, CPU^{L2}, CPU^{Cl} are the cpu utilization in GHz per kHz of Event Building, of L2 ROI collection and Event clears respectively. These values can easily be obtained from the EB and L2 studies in the previous sections. The determination of the cpu load by the 'clear' task can be done using the plateau region in combined systems as presented in Figure 6. The plateaus for large number of L2PUs, originate from the exhaustion of ROS cpu since the throughput from EB and L2 tasks are well below the Gigabit link capacity. Assuming the rest of the computing power goes to clear events from the ROBs, equation 2 allows calculation of the cpu utilization for the handling of clear messages. The values obtained are summarized in Table 2 showing that the 2GHz ROS cannot match ATLAS requirements whereas the 3.06GHz ROS is adequate to achieve the necessary performance.

Figure 6 Combined systems to measure the clear messages

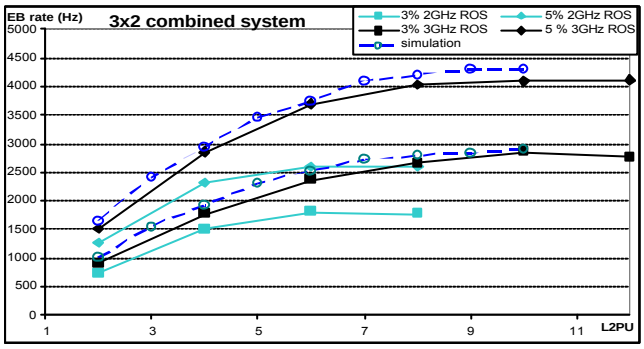


Table 2 ROS CPU utilization for different tasks

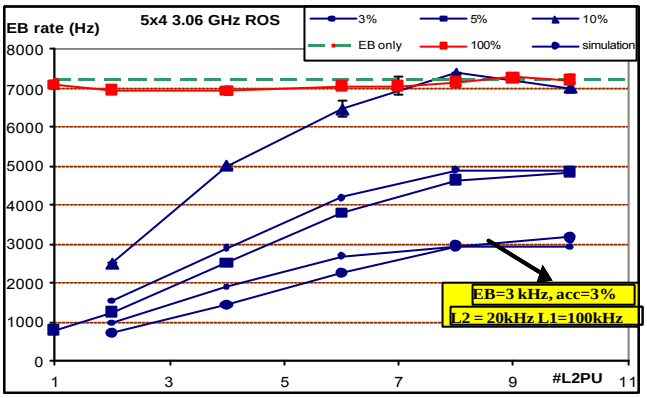
	2GHz CPU	3.06GHz CPU
Max L2 rate (kHz)	33.0	55.5
Max EB rate (kHz)	8.6+	14.5+,*
CPU per 1kHz of L2 Task (GHz)	0.06061	0.05564
CPU per 1kHz of EB Task (GHz)	0.2252	0.20274
CPU per 1kHz of Clear Task (GHz)	0.0074	0.0083

+ Includes clearing event fragments as well.
* Using two NICs and special test requesters (for Gigabit Ethernet).

3 RESULTS AND CONCLUSIONS

The largest possible system in the prototype setup consisted of 18 ROSs, 16 SFIs and 14 L2PUs controlled by 6 L2SVs. The EB rate of this system as a function of the number of L2PUs for different L2 acceptance percentages is shown in figure 7a. Since the results presented in Table 2 show that the 2GHz ROSs used in this large setup cannot deliver the required performance, this exercise was only intended to show that the DAQ performance scales as expected and that larger systems could be operated reliably. The Figure 7b shows that the final requirements can in fact be matched using 3GHz ROS units. This can be seen in the lower curve with circles, where 3kHz EB and 20kHz of L2 ROI collection rate at 100kHz of L1 input rate have been obtained. The dashed curves on the same plot are from the discrete event simulation showing good agreement between the model and the measurements, which also indicates the good emulation of the input channel prototype hardware.

Figure 7 Largest possibly combined systems in the prototype setup with (a) 2GHz ROS (b) 3.06 GHz ROS

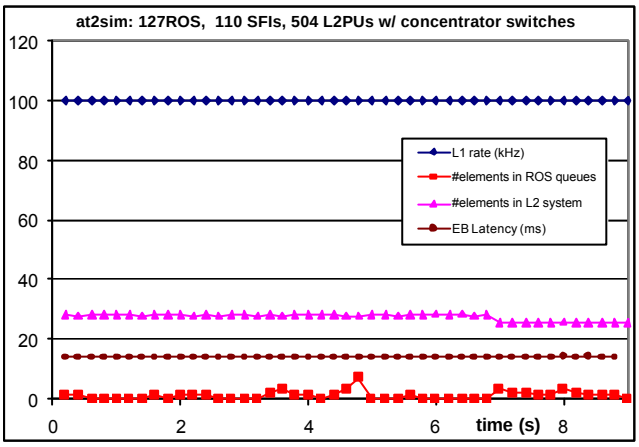


The results of scaling up from the prototype setup to the size of the final ATLAS configuration are presented in figure 8. This simulation contained 127 of the fast (3.06GHz) ROS nodes, 110 SFIs and 504 L2PUs collecting ROI data using a realistic physics trigger menu. 6 L2PUs were connected to the L2 network via concentrator switches and were not running algorithms. The results from this simulation are presented in Figure 8 for the first 9 seconds of a run at 3.5kHz accept rate. The sampling done every 0.2s shows that the overall system runs at the initial 100 kHz of L1 rate as shown by the upper curve. The other curves in the plot justify that no component is under stress: there are no accumulation of events in the L2 or EB systems; there is also no congestion in the network switches. This proves that the requirements of the final ATLAS data acquisition can already be matched using available hardware and software components.

The case studies for slow components have also been performed. For a slow ROS that cannot handle the requests fast enough and introduces long latency in both the L2 and EB systems, two distinct mechanisms exist for slowing down the overall rate to an acceptable level. If all the SFIs in the EB system are busy, an X-OFF message is

sent by the DFM to the L2SVs that momentarily turn off the whole L1 input. This action slows also the L2 subsystem and allows the EB system to recover and eventually turns on the L1 input via an X-ON message. In the other case - if all the L2PUs in the L2 system are busy, first the ROI collection is stopped, and if necessary the L1 input to individual input channels, allowing the L2 system to recover. The simulations have shown that the latter might be a smoother method for avoiding large load oscillations in the DataFlow system.

Figure 8 Simulation of te final ATLAS Dataflow using discrete event simulation model



REFERENCES

[1] ATLAS Collaboration, "ATLAS detector and physics performance technical design report," CERN/LHCC/99–14

[2] ATLAS Collaboration, "ATLAS HLT, DAQ and DCS technical design report" CERN/LHCC/2003–022

[3] http://www.intel.com/xeon

[4] http://linux.web.cern.ch/linux

[5] http://www.foundrynet.com

[6] http://www.batm.de/de/produkte/ip/multilayer/t6

[7] R. Cranfield, P. Golonka, A. Kaczmarska, K. Korcyl, J. Vermeulen and S. Wheeler, "Computer modeling the ATLAS Trigger/DAQ system performance," in Proc. IEEE 13th Conference of Real-Time Computer Applications in Nuclear, Particle & Plasma Physics (RTCA), 2003

[8] HP. Beck *et al.*, "The Baseline DataFlow system of the ATLAS Trigger and DAQ" in Proc. IEEE 13th Conference of Real-Time Computer Applications in Nuclear, Particle & Plasma Physics (RTCA), 2003