A Sequential Processing Strategy for the ATLAS Event Selection*

J. Bystricky, D. Calvet, J. Ernwein, O. Gachelin, T. Hansl-Kozanecka, J. R. Hubbard, M. Huet, P. Le Dû, I. Mandjavidze, M. Mur, M. Smizanska, B. Thooris

CEA Saclay, 91191 Gif-sur-Yvette CEDEX, France

Abstract

A new processing strategy for the ATLAS event selection system is proposed for use downstream of the first level trigger. It aims to achieve a reduction of the event rate to a level suitable for recording on permanent storage via a sequential treatment of the "regions of interest" selected by the level 1 trigger. A possible Trigger/DAQ architecture which exploits this concept is presented. It is based on a single processing farm and a network which transports data and protocol traffic. The resources necessary for the implementation of this architecture are estimated. A set of performance figures obtained by calculations and computer simulations are given assuming the use of ATM networking technology.

I. Introduction

The ATLAS experiment [1] proposed at the CERN Large Hadron Collider (LHC) will place stringent demands on the data acquisition and on-line filtering systems. A sophisticated multi-level selection system will reduce the raw data flow from a few tens of TByte/s to about 100 MByte/s to be recorded on permanent storage for off-line analysis. Several possible data acquisition architectures are currently under study. In any case, a first reduction of the initial data rate will be carried out by fast pipe-lined logic that will retain only those events that satisfy simple geometrical and energy deposition criteria. After this first level of selection the data bandwidth is expected to be of the order of 100 GByte/s.

In this paper we propose a new triggering scheme for use downstream of the first level selection. It consists of a number of sequential steps to reject background events as soon as possible. At each step, only the event data which is necessary to make a decision is acquired and analyzed. The full event reconstruction is performed only when required for physics analysis. This method allows the reduction of the bandwidth and processing power requirements.

We propose an implementation of this scheme using a single processing farm. We present a simple methodology to estimate the processing power required to achieve the targeted rejection factor for high and low luminosity operation of the LHC. A network with several tens of Gbit/s throughput is used to link about 1500 read-out buffers to a similar number of processors. This network carries both data and protocol traffic. We investigate the use of Asynchronous Transfer Mode (ATM) technology [2] to build this high performance network. A simulation model of the proposed architecture has been developed. The performance of the switching network under the expected traffic pattern has been assessed. A selected set of performance figures obtained by simulation are presented.

II. The ATLAS Trigger/DAQ

This section outlines a functional overview of the three level event selection and DAQ system as described in the ATLAS Technical Proposal [1].

The level 1 trigger (LVL1) operates at the LHC beam-crossing rate of 40 MHz. The front-end read-out electronics is designed for a maximum LVL1 output rate of 100 kHz. Present studies estimate this output rate to be about 30-40 kHz. For events accepted by the LVL1, data from all detector front-end electronics is transferred to some 1500 Read-Out Buffers (ROB). The level 1 trigger defines "Regions of Interest" (RoI) to guide subsequent event selection.

The level 2 trigger (LVL2) has access to the full granularity and full precision data from all of the subdetectors. In order to reduce the data transfer requirements and the decision latency, only data belonging to "Regions of Interest" (RoI) is transmitted to the level 2 processors (about 1% of the front-end information). The RoIs are processed to extract "features" such as calorimeter cluster energy or track parameters. Individual particle identification requires combining features from different subdetectors. The LVL2 decision is issued after a topological analysis of the event. The output event rate resulting from the level 2 trigger selection is estimated to be 1.5 kHz.

The level 3 trigger (LVL3) uses full event data to perform an event analysis similar to that of the off-line reconstruction. LVL3 provides a further rejection factor of about 10. Events accepted by the level 3 selection are recorded on permanent storage for subsequent off-line studies.

The ATLAS Technical Proposal describes the "Data-Driven" and "Local-Global" options for implementation of the level 2 trigger. Suggested "Push" data flow protocol implies that data of all RoIs from all subdetectors are systematically sent to the feature extraction processors via dedicated local networks. Even though the RoI concept reduces the requirements on the data transmission and processing power, these requirements still remain high. The final level 2 decision is issued by a processor from the "Global Farm", which communicates with the feature extractors by a "Global" network. A dedicated control network is used to transfer the level 2 decisions to the read-out buffers and the level 3 system. A separate network is used to connect the ROBs to the level 3 processors.

We have studied ways to avoid the complexity of the Trigger/DAQ system, which is due to the parallelism in the processing of RoIs, the separation of the level 2 and level 3 event selection, the presence of several different networks and the "Push" data flow protocol [3]. In this paper we propose a different approach based on a sequential event selection strategy.

*Talk given by P. Le Dû at Nuclear Science Symposium in Anaheim, California, 3 - 9 November 1996

III. Sequential Event Selection

Each RoI identified by the level 1 trigger is characterized by its type (μ, e/γ, jet, etc...), its spatial coordinates and information on energy and isolation thresholds. Two categories of RoIs can be distinguished. The "trigger RoIs" are those that contributed to the level 1 decision. The "secondary RoIs" give additional information on the global topology of the event. Figure 1 shows the distributions of the total number of RoIs and the number of trigger RoIs for di-jet events (about 70% of the events accepted by the level 1 trigger).

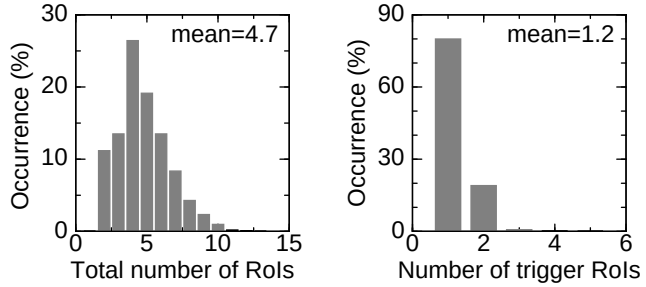

Figure 1: Number of RoIs per event.

A significant event rejection can be achieved by processing only trigger RoIs as most of them are not confirmed with the higher granularity and improved precision of the level 2 algorithms. The calorimeter and/or muon detector data alone can be used to sharpen the energy and momentum cuts and to refine the regions of interest. In addition, confirmed μ and e/γ RoIs can be matched to tracks in the Inner Detector to lower the event rate even further. Table 1 shows the expected background rejection factors due to confirmation of the e/γ trigger RoIs at low (10^{33} $cm^{-2}s^{-1}$) and high (10^{34} $cm^{-2}s^{-1}$) luminosity. Each rejection factor corresponds to the rate reduction obtained for 90% efficiency for isolated electrons at the nominal threshold energy [4].

Table 1
Background rejection factors at 90% efficiency for e/γ trigger RoIs

Luminosity	Low	High
Nominal threshold (GeV)	20	40
Calorimeter algorithm alone	3	10
Calorimeter and Inner Detector algorithms	25	60

Execution times of tracking algorithms are an order of magnitude longer than those of the level 2 calorimeter and muon algorithms [5], [6]. The majority of events can be rejected using the calorimeter and muon features only (see Table 1 for the e/γ RoIs). Therefore, performing track-finding algorithms only for confirmed RoIs allows a reduction of the data transfer bandwidth and the processing power requirements.

A further reduction of the bandwidth and processing power can be achieved by transferring some of the trigger algorithms originally intended for LVL3 to the LVL2 system. This reduction is due to the different nature of the second and the third levels of event selection. The LVL3 selection is based on the full event data (~1 MByte) and code similar to that used for the off-line analysis. On the other hand, level 2 uses partial event data (1-16 kByte) and specialized trigger code, requiring less system resources.

Based on these considerations we propose the sequential event selection strategy for the ATLAS trigger, shown in Figure 2. Event selection proceeds with a number of successive steps. A decision can be issued at each step so that background events can be rejected as soon as possible. The data necessary for the next step is requested only if the analysis is still consistent with at least one set of trigger conditions. Figure 2 also shows estimated input rates for various processing steps.

In the first step, level 1 trigger RoIs are analyzed using data from calorimeter and muon subdetectors only. The event rate is reduced significantly by sharpening the energy and momentum cuts. In the second step the confirmed μ and e/γ regions of interests are matched to tracks found in the Inner Detector.

The next steps in the event selection sequence proceed with the confirmation of secondary RoIs. As for trigger RoIs, data from the calorimeter and muon systems is processed before the data from the Transition Radiation Tracker (TRT) and Semi-Conductor Tracker (SCT) subdetectors.

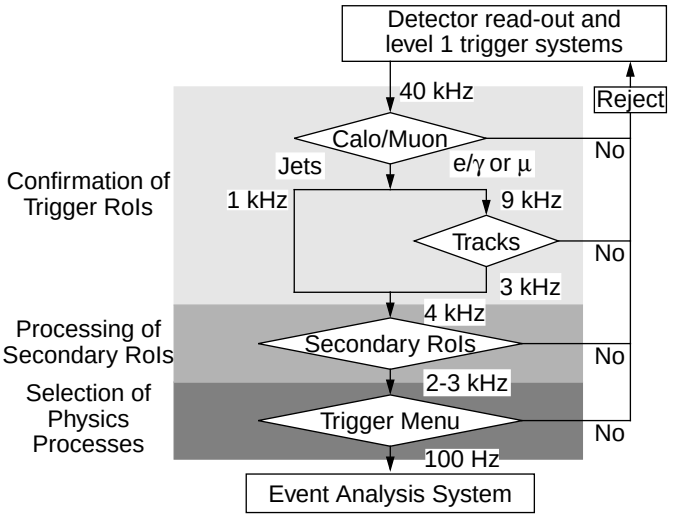

Figure 2. The sequential processing model.

The next event selection step is based on topological analysis of the event. More complex algorithms, such as mass calculations, missing E_T calculation, or a search for secondary vertices, can be used at this stage. The trigger menu is consulted and the decision is made to reject or accept the event for on-line analysis.

For accepted events, a data acquisition system collects the full event data. Partial event collection can be envisaged, if this will be motivated by physics requirements. The events are classified into calibration and various physics streams. Full or partial event data is recorded on permanent storage for subsequent off-line studies.

The actual reductions in data transfer bandwidth and processing power depend on the trigger menu under consideration; they can only be determined after detailed modeling of any proposed processing sequence. Results obtained from our preliminary studies will be given in Section V and Section VI.

IV. Single Farm Architecture

The sequential selection strategy suggests that a single processor executes all triggering algorithms and makes the final decision. We propose to implement this strategy using a single processor farm linked to the ROBs by a single switching network, as shown in Figure 3. Each source groups several ROBs from the same subdetector by a bus in order to reduce the number of network links. We think that 4 to 8 ROBs per source will match the network link bandwidth to the capacity of the bus. Similarly, a destination serves several processors. The network ports are bi-directional, so that control information can flow from the processors toward the ROBs.

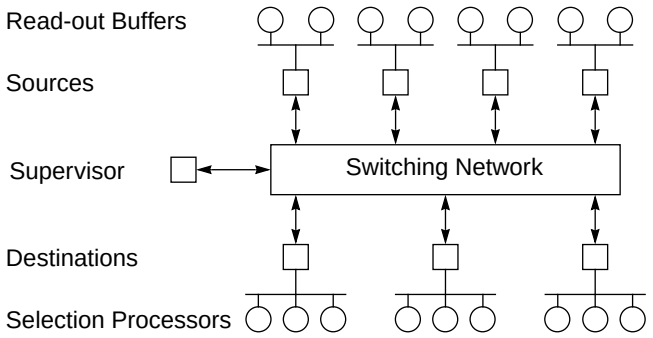

Figure 3: Proposed single farm architecture.

In what follows we describe the operation of the system under the so called “Pull” protocol which allows each destination processor to request data as needed.

For each event accepted by the level 1 trigger the supervisor gets a list of RoIs and sends this information to a destination. A processor within the destination is allocated to perform selection algorithms. It proceeds with the sequential steps described in Section III. At each step the processor sends request messages to the sources that contain the data needed to complete the current selection step. Each source gets the data from the ROBs, preprocesses it, and sends it to the destination. The destination passes the data to the processor, which executes the algorithm. After each step, the processor decides to continue with the event selection or to stop. If the processing is to continue, new data is requested from the sources. While waiting for the requested data the processor can execute selection algorithms for another event. If the processing is to stop, the trigger decision is sent to the supervisor that broadcasts it to all sources. The sources forward the decision to the ROBs. Rejected events are discarded. Accepted events are kept in the ROBs until their transfer to the DAQ system for event building and on-line analysis. It is possible that the same processor will perform the on-line analysis algorithms. Another option is to allocate a processor for this task from a farm dedicated to analysis (not shown in Figure 3).

V. Paper Model

A “paper model” [7] of the single farm architecture has been developed to help in the design effort and to allow comparisons between different event selection strategies. It uses average values for data volumes and algorithm execution times, as well as measured and estimated processing and data transfer overheads. The input data rates are given by trigger menus based on a catalog of physics processes expected at LHC. The model is used to evaluate system resource requirements and trigger decision latencies for various selection sequences.

A set of working assumptions have been used to model the “final” ATLAS Trigger/DAQ architecture. They are based on our current knowledge of the detector, on measurements performed on today’s hardware and on performance estimates of future hardware and software. In the paper model the PCI bus (Peripheral Component Interconnect - [8]) is used to group several ROBs in sources; an ATM network connects sources and destinations; each destination processor has ~500 MIPS computing power. We assume that the raw data in the ROBs is preprocessed before sending it to the selection processors. The computing power per ROB is equivalent to 100 MIPS.

The read-out organization of the detector described in [9] was used to calculate the data volume per ROB. Based on extensive studies of the ATLAS trigger menus [10] and processing sequences, measurements [5], [6], and estimates of algorithm execution times, we evaluate data transfer bandwidth and processing power requirements. This work allowed us to determine the dimensions of the event selection and analysis systems. We tried to match the capabilities of various elements such as the PCI bus, the ATM network and the processing farms to the detector read-out characteristics. Figure 4 shows the model of the ATLAS sequential Trigger/DAQ system.

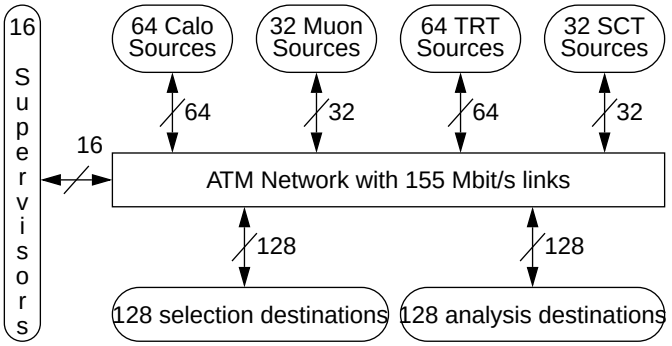

Figure 4: Model of the ATLAS sequential Trigger/DAQ system.

The actual number of ROBs grouped in the sources varies between 4 and 8 for different subdetectors. Each destination in the farms serves 8 processors.

In the paper model we calculate rates for RoI data requests and responses which result from execution of various selection algorithms, such as muon, calorimeter and tracker feature extraction, TRT full scan, etc. These rates are used to estimate the occupation level of ROBs, sources, switching network, destinations, processors and supervisors. The paper model has been evaluated for a sample ATLAS trigger menu at low luminosity. The hardware is far from saturation, except for the TRT ROBs and sources, which suffer from high occupation due to the 4 kHz rate of the TRT full scan algorithm required for B-physics processing [11].

We followed the operation sequence described in

Sections III and IV to evaluate event selection latencies. We estimated execution times for each processing step in all components of the system. We found that the average trigger decision latency for the standard level 2 trigger menu described in [1] amounts to about 10 ms. This average latency increases by a factor of about two when the processors also perform the selection algorithms originally intended for LVL3.

The paper model is not restricted to sequential event selection. In order to compare different event selection strategies we have calculated the corresponding system requirements and performance figures for the parallel processing scheme described in the ATLAS Technical Proposal. Processing power and data transfer requirements for the parallel scheme are at least twice as high as those for the sequential scheme proposed in this paper. On the other hand, the read-out buffer requirements are about three times higher for the sequential scheme; the memory size should not be an issue for the ATLAS Trigger/DAQ system.

Another conclusion of this study is the following: it is advantageous to perform all of the trigger selection algorithms sequentially in a single processor, instead of splitting them into two separate parts (LVL2 and LVL3). This reduces the data transfer requirements by a factor of five.

VI. Simulation Model

The "paper model" allows us to estimate important parameters of the single farm architecture, such as processing power, data bandwidth and trigger decision latency. However, the paper model is based on somewhat simplified assumptions. In particular, the data transfer queueing delays due to contention in the network have not been taken into account.

A C++ simulation model has been developed to study the influence of contention in the network on the performance of the system. It does not deal with individual ROBs and processors, but with sources and destinations. The simulation model allowed us to assess the performance of the ATM network under the expected traffic pattern. In this way it complements the paper model.

The simulation model for the ATLAS Trigger/DAQ architecture is shown in Figure 4. We have modeled the behavior of ATM Segmentation And Reassembly (SAR) interface chips [12], [13] for the source interface to the network. Specific features of the SAR, such as static and/or dynamic bandwidth allocation and servicing priorities, have been implemented. In the model a destination is responsible for sending requests for data via the network and formatting the data received (e.g. reassembling the RoIs). The execution of selection algorithms has not been emulated, because this does not influence the network behavior. At present the supervisors have not been modeled.

The 512 port network is a two stage regular interconnection of 16 x 16 bi-directional switches. Thirty-two switches are used. Each switch is based on a time division multiplex bus with output buffers [14]. The buffer sizes in the switching fabrics are programmable.

A. Physics Input

We used the set of triggers and estimated rates listed in the ATLAS Technical Proposal to derive the rates of trigger RoIs for high luminosity operation. We present them in Table 2, scaled up to account for the maximum level 1 trigger rate of 100 kHz. The rate for the trigger tracks corresponds to the confirmed μ and e/γ trigger RoIs. Table 2 also shows the amount of data for various types of RoIs and the number of sources which contain this data.

Table 2
Rates and data volumes for event selection and analysis

Trigger objects	Rate (kHz)	Size (kByte)	Sources
μ RoI	15.8	1.0	2 Muon
e/γ RoI	73.7	3.0	2-4 Calorimeter
Jet RoI	7.9	3.0	3-6 Calorimeter
Track	31.2	0.3	2 TRT
		0.2	2 SCT
Missing E_T	2.5	16.0	All Calorimeter
Full event building	0.5	1 275	All

The simulation program generates different types of RoIs according to Table 2. Events with an average of four secondary RoIs are produced at a 4 kHz rate. The μ, e/γ and jet trigger RoIs are assigned in a round-robin fashion to the destinations in the event selection farm. The decision to continue a selection algorithm can be taken in any destination. Therefore, TRT and SCT track RoIs, as well as secondary RoIs and missing E_T data are assigned randomly to the destinations. In the simulations we assume that full event building is performed at 500 Hz. These events are assigned in a round-robin fashion to destinations in the analysis farm. Simulation runs correspond to about 5 seconds of LHC operation.

B. Switching Network Performance

During the simulations the bandwidth utilization of each network link is monitored. Most of the event selection traffic (i.e. RoI data) originates from the calorimeter sources (roughly 40 Mbit/s for each of the 64 calorimeter links). The full event building requires an additional 30-45 Mbit/s bandwidth per link. The data traffic from the sources creates a 15% load on the 155 Mbit/s links for the selection farm and a 30% load for the analysis farm.

Concentration of many long event data fragments (about 5-10 kByte) of the analysis traffic towards the same destination may create severe contention in the network. In order to avoid congestion, the rate division traffic shaping technique [3] has been used. This technique can be implemented naturally using ATM by establishing Constant Bit Rate (CBR) virtual connections between the data sources and the analysis destinations. In our simulations each source grants 50% of the available link bandwidth for the full event building traffic. This bandwidth is equally shared between all CBR virtual connections.

The sources establish Variable Bit Rate virtual connections with each destination of the event selection farm. They are used to transport relatively short RoI data fragments (about

256-1024 Byte). The packets on the virtual connections are sent on a FIFO basis and can use up to 100% of the link bandwidth. In addition the sources service the trigger RoI data at a priority higher than that of the secondary RoIs. This model of source avoids blocking the event selection traffic with long fragments of event analysis data. It also minimizes the mutual influence of these two types of traffic and reduces congestion probability in the switching network.

The occupancy of the switching fabric output queues reflects the contention within the network (Figure 5). At present, switching fabrics with buffers of more than 1000 cells per output queue are commonly available [14], [15]. In this case, our simulations predict a cell loss probability of less than 10^{-8}.

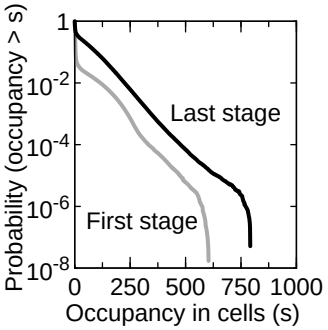

Figure 5: Occupancy of switching fabrics output queues.

On average the cell transfer latency across the network is about 100 μs.

C. Influence of Contention on System Performance

We define the RoI building latency as a time necessary to request RoI data and to gather it into the selection destination. Table 3 presents this latency for several types of trigger regions of interest.

Table 3
Trigger RoI building latencies

Type of RoI	min. (μs)	mean (μs)	99% quantile (μs)
μ	80	135	410
e/γ	200	500	1160
Jet	200	460	1100
TRT track	50	77	340
SCT track	50	83	350

As can be seen, the e/γ RoI building latency amounts to ~200 μs when there is no contention in the network. This value is comparable to what is calculated in the "paper model". In the presence of contention in the switching network, however, this latency is ~2.5 times longer.

The time required to gather all data fragments for a given event into an analysis destination is referred to as the event building latency. We found that it amounts to ~220 ms. This is approximately 3 times longer than the time to transmit the 1 275 kByte of event data over a 155 Mbit/s link. The difference is due to the source organization (50% of link bandwidth for CBR traffic) and the distribution of event fragment sizes.

VII. Summary

In this paper we have proposed a sequential processing strategy for the ATLAS event selection system for use downstream of the first level trigger. It is motivated by the LHC physics requirements. Event selection algorithms are executed step-by-step, requesting data for the next step only if the analysis is still consistent with at least one set of trigger conditions. A decision can be issued at each step and background events can be rejected as soon as possible. We have presented an ATLAS Trigger/DAQ architecture which exploits this concept, based on a single processing farm and a network which transports both data and protocol traffic.

Paper models and computer simulations were used to evaluate the resources necessary for the implementation of this architecture, and to estimate its performance. We have compared different event selection strategies. We see a clear advantage for the sequential scheme over the parallel processing scheme described in the Technical Proposal. It requires significantly less data transfer bandwidth and processing power. The proposed architecture is equally suited for operation at low and high luminosities.

Our simulation studies with ATM networking technology show encouraging results. Specific features of ATM allow the construction of a high performance network capable of transporting simultaneously the various types of traffic specific to this application. The traffic shaping schemes implemented with industrial ATM components minimize the influence of the network contention on the Trigger/DAQ system performance.

More work is needed to refine our models and determine various parameters for the design of the "final" ATLAS Trigger/DAQ system. However, our studies have already helped us to find some bottleneck areas in the single farm architecture and to begin an optimization procedure. We plan to investigate and validate the concepts introduced in this paper on a small scale demonstrator. It will allow us to compare the measurements performed on real hardware with results predicted by calculations and simulations.

VIII. Acknowledgments

Authors would like to thank members of the RD-31 collaboration, especially M. Costa, J.-P. Dufey, M. Letheren and C. Paillard for their contribution to our understanding of ATM technology.

IX. References

[1] ATLAS collaboration, "Technical Proposal for a General-Purpose pp Experiment at the Large Hadron Collider at CERN", CERN/LHCC/94-43, December 1994.

[2] L. G. Guthbert, J-C. Sapanel, ATM - the Broadband Telecommunications Solution, IEE Telecommunications Series 29, ISBN 0 85296 815 9, London, 1993.

[3] D. Calvet et al., "A Study of Performance Issues of the ATLAS Event Selection System Based on an ATM Switching Network", *IEEE Transactions on Nuclear*

Science, Vol. 43, No. 1. February 1996, pp. 90-98.

[4] T. Hansl-Kozanecka, J. R. Hubbard, B. Thooris, “A Detailed Study of the Selection Criteria for the Level 2 Electron Trigger”, ATLAS Internal Note in preparation.

[5] R. Hauser, I. Legrand, “EAST note 94-37: Algorithms in second-level triggers for ATLAS and benchmark results”, ATLAS internal Note DAQ-NO-27, December 1994.

[6] S. Sivoklokov, R. Dankers, J. Baines, “Second Level trigger in the Forward Region of the ATLAS Inner Tracker”, ATLAS Internal Note INDET-NO-111.

[7] J. Bystricky et al., “A Model for Sequential Processing in the ATLAS LVL2/LVL3 Trigger”, ATLAS Internal Note DAQ-NO-55, June 1996.

[8] “PCI Local Bus Specification”, Rev. 2.1, April 1993. Available from PCI Special Interest Group, Portland, OR 97214, USA.

[9] R. Bock, P. Le Dû, “Readout data specifications for modeling a level-2 trigger using regions of interest”, ATLAS Internal Note DAQ-NO-25, December 1994.

[10] J. Bystricky et al., “ATLAS Trigger Menus at Luminosity 10^{33} $cm^{-2}s^{-1}$”, ATLAS Internal Note DAQ-NO-54, June 1996.

[11] M. Smizanska, “Second Level TRT Trigger for B-Physics”, ATLAS Internal Note PHYS-NO-89, July 1996.

[12] IDT Inc., Santa Clara, California, USA, “IDT 77201 NicStar”, User Manual, version 2.0, November 1995.

[13] Transwitch Corp., Shelton, Connecticut, USA, “SARA chip set”, Technical Manual, version 2.0, October 1992.

[14] S. Bush et al., “Switching to ATM”, *Network World Magazine*, February 1994, pp. 37-39.

[15] Newbridge Networks Corp., Ontario, Canada, “ATM Virtual Router Architecture”, April 1994.