The Baseline DataFlow System of the ATLAS Trigger & DAQ

M. Abolins[1], A. Dos Anjos[2], M. Barisonzi[3,4], H.P. Beck[5], M. Beretta[6], R. Blair[7], J. Bogaerts[8], H. Boterenbrood[3], D. Botterill[9], M. Ciobotaru[8], E. Palencia Cortezon[8], R. Cranfield[10], G. Crone[10], J. Dawson[7], B. Di Girolamo[8], R. Dobinson[8], Y. Ermoline[1], M.L. Ferrer[6], D. Francis[8], S. Gadomski[5,11], S. Gameiro[8], P. Golonka[8], B. Gorini[8], B. Green[12], M. Gruwe[8], S. Haas[8], C. Haeberli[5], Y. Hasegawa[13], R. Hauser[1], C. Hinkelbein[16], R. Hughes-Jones[14], P. Jansweijer[3], M. Joos[8], A. Kaczmarska[15], E. Knezo[8], G. Kieft[3], K. Korcyl[15], A. Kugel[16], A. Lankford[17], G. Lehmann[8], M. LeVine[18], W. Liu[6], T. Maeno[8], M. Losada Maia[2], L. Mapelli[8], B. Martin[8], R. McLaren[8], C. Meirosu[8], A. Misiejuk[12], R. Mommsen[17], G. Mornacchi[8], M. Müller[16], Y. Nagasaka[19], K. Nakayoshi[20], I. Papadopoulos[8], J. Petersen[8], P. de Matos Lopes Pinto[8], D. Prigent[8], V. Perez Reale[5], J. Schlereth[7], M. Shimojima[21], R. Spiwoks[8], S. Stancu[8], J. Strong[12], L. Tremblet[8], J. Vermeulen[3]*, P. Werner[8], F. Wickens[9], Y. Yasu[20], M. Yu[16], H. Zobernig[22], M. Zurek[15]

ATL-DAQ-2003-032
16 October 2003

Abstract

In this paper the baseline design of the ATLAS High Level Trigger and Data Acquisition system with respect to the DataFlow aspects, as presented in the recently submitted ATLAS Trigger/DAQ/Controls Technical Design Report [1], is reviewed and recent results of testbed measurements and from modelling are discussed.

I. Introduction

In the ATLAS experiment the data from events accepted by the first-level (LVL1) trigger will flow from the Front-End electronics into the Read-Out Drivers (RODs). The RODs are subdetector specific and assemble the event data received into fragments, which are passed via the Read-Out Links (ROLs) to the Read-Out Buffers (ROBs). In total there are 1600 ROLs and ROBs. Each ROL is an S-LINK capable of transferring 160 MByte/s [2]. The design of the ROBs is identical for all subdetectors. The ROLs cross the boundary between detector specific electronics and the High Level Trigger and DAQ system. In this paper the movement of data - the "DataFlow" - in the baseline design of this system is discussed. The baseline design is described in the Technical Design Report [1], recently submitted to the LHCC. In Figure 1 a schematic layout of the Trigger and DAQ system of ATLAS is presented.

An important parameter for the system is the accept rate of the first-level trigger. This is at maximum 75 kHz and required to be upgradable to 100 kHz. The rate with which the second-level trigger has to produce decisions is equal to this accept rate. These decisions are taken using a small fraction (of the order of 1 - 2%) of the event data, which are requested on the basis of information provided by the LVL1 trigger. Complete events are built with the second-level trigger accept rate of about 3 kHz at maximum. The average event size is expected to be about 1.5 MByte. The bandwidth required for transporting data to the High Level Triggers, i.e. the second-level (LVL2) trigger and the Event Filter, which operates on fully built events, is therefore several GByte/s at nominal

1. Michigan State University, Department of Physics and Astronomy, East Lansing, Michigan
2. Universidade Federal do Rio de Janeiro, COPPE/EE, Rio de Janeiro
3. NIKHEF, Amsterdam
4. Universiteit Twente, Enschede
5. Laboratory for High Energy Physics, University of Bern
6. Laboratori Nazionali di Frascati dell' I.N.F.N., Frascati
7. Argonne National Laboratory, Argonne, Illinois
8. CERN, Geneva
9. Rutherford Appleton Laboratory, Chilton, Didcot
10. Department of Physics and Astronomy, University College London, London
11. On leave from Henryk Niewodniczanski Institute of Nuclear Physics, Cracow
12. Department of Physics, Royal Holloway and Bedford New College, University of London, Egham
13. Department of Physics, Faculty of Science, Shinshu University, Matsumoto
14. Department of Physics and Astronomy, University of Manchester, Manchester
15. Henryk Niewodniczanski Institute of Nuclear Physics, Cracow
16. Lehrstuhl für Informatik V, Universität Mannheim, Mannheim
17. University of California, Irvine, California
18. Brookhaven National Laboratory (BNL), Upton, New York
19. Hiroshima Institute of Technology, Hiroshima
20. KEK, High Energy Accelerator Research Organisation, Tsukuba
21. Department of Electrical Engineering, Nagasaki Institute of Applied Science, Nagasaki
22. Department of Physics, University of Wisconsin, Madison, Wisconsin

* Corresponding author: J.Vermeulen@nikhef.nl

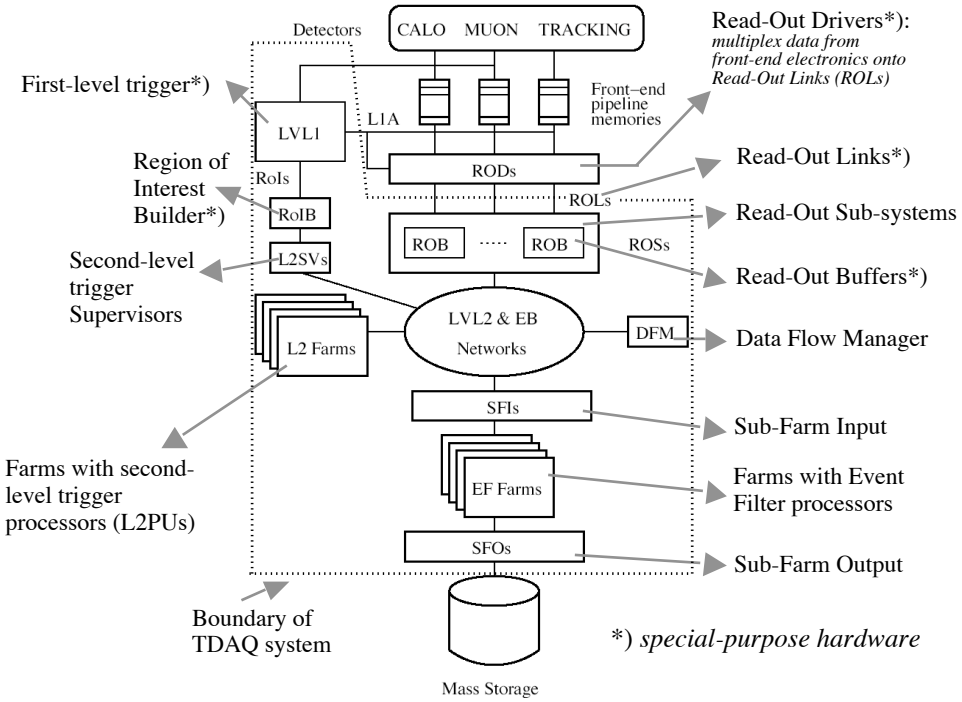

Figure 1: Schematic layout of the Trigger and DAQ system of ATLAS

operating conditions and up to about 7 GByte/s at 100 kHz LVL1 accept rate. The final accept rate is estimated to be about 200 Hz. Data from accepted events will be sent to the CERN computer centre for permanent storage.

II. Implementation Aspects of the Base line DataFlow System

The Read-Out Subsystems (ROSs) are planned to be built from industrial rack-mounted PCs (4U high) with dedicated PCI cards implementing ROB functionality. In the current prototypes each PCI card (ROBIN) contains two ROBs buffering the data from two ROLs. It is anticipated that the production version of the ROBIN will contain 4 ROBs. The event data can be output via a Gigabit Ethernet interface on the card as well as via the PCI bus. Each PC is planned to have two Gigabit Ethernet interfaces, one intended for receiving requests from and sending event fragments to the second-level trigger system, the other one for receiving requests from and sending event fragments to the Event Builder. The Ethernet interfaces of the ROBIN cards and of the PC provide alternative data paths. In the "bus-based read-out" option of the baseline design all event data flows through the PCI busses of the PCs. There are 144 of these PCs, mounted in about 15 racks, with each PC handling the data from 12 ROLs. The ROSs are located underground in the USA15 area. The ROBIN is discussed in more detail in [3].

The Region of Interest (RoI) Builder builds per event a message (RoI record, see also section III) from data sent to it by various parts of the LVL1 trigger (via S-LINKs). A RoI record is sent to one of the second-level trigger supervisor processors via a dedicated S-link. The RoI builder consists of 9U VME cards fitting in a single 9U crate [4].

For the LVL2 and Event Builder networks Gigabit Ethernet technology will be used. For "bus-based read-out" each ROS PC connects directly to a central LVL2 switch and a central Event Builder switch. In case of direct connection of ROBINs to the network, "concentrating switches" (located in USA15) connect groups of ROBINs via "uplinks" (two per switch) to the LVL2 and Event Builder networks. In Figure 2 a schematic overview of the Ethernet network architecture is presented, the central switches will be located in the SDX15 building at the surface, as well as e.g. the LVL2 processor farms and the Sub-Farm Interfaces (SFIs) taking care of event building. Switches of the size of the central switches (with about 250 ports) are commercially available now.

The protocol to be applied in the LVL2 and Event Builder networks is still subject of investigation, raw Ethernet or UDP are the most likely candidates. It is possible to use these protocols due to the request driven transfers of event data in the system. As discussed in section V, by limiting the number of outstanding requests and by controlling the request patterns, queue overflow causing frame loss can be avoided. Hence a protocol with support for automatic retransmission in case of frame loss is not needed. TCP/IP is to be used in the Event Filter farms.

The High Level Triggers are built from farms of PCs. The estimates for the required number of 8 GHz dual CPU machines, indicated in Figure 2, are based on an extrapolation of measurement results obtained with 2 – 2.4 GHz dual Xeon PCs. Other components in Figure 1 and 2, discussed in the next section, are also implemented with PCs. All PCs make use of Linux as operating system. The DataFlow software consists of multi-threaded applications written in C++. A detailed discussion on the design of the software and on experience with the use of threads in combination with the Standard Template Library (STL) can be found in [5].

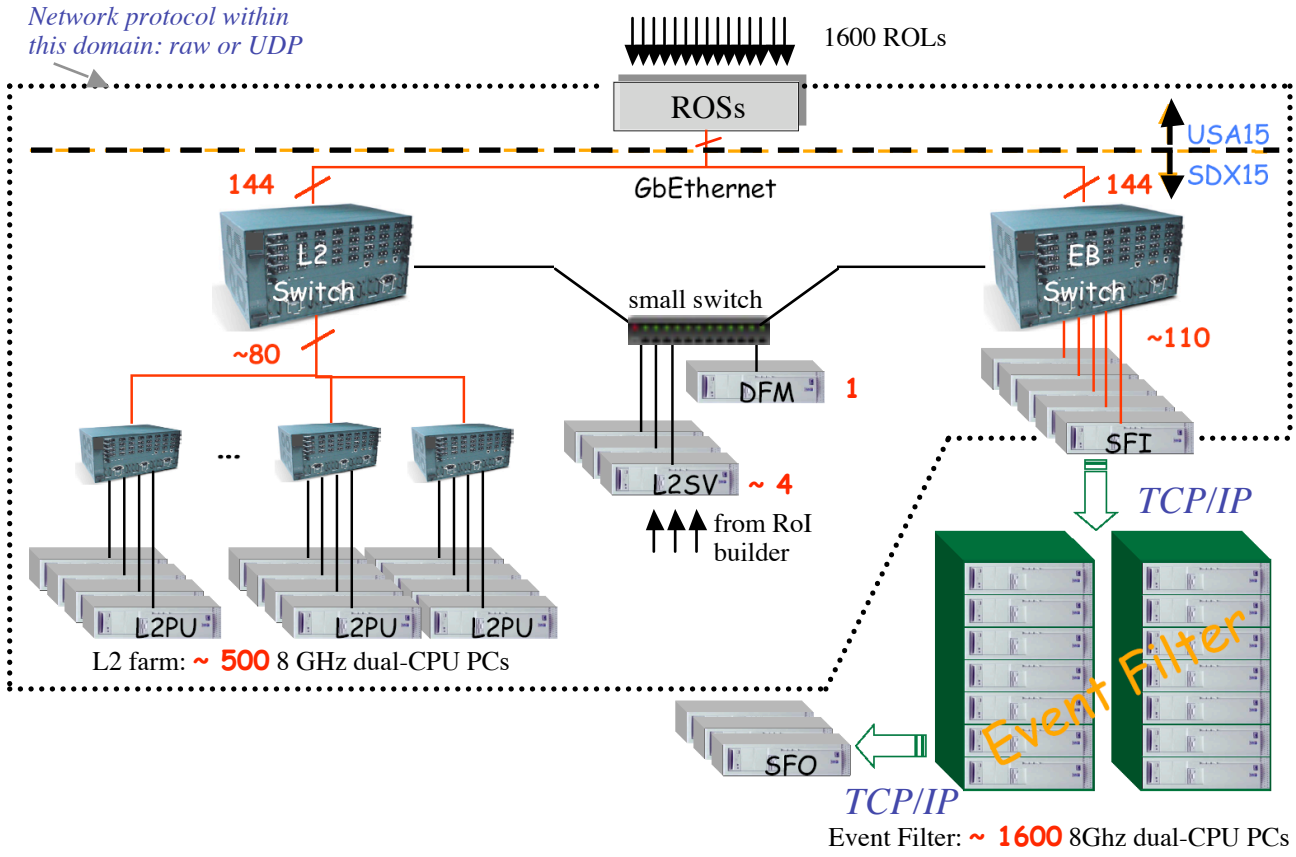

Figure 2: Schematic overview of the network architecture

III. DataFlow: Requests and Responses

The flow of data in the ATLAS trigger/DAQ system is controlled by the LVL2 supervisors for the LVL2 system and by the DataFlow Manager (DFM) for the Event Building system. In this section the pattern of requests and responses is reviewed.

A LVL2 Supervisor sends, after receiving of a RoI record from the RoI Builder, RoI information to one of the second-level trigger processors (L2PUs). The RoI information indicates which data has to be requested from the ROSs as input for the LVL2 selection. An L2PU will request data corresponding to a RoI in steps, e.g. for an electron/gamma RoI first data from the electro-magnetic calorimeter are required, next from the hadron calorimeter and then from the inner detector. Only a fraction of the events are selected at each step, for the example the expectation is about 19% at the first step and 11% of the original number at the second step. After production of a decision by an L2PU, the decision is communicated to the Supervisor that sent the RoI request. For accepted events, data produced by the trigger algorithms are also passed to the "pseudo ROS" (pROS, not shown in the figures. The pROS connects to the small switch in Figure 2 to which also the LVL2 Supervisors and the DFM connect). Decisions are passed to the DFM.

For each event accepted by LVL2 the DFM sends a build request to an SFI. This in turn sends requests for data to the ROSs (including the pROS). The ROSs return the fragments (identified by the LVL1 id) requested. After completion of event building an EoE (End of Event) message is sent by the SFI to the DFM. The DFM stores these and LVL2 reject messages until ~ 300 have been received. Event clear commands for the LVL1 ids associated with the EoE and LVL2 reject messages are then sent to the ROSs, with ~ 300 of these commands in a single message. These messages are multi-cast, and the rate is the LVL1 accept rate divided by the grouping factor (330 Hz for a LVL1 accept rate of 100 kHz).

After building the event it is delivered on request to one of the Event Filter processors. A further decision is taken on acceptance or rejection. The data of accepted events are passed to the "Sub-Farm Outputs" (SFOs), where the events are buffered and passed to central mass storage in the CERN computer centre.

IV. Rates

The message frequencies and data volumes associated with event building can be calculated in a straightforward way. Input for such a "back-of-the-envelope" calculation or "paper model" are the event building rate and for the data volumes the average fragment sizes per ROS (which are detector specific).

A paper model for the LVL2 system is more complicated and has been initially implemented in a spreadsheet. The current implementation of a paper model of the full system is in the form of a small C++ program (see the appendix of [1]). A

basic assumption made in it is that the RoI rate does not depend on the η and ϕ of the centre of the RoI, only on the area in η-ϕ space associated with the RoI. The RoI rates can then be obtained with a straightforward calculation using:
- the LVL1 accept rate,
- the exclusive rates of the various LVL1 trigger menu items,
- the number of RoIs associated with each trigger item,
- the η-ϕ area associated with each possible RoI location.

The request rates are then obtained using:
- information on the mapping of the ROLs onto the detector,
- the acceptance factors of the various LVL2 trigger steps,
- the size of the η-ϕ areas from which data is requested (RoI and detector dependent).

It is found that an L2PU generates on average about 1.6 times per event requests for data within one RoI from one or several subdetectors. With the method outlined the average LVL2 request rate per ROS can be calculated, results are shown in Figure 3 for the expected initial luminosity of the LHC (2.10^{33}) and a LVL1 accept rate of 100 kHz. Results for the average data volume to be transferred per ROS to the LVL2 system under the same conditions are shown in Figure 4. The entries above 10 MByte/s are all due to ROSs handling data from the calorimeters, the entries below 10 MByte/s are associated with the other detectors.

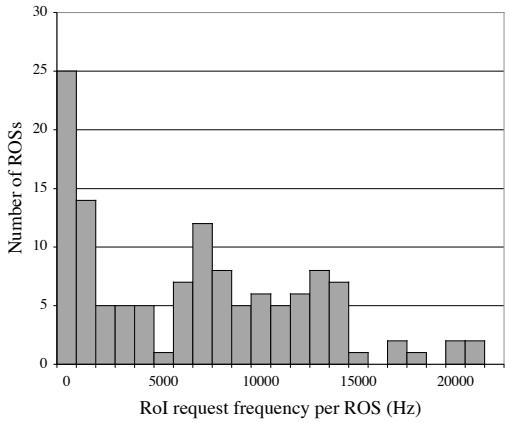

Figure 3: Number of ROSs associated with different RoI request frequency intervals (of 1000 Hz) for low luminosity, for ROSs buffering data that can be used by the second-level trigger

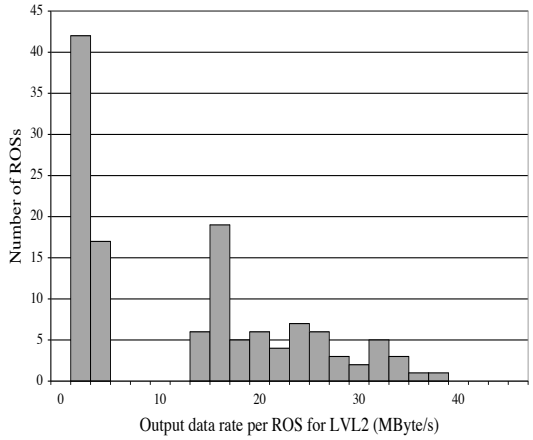

Figure 4: Number of ROSs associated with different data rate intervals (of 2 MByte/s) for low luminosity, for ROSs buffering data that can be used by the second-level trigger

V. Queuing and Load Balancing

Queuing in the switches will occur in particular in the output ports of the central switches connecting to the LVL2 subfarm switches and to the SFIs. Also in the output ports of "concentrating switches" connecting individual ROBINs to the LVL2 and Event Builder networks, queuing may occur. Without flow control active, queuing in input ports is less likely to occur as present switches in general have high internal bandwidths preventing blocking of data transfers.
The lengths of the queues in the output ports can be minimized by:
- limiting the number of outstanding requests for each L2PU or SFI,
- limiting or minimizing for each L2PU or SFI the number of events handled simultaneously,
- taking into account the LVL2 subfarm structure for the pattern of assignment of events to the L2PUs and for the sequence of request patterns generated by the SFIs how the "concentrating switches", if present, are connected to the central EB switch.

Avoidance of queue overflow is important for preventing frame loss. Flow control also provides an effective measure against frame loss, but may lead to undesired blocking of data transfers other than the transfer that has to be temporarily halted [6].

Minimizing for each L2PU the number of events handled simultaneously also results in good load balancing. It is straightforward to implement this strategy in the LVL2 Supervisors, long tails in the LVL2 decision time distribution are avoided in this way and high average utilizations of the L2PUs are feasible, as has been demonstrated with discrete event simulation. The simulation also provides information on the maximum number of event fragments to be stored in each ROB; this is about 3000 for the current estimates of the parameter values. With a maximum size of 1.6 kByte per fragment this corresponds to about 5 MByte, an order of magnitude lower than the buffer memory in the current ROBIN prototype (64 MByte) [3]. Relevant results on the distribution of the LVL2 decision time as well as results on queue formation can be found in ch. 14 of [1] and in [7].

VI. Testbed Measurements

Testbeds have been set up to validate the design choices made. One setup includes 128 traffic generators for Fast Ethernet and up to 16 traffic generators for Gigabit Ethernet [8]. Results of studies performed with this setup have been reported in ch. 14 of [1], together with a comparison to "computer model" (simulation) [7] results. With the traffic generators the data entered via 1600 access points (via concentrating switches) into the Event Builder network was emulated. Scaling of the event building rate with the number of SFIs used was demonstrated for 1 – 11 SFIs and up to saturation of the output bandwidth of the traffic generators. Per SFI an event building rate of about 30 Hz was observed, i.e. with 100 SFIs as used in the testbed an event building rate of 3 kHz should be possible. This is also predicted by the model with which the testbed results were described correctly.

Since the submission of [1] also measurements on event building with bus-based ROSs (2.2 GHz PCs) with emulated ROBINs (handling 1.4 kByte fragments from 12 ROLs) have been performed. In Figure 5 results are presented for a testbed with 6 ROSs, for two different types of switches and together with modelling results obtained with at2sim [7] using calibrated models. It can be seen that the building rate is limited to 6 kHz due to the performance of the ROSs. This is close to the maximum rate in the testbed of about 7 kHz, as imposed by the available bandwidth of a single Gigabit Ethernet link. Raw Ethernet protocol has been used, flow control was switched on and the maximum number of outstanding requests per SFI was 20. Results for more ROSs and also using UDP and with flow control switched off are being obtained at the time of submission of this paper.

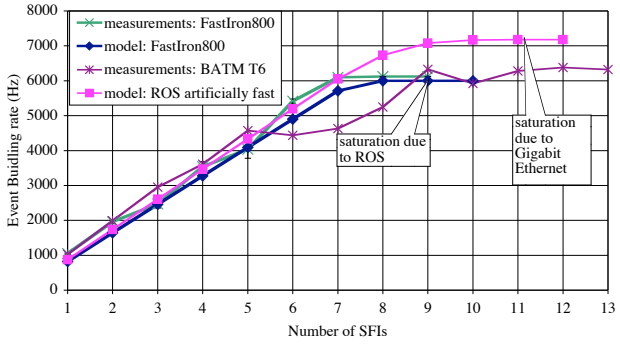

Figure 5: Event building rate for 6 ROSs with emulated ROBINs and 12 ROLs per ROS (1.4 kByte per ROL), raw Ethernet protocol, flow control on and a maximum of 20 outstanding requests per SFI

Results for the maximum RoI request rate in a testbed setup with 4 ROSs (2.2 GHz PCs) with emulated ROBINs (12 ROLs per ROS) and 1-11 L2PUs are presented in Figure 6. About 20 kHz request rate per ROS is possible if the data from a single ROL are requested. This rate is about the maximum RoI request rate as obtained from the paper model.

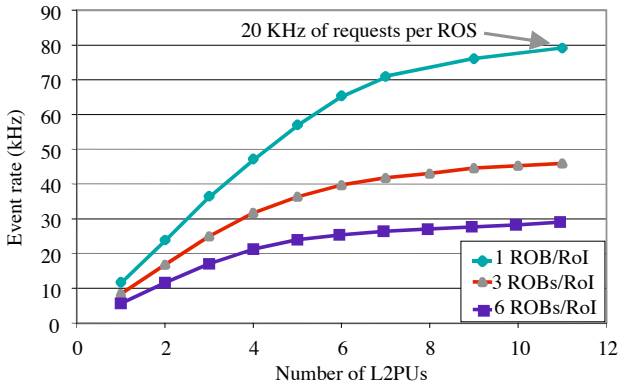

Figure 6: Maximum RoI request rate for 1-11 L2PUs (not running algorithms) fetching data from 4 ROSs (12 inputs each). Different curves corresponds to different ROI sizes, 1.4 kByte per ROB was requested using raw Ethernet protocol, the ROS PCs were 2.2 GHz machines, the response of real ROBINs was emulated in software

The results presented in Figure 7 show that for a 3 GHz PC with three hardware ROBIN emulators (4 ROLs per ROBIN), from which data are collected via the PCI bus and a single network interface, a RoI request rate of about 20 kHz (for 1.4 ROL/RoI and l kByte event fragments) can be combined with

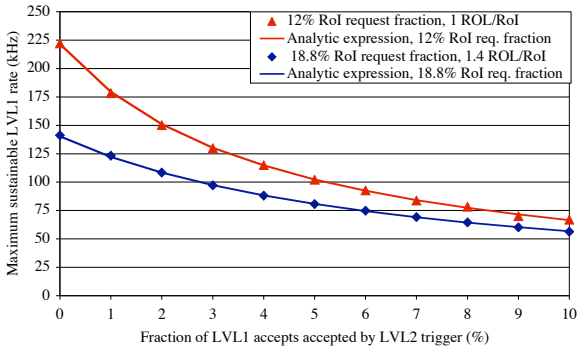

Figure 7: Maximum LVL1 accept rate as function of the LVL2 accept fraction for two different RoI request fractions for a single ROS consisting of a 3 GHz PC, further details are given in the text

3 kHz event building for a LVL1 accept rate of 100 kHz. Requests were generated with a PC running a dedicated program. The lines labeled with "analytical expression" result from describing the inverse LVL1 accept rate as a linear function of the RoI request and accept fractions.

VII. CONCLUSIONS

The implementation of the baseline design of the ATLAS DataFlow system of the ATLAS Trigger and DAQ system is based on the use of standard rack-mounted PCs running Linux, with multi-threaded application software written in C++. The PCs are interconnected using Gigabit Ethernet. Only for the RoI Builder and for the ROBINs dedicated hardware will be used.

The system design is complete, but optimization of the I/O at the Read-Out System level and of the deployment of the LVL2 and Event Builder networks is possible. The architecture allows for deferring the purchase of part of the system and upgrading its rate capability in a later stage.

Testbed and modelling results validate the baseline design of the DataFlow system (as well as the deployment - not reported in this paper - of the design at the ATLAS H8 test beam). Further testbed and modelling studies are under way to ensure the absence of potential problems.

VIII. REFERENCES

[1] ATLAS-TDR-016; CERN-LHCC-2003-022, http:// cdsweb.cern.ch/search.py?recid=616089

[2] A.Ruiz Garcia, HOLA Specification, EDMS Note, 330901 (2001), https://edms.cern.ch/document/330901/1

[3] M.Müller et al., "A RobIn ProtoType for a PCI-Bus based Atlas Readout-System", these proceedings

[4] R.Blair et al., EDMS Note 367638 (2003), https://edms.cern.ch/document/367638/2.0

[5] S.Gadomski et al., CHEP2003 Proceedings, ATL-DAQ-2003-007, http://cdsweb.cern.ch/search.py?recid=621381

[6] S.Stancu et al., CHEP2003 Proceedings, ATL-DAQ-2003-003, http://cdsweb.cern.ch/search.py?recid=622032

[7] R.Cranfield et al., "Computer modeling the ATLAS Trigger/DAQ system performance:", RT2003 Conference Proceedings, accepted for publication in IEEE Trans. on Nucl. Sci.

[8] M.J.Levine et al., ATL-DAQ-2003-009, http://cdsweb.cern.ch/search.py?recid=621754