ATLAS Internal Note

DAQ-No 67

02 April 1997

A Study of Performance Issues of the ATLAS Event Selection System Based on an ATM Switching Network

D. Calvet, K. Djidi, P. Le Dû, I. Mandjavidze
CEA Saclay, 91191 Gif-sur-Yvette CEDEX, France

M. Costa, J.-P. Dufey, M. Letheren, C. Paillard
CERN, 1211 Geneva 23, Switzerland

A. Manabe
National Laboratory for High Energy Physics, Oho 1-1, Tsukuba 305, Japan

Abstract

Asynchronous Transfer Mode (ATM) is a candidate technology to implement the high performance network in the data collection system for the ATLAS experiment. This work presents the results of modelling and simulation studies which aim at integrating the detailed organization of the detector read-out, the trigger requirements and the capabilities of ATM switching networks.

The status of hardware development of small scale demonstrators is outlined.

I. Introduction

The next generation of High Energy Physics experiments, ATLAS [1] and CMS [2], proposed at the CERN Large Hadron Collider (LHC), will place heavy demands on the data acquisition and on-line filtering systems. A variable portion of the 10^6-10^8 detector channels in those experiments will be fired by tens of interactions created by two bunches of hadrons colliding at a 40 MHz rate. Sophisticated multi-level selection systems will reduce the raw data flow from a few tens of TBytes/s to the several tens of Mbyte/s that will then be recorded on tape for subsequent off-line analysis. A first reduction of this data will be carried out by fast pipe-lined logic that will retain only those events that satisfy some simple geometrical and energy deposition criteria.

After this first level selection the remaining data bandwidth is expected to be of the order of ~1000 Gbit/s. Traditional bus-based data acquisition (DAQ) systems are not adequate to handle this high bandwidth. Several data acquisition conceptual models have been proposed for use downstream of the first level trigger ([1], [2]). The RD-31 project [3] aims at evaluating a new, parallel approach to data acquisition based on the use of standard Asynchronous Transfer Mode (ATM) packet switching technology [4]. This technology holds the promise of becoming a "universal" communication standard, unifying the telecommunications and local area network markets on the time scale of the LHC.

A group of collaborators within RD-31 is involved in the ATLAS experiment. It focuses its efforts on the architecture design and simulation studies adapted to the ATLAS trigger system based on the ATM technology.

In this paper we propose an integrated architecture for the level 2 and level 3 selection and data read-out systems. It is based on the so-called data "Pull" control strategy. We discuss the relative merits of this approach and evaluate its performance by means of simulations.

This paper is organized as follows. Section II describes the principles of the ATLAS event selection and data read-out models. Some of the system bandwidth requirements are presented in section III. The motivations leading to the choice of ATM for our application are given in section IV. Our proposed "integrated Pull architecture" is described in section V. The system that we have modelled and the corresponding simulation results are presented in section VI. The status of two hardware demonstrators is out-lined in section VII. Future plans and a summary are presented in sections VIII and IX.

II. Event Selection and Data Read-out in ATLAS

A. *The Basic Principles*

The ATLAS trigger consists of three logical levels, shown schematically in Fig. 1. Beam crossing interactions occur at a rate of 40 MHz. At the nominal luminosity of 10^{34} cm^{-2}s^{-1}, the input event rate resulting from the level 1 trigger threshold cuts is estimated to be approximately 30-40 kHz. A safe design value of 100 kHz has been adopted. The level 1 trigger is without deadtime, because all the data are pipelined during the fixed 2 μs latency needed to decide whether to accept or reject the event candidates.

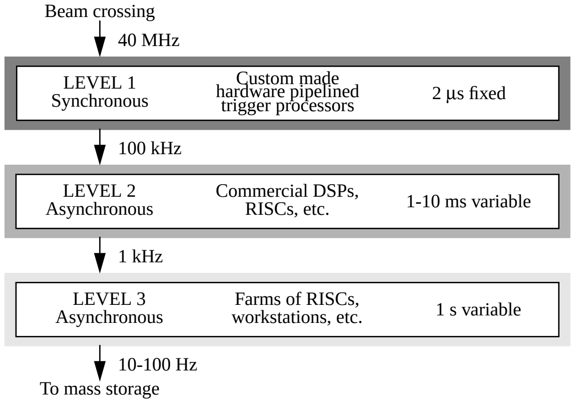


Fig. 1. The ATLAS multi-level trigger selection scheme.

The event rate that can be recorded on tape is estimated to be in the range 10-100 Hz. Thus a further reduction of the order of 10^3-10^4 is necessary. This important rejection factor can be achieved by using two more logical sequential steps

Presented at IEEE Conference on Real Time Computer Applications, Michigan State University, East-Lansing, 24 - 28 May 1995, also in IEEE Transactions on Nuclear Science, vol. 43, No. 1, February 1996.

with variable decision latency. A rejection factor of ~100 is expected to be achieved by the level 2 selection. It can have access to the full granularity data as well as to new information from other subsystems that cannot be available in time for level 1. In order to reduce the decision latency and the data transfer requirements, only data belonging to “Regions of Interest” (RoI) are transmitted to the level 2 processors, thus representing less than 2% of the front- end information. A list of pointers to the RoIs is provided by the level 1 trigger system for each event. First, the RoIs within an event are processed individually to identify particle candidates such as electrons, jets or muons. Then a topological analysis of the event is performed by the combination of the previously identified particles. The trigger decision is issued accordingly.

An additional event rejection factor of ~10 is expected to be achieved by the level 3 trigger that executes sophisticated algorithms and selects events on the basis of physics signatures. The level 3 system should provide access to the complete event data in order to perform full event analysis similar to that applied off-line. Events accepted by the level 3 selection are recorded on tape for subsequent off-line studies.

B. *Data Flow*

The logical organization of the data flow for the read-out and the level 2 (LVL2) and level 3 (LVL3) triggering systems is depicted in Fig. 2. For each bunch crossing the signals from all subdetectors are stored locally in pipeline memories (digital or analog) during the level 1 processing. For events accepted by level 1 (LVL1) the data from all the detector front-end memories are transferred via optical links to about 2000 read-out cards located in the counting room. The read-out cards contain a data buffer to store events during subsequent triggering steps. In addition they possess enough processing capabilities to preprocess and format data for the level 2 and level 3 selection systems. The data is transmitted to those systems via a dedicated port of the read-out cards.

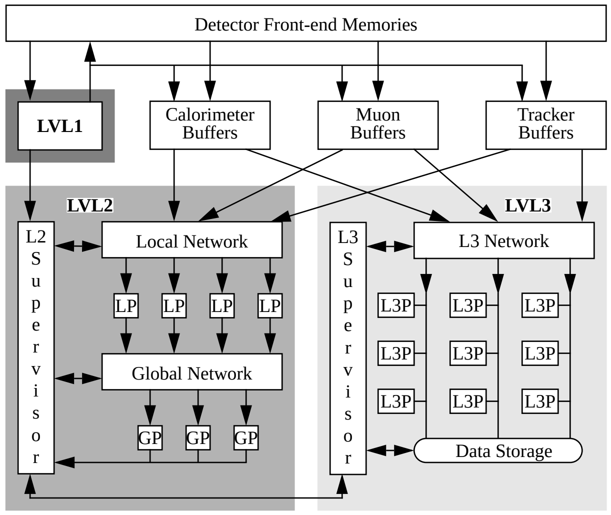


Fig. 2. The logical model of the read-out data flows.

The data transferred to the LVL2 system corresponds to the regions of interest selected by the LVL1 trigger. It is estimated that an event will contain 5 RoIs on average. The RoIs within each subdetector are processed individually and in parallel. One Local Processor (LP) per RoI per subdetector is allocated (see Fig. 2). In a given subdetector the data for a RoI can be spread across several read-out cards. Therefore a network providing local connectivity (Local Network, Fig. 2) is needed to gather the RoI data into the local processor. However, full connectivity at this stage provides more flexibility, better load balancing among the local processors [5] and might be simpler.

For each event the results of the local processing (referred to as “features”) are combined in a Global Processor (GP) via a Global Network for the subsequent topological analysis (Fig. 2). The Global Processor issues the LVL2 trigger decision. The LVL3 selection starts for the accepted events. Rejected events are discarded.

The LVL3 selection algorithm may require the complete event data from all subdetectors. The duration of the algorithm is estimated to be of the order of 0.1-1 second. Therefore, a farm of processors is needed in order to cope with the expected 1 kHz LVL2 rate since one processor is allocated per event. The necessary connectivity between read-out cards and the processing farm is provided by the L3 network (Fig. 2).

The control and management of the LVL2 and LVL3 systems is performed by the L2 and L3 supervisors respectively.

III. Bandwidth Requirements

The expected average event data sizes for all subdetectors is given in Table 1. The table also shows the aggregate bandwidth requirements for data transmission to the LVL2 and LVL3 systems. This corresponds to level 2 and level 3 input rates of 100 kHz and 1 kHz respectively.

Table 1: Estimated bandwidth requirements

Subsystem	Event Size KBytes	L2 Bandwidth Gbit/s	L3 Bandwidth Gbit/s
Tracker	770	10	6
Calorimeter	400	5.6	3.2
Muon chamber	200	2.4	1.6
Total	1 370	18	10.8

It can be seen that the aggregate bandwidth required for the triggering system is of the order of several tens of Gbit/s. This high bandwidth cannot be handled by the traditional bus-based data acquisition systems. It is expected that switching technology will allow implementation of the high performance, cost-effective and expandable network, required for this challenging application. However, building this high performance network is not a trivial task. In order to reduce the overall development phase and facilitate maintenance of such a complex system, it is desirable to use commercially available components wherever possible. Compliance with widely adopted industrial standards ensures the inter-operability (software and hardware) of equipment from various vendors.

The Asynchronous Transfer Mode (ATM) is the main candidate technology currently evaluated within the RD31 collaboration [3].

IV. Asynchronous Transfer Mode

The International Telecommunication Union's standardization body, the ITU-T (formerly known as the CCITT), has recommended the use of ATM as the switching technology for the future broadband integrated services digital network (B-ISDN). The ITU's B-ISDN standards [7] were originally targeted at telecommunications and wide area networking (WAN) applications.

ATM technology is also being adopted for high-performance local area networking (LAN) applications, and all major workstation companies are actively engaged in developing the technology (typical activities are the development of LAN hubs based on ATM switches, ATM interfaces to workstations, and the implementation of the internet TCP/IP protocol over ATM). Efforts in this area are coordinated by an industry association, the ATM Forum [8], which parallels the ITU's standardization efforts, while focussing on the needs of the workstation/LAN industry.

It appears likely that ATM technology will dominate both high-performance WAN and LAN networking throughout the time-span of experiments at the LHC. The growth of multi-media applications and the adoption of ATM by the more cost-competitive LAN industry suppliers are expected to render ATM affordable on the time scale of the LHC.

V. Proposed Architecture

A. *The Principles*

Currently several different implementations of the proposed read-out scheme (Fig. 2), based either on "Push" or on "Pull" strategies, are under evaluation within the ATLAS collaboration. In the "Push" approach the sources send their data to the processors as soon as the data is ready and the destination processor entity is known. This scheme implies that sources must know, prior to the execution of the selection algorithm, which data will be needed by processors. Therefore, for LVL2 all RoIs from all subdetectors should be sent to the local processors and examined in parallel, even if this is not necessary. By contrast, in the "Pull" strategy the destination processors request data from the sources as it is needed. This approach allows the implementation of a sequential LVL2 algorithm: the RoI data for the particular subdetector is sent to the local processors only if it is required by the subsequent steps of the algorithm. The sequential steps of the level 2 selection can significantly reduce requirements for the switching network aggregate bandwidth and processing power.

At present it is not decided whether the full event data will be presented to the LVL3 or whether the selection will be based on partial event data. It should be mentioned that in the LVL3 triggering system based on the "Push" approach, the full event data will always be sent to the destination processor. The event flow control in this case might be relatively simple at the expense of a higher demand on the aggregate bandwidth of the switching network. On the other hand, the architecture, based on the "Pull" strategy will allow implementation of both partial read-out and/or full event building schemes.

In any DAQ architecture, control information should be exchanged between various parts of the system (e.g between the data sources and the destination processors, etc.). The control information can either pass via a dedicated network, or it can use the same network as the one used for data transmission. The main advantages of the second approach are the simplification of using a common medium for all types of traffic (LVL2 / LVL3 data and control) and the requirement of a single network adapter per node.

B. *An Integrated "Pull" Architecture*

In this section we describe a possible LVL2 / LVL3 selection and data read-out system based on the "Pull" principle and a single physical network. This physical network will support the "local", "global" and "L3" networks of the logical model shown in Fig. 2.

We assume that the information on each event accepted by level 1 (number of RoIs and their position in the (η, ϕ) coordinate system[1]) will be delivered to the L2 supervisor via a dedicated path. One of the tasks of the supervisor is to allocate processing resources for this event, e.g. assign a local processor per RoI and a global processor for the LVL2 decision. Currently we propose a very simple destination processor assignment scheme: we estimate that simple sequential allocation is adequate [5]. More sophisticated algorithms are not excluded.

The control and data flow in the proposed system is shown in Fig. 3.

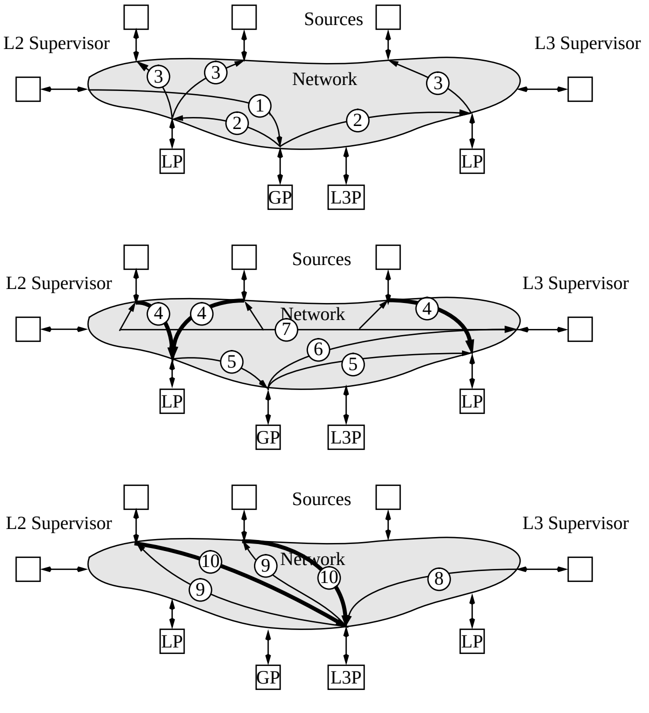


Fig. 3. Possible scheme for the ATLAS LVL2 and LVL3 selections.

The global decision processor receives a notification message from the L2 supervisor (the flow labelled 1 in Fig. 3). This message contains the event ID, a list of RoIs and a list of IDs of the local processors assigned to these RoIs. The global processor sends a notification message to each local processor allocated for this event (flow 2). The message contains the event ID, a RoI ID and the Global Processor ID, etc... Therefore, global and local processors know their respective partners for the event, e.g. the global processor knows from which local

[1] η axis: direction parallel to the beam; ϕ axis: direction radial to the beam

processor it has to expect features data. This can simplify error detection and recovery.

From the RoI ID information the local processor knows which sources contain data for the particular RoI (deduced by table look-up in the L2 supervisor, the global processor or even in the local processor itself). The local processor will send a request message to each source concerned (flow 3). In response to message (3), sources send the requested data (flow 4) to the local processor after preprocessing (if needed) and formatting. It is proposed that when all data for a given RoI have been delivered to the local processor, it starts to execute the feature extraction algorithm. Features are sent (flow 5) to the global processor. The global processor executes an algorithm based on the collected features for the event. Note, that neither the local processors nor the global processor need to idle while waiting for the data. For example, they can work on the previous events. A LVL2 "Yes/No" decision is issued when the global algorithm completes.

We consider two possibilities for the treatment of the level 2 decisions. In one case, the sources are notified only if the event has been accepted by the LVL2 selection. Only the LVL2 decision "Yes" is sent to the L3 supervisor (flow 6), which then multicasts it to all sources (flow 7). No immediate action is taken for the events which did not pass the LVL2 selection. The oldest event is simply overwritten in the source buffer when a new event is read from the front-end modules. This scheme is attractive because it does not generate unnecessary traffic in the network (99% of the LVL2 decisions are expected to be "No"), it simplifies control logic in the data sources and requires less actions per event in the system. An alternative solution is to send both LVL2 decisions "Yes" and "No" to the L3 supervisor (flow 6).

It is possible, that the same processor, which performed the LVL2 global decision will continue to work on the LVL3 selection for the event (because it already possesses a substantial information about it). Another option is to use different processors for those tasks. In this case the L3 supervisor allocates a processor for the LVL3 selection (flow 8). As for the LVL2, the allocation algorithm can be either simple sequential (e.g. round-robin) or can use a more sophisticated discipline.

The allocated L3 processor sends request messages to the concerned sources (flow 9). In response, sources send requested data (flow 10) to the L3 processor after preprocessing (e.g. zero suppression) and formatting. When the required data is available, the L3 processor executes the LVL3 selection algorithm. For accepted events, if necessary, the remaining part of the event data is collected prior to writing the event to mass storage. Rejected events are discarded.

As can be seen from the above description, several different types of traffic are transported by the network. Each of them has its own properties. For example, control traffic uses short messages; the LVL2 data has to be delivered at high rate; the level 3 traffic requires continuous concentration of data flow from many front-end sources to a destination processor, etc.

The ATM technology has been designed to carry simultaneously various types of traffic having different service requirements (e.g. real-time video, audio, data) on a common physical medium. Therefore, we propose to investigate whether ATM is adequate in our application to handle efficiently the LVL2 / LVL3 data and control traffic.

VI. Simulation Studies

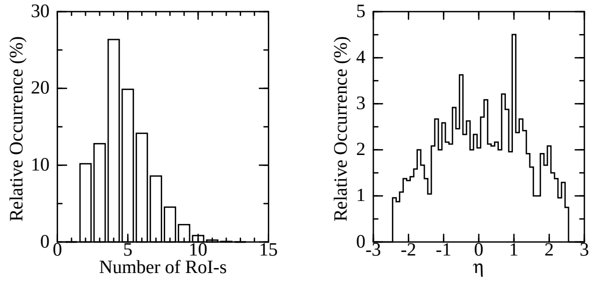
The ATLAS detector is composed of three main sub-systems: calorimeter, muon chambers and tracker (see Table 1). We decided to concentrate our simulation efforts on the LVL2 / LVL3 trigger for the calorimeter sub-system because the ATLAS Saclay group is strongly involved in calorimetry; therefore we had direct access to relevant information concerning the physics and the read-out organization of the calorimeter.

A. Physics

As previously mentioned, the ATLAS triggering system is based on the concept of regions of interest (RoI). The LVL2 selection uses only data from RoIs. The number of RoIs within events and their properties (size, amount of data, etc.) are important parameters for the design of the overall triggering system. For example, the system described in Fig. 2 may not be practical if the number of RoIs within events is very large (no reduction of the required aggregate bandwidth) or if it is too small (no cost-effective gain from parallel processing of the RoIs).

Extensive Monte Carlo simulation studies have been made by physicist groups in order to evaluate trigger performance [6]. A sample of ~1000 di-jet events which passed LVL1 electron / gamma selection has been produced. Di-jet events are expected to give the largest contribution (~60%) to the level 1 trigger rate. Samples of events which passed other LVL1 selection criteria (muon, jet and missing energy triggers) have been produced and are currently under analysis.

The distribution of the number of RoIs per event, shown on Fig. 4.a, has been derived from the analysis of those events. It can be seen that each event contains an average of 5 RoIs. However, the maximum number of RoIs can be as high as 12. It should be mentioned that those parameters depend on the thresholds used at the LVL1 selection.



a) Distribution of the number of RoIs

b) Distribution of e/g RoIs η axis

Fig. 4. Some characteristics of physics events.

For each RoI, the LVL1 trigger indicates its geographical position in the η, ϕ coordinates, its type (electron/gamma, jet, muon, etc.) and possibly some other information. The distribution of electron / gamma RoIs along the η coordinate (direction parallel to the beam axis) is depicted in Fig. 4.b. The detector occupation is higher in the central region (Barrel) than in the extremities (left and right End-caps). The distribution of RoIs along the radial (ϕ) coordinate is flat (not shown). The amount of data depends on the type of RoI and the granularity of the detector segmentation. The calorimeter subdetector consists of four different parts: a PreSampler (PS), an ElectroMagnetic (EM), a Hadronic (HAC) and an Integrated

Forward (IFC) calorimeter. Table 2 shows the expected data volumes for different types of RoIs.

Table 2: Data volumes for different RoI types in the calorimeter barrel.

ROI type	RoI Size (Δφ x Δη)	EM + PS data (Byte)	HAC data (Byte)
Muon	0.4 x 0.4	128	96
Elect./Gamma	0.3 x 0.3	1080	54
Jet	0.8 x 0.8	512	384

For simplicity Table 2 presents data volumes only for the barrel part of the calorimeter. Due to the changes of the segmentation granularity and other irregularities (overlaps), the amount of data per RoI is variable. This has been carefully modelled.

B. *Calorimeter Read-Out Organization*

According to the read-out scheme proposed in [1], for each event accepted by level 1, data are transmitted from the calorimeter (PS, EM, HAC, IFC) front-end boards to the read-out cards. Several such cards will be housed in a crate. In our model we assume that there are 26 crates of 16 cards for the calorimeter:

- 16 crates for the PS and EM calorimeters,
- 8 crates for the Hadronic calorimeter,
- 2 crates for the IF calorimeter.

We consider that one link per crate is used to transmit the data from the read-out cards to the LVL2 and LVL3 triggering system. At present, the necessary connectivity inside of a crate is provided by a back-plane bus. The simulated architecture is presented in Fig. 5. As can be seen the model comprises:

- 26 source modules,
- 16 farms of local processors,
- 14 farms of LVL2 global and LVL3 processors,
- L2 and L3 supervisors,
- An ATM switching fabric with 64 bidirectional ports.

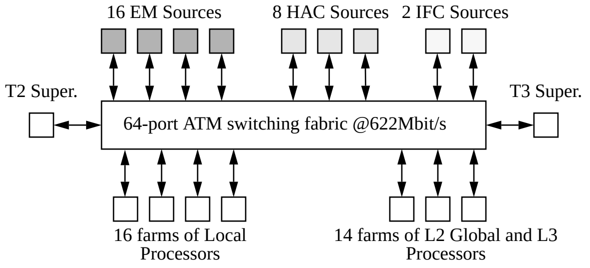

16 EM Sources
8 HAC Sources
2 IFC Sources
T2 Super.
T3 Super.
64-port ATM switching fabric @622Mbit/s
16 farms of Local Processors
14 farms of L2 Global and L3 Processors

Fig. 5. Model of the calorimeter subsystem.

We next describe each element of the model in more detail.

B.1. *Sources*

We first evaluated the bandwidth required to transfer LVL2 and LVL3 data from the source modules to the destination processors. For the LVL2 bandwidth estimates, we assumed a 100 kHz level 1 trigger rate, 5 RoIs per event and ~1 Kbyte data per RoI. The average required bandwidth for each EM/PS source then equals ~250 Mbit/s. For each HAC source it amounts to ~200 Mbit/s. The IFC sources do not participate in the LVL2 selection.

For LVL3 data, assuming a 1 kHz L2 trigger rate and a full event read-out, the bandwidth requirements are ~200 Mbit/s for each EM source and ~20 Mbit/s for each HAC and IFC source.

This bandwidth estimation indicates that a 622 Mbit/s link would be adequate to carry LVL2 / LVL3 data for EM and HAC sources. Even though slower links could be used for IFC sources, our model uses a 622 Mbit/s link for all sources.

For the source link interface, we have modelled the behavior of an ATM Segmentation And Reassembly (SAR) interface chip set, available from industry ([9], [10], [11]). Specific features of the SAR, such as static and/or dynamic bandwidth allocation and servicing priorities, have been implemented.

In the case of the level 3 full event building, data fragments for LVL3 are ~10 times bigger than RoI data for LVL2. In order to guarantee fast servicing times for the level 2 data and protocol packets, we intend to use a higher priority for these types of message. A lower priority will be assigned to the LVL3 data. The funneling of the large LVL3 data packets towards a destination processor induces severe contention in the switching network. This contention can be reduced by an appropriate bandwidth allocation scheme, as described below. The model of a source is shown in Fig. 6.

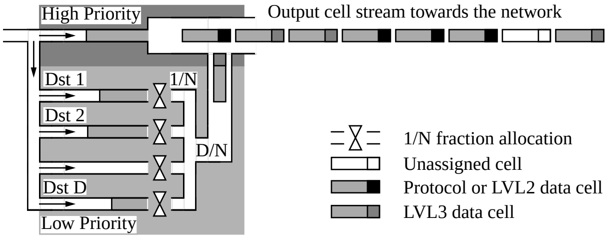


Fig. 6. Logical model of a source.

The source maintains the necessary number of semi-permanent Virtual Connections (VC), providing a connection path to all farms of local processors. The VCs are associated with a high priority logical queue. They are serviced in FIFO order at a full link bandwidth. The protocol and LVL2 data (i.e. AAL5 packets) are placed in this high priority queue.

Lower priority queues, dedicated to the LVL3 data, are serviced whenever the high priority queue is empty. The source manages D low priority logical queues. Each queue contains the event fragments to be transferred to one of the D level 3 destinations. A semi-permanent VC to a destination is associated with each queue. Rate control is used in the sources in order to limit the traffic on each virtual connection so that the aggregate bandwidth of all traffic to a given destination does not exceed the available bandwidth of an output port. A rate control mechanism ensures that one cell is read from the head of each logical queue periodically. The period for servicing the logical queues can be chosen to be N times the cell transmission delay (i.e. 0.68 µs @ 622 Mbit/s), where N is a programmable parameter (N>=D). The fraction of the available bandwidth allocated per VC is 1 / N. Therefore, the peak LVL3 bandwidth per source is D / N of the 622 Mbit/s link rate.

The required functionality of a source can be implemented with available industrial ATM components.

B.2. *Destinations*

The same destination model has been used for the farms of local processors and global/LVL3 processors. A destination contains a master unit and several processors. The master unit is responsible for:

- sending requests for data via the network,
- formatting the received data (e.g. reassemble RoI),
- distributing the formatted data to the processors,
- handling the results of the processors.

The processors in a farm perform the actual execution of the appropriate algorithm (feature extraction for RoI, Global LVL2 decision, etc.). If all processors in a farm are busy, events ready to be processed are queued.

Assuming a 100 kHz level 1 trigger rate, 5 RoIs per event, 128 µs feature extraction algorithm duration and 50% processor occupancy, 128 local processors are needed. Our model contains 16 farms of 8 processors. At present we do not model the LVL2 Global and LVL3 algorithms execution, because we simulate only one subdetector.

B.3. *L2 and L3 supervisors*

The main tasks of the supervisors were described in section V.B. Our estimates show that ~3 ATM cells are needed to deliver the list of RoI pointers and allocated processors to a global processor. Assuming a 100 kHz level 1 trigger rate, the corresponding average bandwidth is 125 Mbit/s. In our model 622 Mbit/s links were used to connect the L2 and L3 supervisors to the network.

B.4. *Switching fabric*

The switching fabric is a regular interconnection of switching elements. The topology can be either *Banyan* or *Omega*. Switching elements can be of variable sizes *(2x2, 4x4, etc.)*. Internal contention resolution can be selected from one of the following methods:

- *shared media with no link-level flow-control*
  (Fore Systems type [12])
- *shared memory with no link-level flow-control*
  (Alcatel HSS type [13])
- *output queueing with link-level flow-control*
  (AT&T Phoenix type [14])
- *shared memory with link-level flow-control*
  (IBM Prizma type [15])

The buffer sizes in the switching elements and the bit-rates of the fabric's external and internal links are programmable.

Semi-permanent virtual connections are used to provide the required connectivity in the system. The connections are not established dynamically. This avoids the complexity of signalling and admission control. At present switching fabrics supporting up to 4,000 VCs per link are available [16].

C. *Simulation Results*

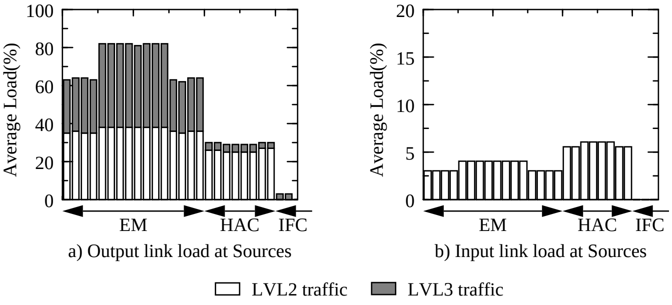
Two completely independent simulation programs have been developed in concurrent object oriented languages, Modsim [17] and µC++ [18]. The same set of input parameters was used for both programs to cross-check results. The results obtained from the two different codes agree within ~1%.

Our queueing models do not take into account various overheads (e.g. processor I/O, software, etc.). We plan to refine our models with the measurements performed on the hardware demonstrator systems (see, section VII). However, we believe that our models are adequate to evaluate the performance of a single network when it is used to carry both data and protocol traffic for the LVL2 and the LVL3 systems. They allow us to study interference between the two types of traffic in the system, and to evaluate methods to minimize it. The ability of ATM networks to carry various types of traffic, specific to our application, and the influence of fabric architecture have been investigated.

Performance evaluation of the LVL2 and LVL3 triggering systems requires to pass a large number of events through the simulation program to accumulate enough statistic (one LVL3 accepted event corresponds to ~1000 initial LVL1 events). At present, the number of LVL1 accepted events available from the Monte Carlo studies is ~1000. To rapidly produce large sets of events, we developed an event generator which possesses characteristics similar to those of the physics events (number of RoIs, their distributions, etc.). In our simulations 50,000 level 1 accepted events passed through the system. This corresponds to ~0.5 seconds of the LHC operation. In what follows we present our simulation results. Unless otherwise specified, all simulation results correspond to an average of 100 kHz LVL1 and 1 kHz LVL2 Poisson distributed trigger rates and LVL3 full event building.

C.1. *Network bandwidth utilization*

During the simulation the load on each link of the network is monitored. Figure 7 shows the bandwidth utilization for each source. As can be seen on fig.7.a, the LVL2 data traffic (RoI) from the 26 sources to the 16 local processor farms requires ~35% of the sources' 622 Mbit/s output link bandwidth. This traffic creates ~50% load on the destinations' input links.



a) Output link load at Sources

b) Input link load at Sources

Fig. 7. Source link utilization.

The LVL3 data traffic adds at most another ~30% load on the sources' output links and requires ~45% of the global / LVL3 processors' input link bandwidth. Due to the different granularity of various parts of the calorimeter, the contribution of EM sources to the LVL3 data traffic is significantly bigger than that of the HAC and IFC.

The traffic which delivers features from the local processors to the global processor uses ~19 Mbit/s of the bandwidth of the global processors' input link and increases their load up to 50%. The request messages for the RoI and LVL3 data use less than 5% of the available bandwidth of the source input links (Fig. 7.b). On average 25% utilization of the available switching fabric aggregate bandwidth is observed.

C.2. The rate division technique for level 3 traffic

In our simulations, the event data fragments for LVL3 are ~10 times bigger than RoI data fragments. Therefore, if each source segments LVL2 and LVL3 data packets sequentially, the RoI data can be blocked during a long time while waiting for a LVL3 packet transmission to terminate. When segmentation of the LVL2 and LVL3 packets was not interleaved, we observed an unacceptably high latency for the LVL2 traffic. However, the ATM technology allows concurrent segmentation of packets belonging to different virtual connections. Therefore, in our model, cells carrying the LVL2 data can be interleaved in the cell stream of a LVL3 packet.

Furthermore the concentration of many long LVL3 cell streams towards the same outlet creates severe contention inside of the switching fabric. This introduces long latencies for LVL3 traffic. If there is no support for different levels of routing priority inside of the switching fabric, control traffic and LVL2 traffic will also be affected. In order to prevent the sources saturating the switching fabric with LVL3 data, a rate division technique is used for the LVL3 traffic.

This technique will be described for a system with S sources and D destinations. We consider only LVL3 event building traffic. As was presented in section VI.B.a, each source maintains a semi-permanent virtual connection for each L3 destination. A programmable fraction of the available link bandwidth is allocated to each virtual connection. In our application LVL 3 events are evenly distributed among destinations. Therefore all the D virtual channels within a source should be granted an identical bandwidth. The sum of bandwidth for all VCs in a source cannot exceed the link bandwidth at the input to the switch. Hence, the average fraction of the available bandwidth used by any VC will not exceed 1/D. Our model includes 26 sources and 14 L3 destinations. Therefore 1/14 of the 622 Mbit/s link rate can be allocated to each VC (i.e. ~44 Mbit/s) within a source.

As many sources concurrently send event data to the same destination, on average the sum of their traffic contributions cannot exceed the available output link bandwidth. The simplest scheme is to allocate the same fraction of bandwidth to each virtual connection in the system, provided that it does not exceed 1/D. For a system with S sources and D destinations, this fraction would be 1/S. This guarantees that the output links will not be saturated. Therefore, for our model with 26 sources and 14 L3 destinations, only 1/26 of the 622 Mbit/s link rate (i.e. ~24 Mbit/s) can be granted to each VC within a source. In this case 14 * 24 = 336 Mbit/s will be allocated in each source.

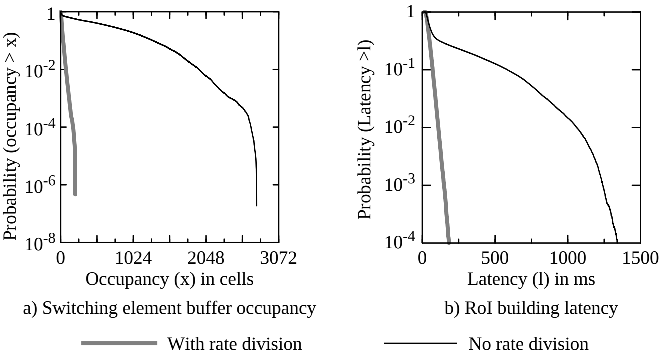
This equal bandwidth allocation scheme is adequate if all sources have approximately the same amount of data to send. However, it can be seen from Table 3 that our system is very un-balanced. The EM sources have to send event data fragments ~10 times larger than others.For a 1 kHz LVL2 trigger rate, each EM barrel source needs at least 250 Mbit/s to send its data. The required bandwidth for each VC in the different sources is also given in Table 3. For example, each VC for EM barrel sources needs at least 250 / 14 = 18 Mbit/s. On the other hand, VCs in HAC sources only need 1.5 Mbit/s. The equal bandwidth allocation scheme discussed above will assign 24 Mbit/s to all VCs. Hence EM barrel source VCs will operate at 18 / 24 = 0.75 load. If possible, it is desirable to allocate more bandwidth for the VCs in EM sources.

Table 3: LVL3 data volumes and bandwidth requirements

	8 EM Barrel Sources	8 EM End-cap Sources	8 HAC Sources	2 IFC Sources
Data per source (KBytes)	31	20	2.6	1.8
Required source bandwidth (Mbit/s)	250	160	21	15
Required bandwidth per VC (Mbit/s)	18	11.5	1.5	1.1
Allocated bandwidth per VC (Mbit/s)	34	22.5	4.2	4.2

One can try to perform load balancing by distributing the available output link bandwidth among sources in proportion to their relative contributions. The total amount of data corresponding to calorimeter full event building is ~430 kbytes per event. One EM barrel source event fragment is ~31 kbytes (i.e. 7.2% of the total). The corresponding allocated bandwidth could be 622 * 0.072 = 45 Mbit/s. However, since the LVL3 events are evenly distributed among the 14 destinations, the average fraction of the available bandwidth used by any VC cannot exceed 622 / 14 = 44 Mbit/s. The bandwidth allocation used in our simulations is shown in Table 3. It can be seen that each virtual connection uses ~50% of its allocated bandwidth. One can check that the allocated bandwidth for each input and output link does not exceed its capacity.

We compared the system performance with and without the rate division technique. In both cases, a 2-stage Omega network composed of 8x8 switching elements has been used to model the 64-port ATM network. The switching elements operate with output queues configured as a dynamically shared memory [13]. The buffer occupancy of switching elements reflects the contention within the fabric. Figure 8.a shows the *tail distribution* of the occupancy of the shared buffer memory in the switching elements (the tail distribution indicates the probability of buffer overflow as a function of the switching element buffer size).



a) Switching element buffer occupancy

b) RoI building latency

Fig. 8. Switching element buffer occupancy.

It can be seen that the rate division technique significantly reduces internal buffer occupancy and contention in the fabric. In switching fabrics in which there is no hardware link-level flow control mechanism the internal buffers can overflow; in this case cells will be lost. At present, switching elements with a shared memory large enough to buffer 256 cells are commonly available. For that case, our simulation predicts 10^{-8} cell loss probability if the rate division technique is used.The time necessary to gather RoI data, distributed among several sources, into a local processor is referred to as RoI building latency. The probability that this latency exceeds a given value (the tail distribution) is plotted in Fig. 8.b for the two different cases considered. The average RoI building latency amounts to 58 µs and 183 µs with and without rate division respectively.

The time required to gather all LVL3 event data fragments into a L3 processor is referred to as event building latency. Longer event building latencies (average of 18 ms compared to 13 ms) have been observed when no rate division was applied. In this case, in order to reduce contention in the fabric, the L3 processor requested LVL3 data fragments one by one from each source sequentially. If no rate division is applied and if all sources send their data to the L3 processor simultaneously, the system is immediately saturated.

C.3. Influence of the LVL3 traffic on the LVL2 traffic

As was mentioned in the previous section, a degradation of the LVL2 performance is observed when no rate division is applied for the LVL3 traffic. We investigated the influence of the LVL3 traffic on the LVL2 traffic when this technique was used. The switching fabric was a 6-stage Banyan network of 2x2 switching elements with hardware link-level flow control mechanism (AT&T like [14]).

The RoI building latency tail distribution is plotted in Fig. 9 for three different cases.

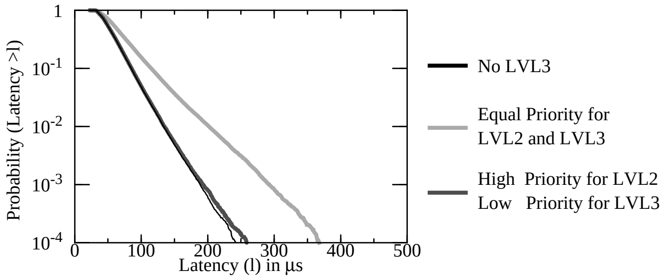
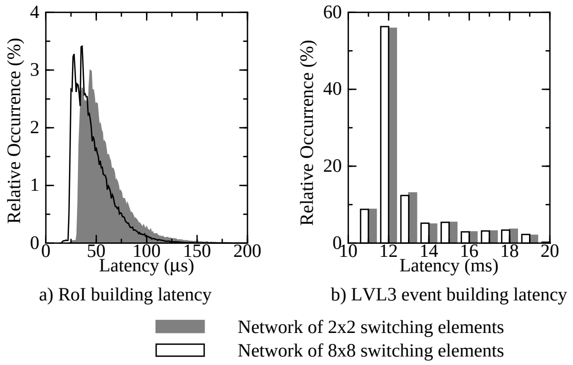


Fig. 9. Influence of L3 traffic on RoI building latency.

In the first case on fig.9 the LVL3 traffic is not present in the system and the average RoI building latency amounts to 58 µs. When the rate division technique is applied to the LVL3 traffic, a 25% increase of the average latency has been observed. However, if the LVL2 data is serviced at a higher priority in the sources, the LVL3 traffic has no significant influence on LVL2 traffic.

C.4. Influence of the architecture of the switching fabric

We conducted simulations with various types of switching fabrics. Figure 10.a shows the RoI and LVL3 event building latencies for two different switching fabric types. Curve 1 corresponds to the 6-stage Banyan network of 2x2 switching elements. Curve 2 relates to the 2-stage Omega network composed of 8x8 switching elements. As can be seen, the shapes for the RoI building latency distributions are identical. The observed shift of ~10 µs is due to the longer cell transfer time through the 6-stage fabric.



a) RoI building latency

b) LVL3 event building latency

Fig. 10. Sensitivity to internal architecture of switching fabric.

Figure 10.b shows that the event building latency for the two types of switching fabrics is identical. It amounts to 13 ms on average. Our simulations show that, when the rate division technique is applied, the internal architecture of the switching fabric has minor influence on the performance of the system.

C.5. Push vs. Pull

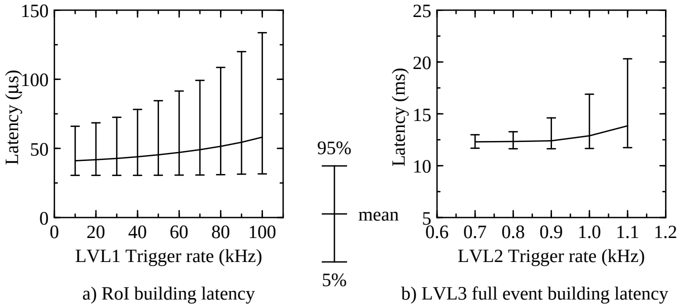
We have simulated the previously described data flow control strategies, namely “Push” and “Pull”. In the push approach, the sources send RoI and LVL3 data to the allocated processors as soon as they become available. In our model, the time required to distribute the necessary information (RoI pointers, allocated processors) to the sources was not taken into account.

In the pull approach, the local and L3 processors request the necessary data from the sources when needed. The transfer delay for the protocol traffic through the fabric has been modelled. Our simulations indicate that the average latency introduced by the network for the request messages amounts to ~8 µs.

The average RoI building latency for the pull approach is ~10 µs more than that for the push data flow strategy. The difference is due to the latency of the request traffic. The average event building latencies for both cases are identical, since the influence of the protocol traffic delay is negligible.

C.6. Influence of trigger rates.

We have evaluated the system performance for various LVL1 and LVL2 trigger rates. The simulation results are shown on Fig. 11.



a) RoI building latency

b) LVL3 full event building latency

Fig. 11. Influence of trigger rates.

In one case, the LVL1 trigger rate was varied while the average LVL2 rate was kept constant at 1 kHz. As can be seen from fig.11.a, the system behavior is satisfactory up to the targeted 100 kHz LVL1 trigger rate.In the second case, the average LVL1 trigger rate was kept constant at 100 kHz while the LVL2 trigger rate was varied. Figure 11.b shows that, even for the LVL3 full event building, the system is not saturated for LVL2 trigger rate up to 1.1 kHz (targeted rate is 1 kHz).

VII. Hardware Demonstrators

A general purpose hardware demonstrator system based on the Alcatel multi-path self routing switching fabric [16] is under evaluation. The switch has been delivered together with embedded operations and management software and an operator interface which runs on a SUN workstation. The demonstrator includes several data generator [19] sources and the destination VME/ATM interfaces [20] developed within the RD31 project. A Hewlett Packard broadband test system [21] is used for ATM protocol validation and also allows performance measurements and comprehensive error stressing. The inter-operability of all the components constituting the demonstrator system has been successfully tested. The software protocol layers and management functions, required for event building, are currently being developed and tested. Some preliminary measurements of the software protocol overhead have been made [3].

A second demonstrator based on the Phoenix AT&T switch [14] is foreseen. At present, we have installed two SBus/ATM [22] and one VME/ATM [23] interfaces. We evaluate performance of the interfaces at various protocol levels (e.g. LAN emulation, AAL5 layer) and under real-time requirements. An 8-port AT&T switching fabric will be delivered soon.

VIII. Future Work

We will continue the architecture design and simulation studies of the ATLAS trigger system. We plan to extend our simulation code to include models of other subdetectors. More Monte Carlo simulated physics events are under production. We are going to use those events to generate a more realistic data traffic pattern in our simulations. Realistic trigger algorithm processing times will be included.

The demonstrators will allow us to compare the measurements performed on real hardware against the results predicted by simulation and to refine the models. We plan to investigate and validate the concepts introduced in this paper on our demonstrators.

IX. Summary

The ATLAS Collaboration proposes to built a general-purpose proton-proton detector which is designed to exploit the full discovery potential of the Large Hadron Collider (LHC) at CERN (Geneva).

Asynchronous Transfer Mode (ATM) packet switching network technology has been proposed as the interconnect for building high-performance data acquisition architectures for future physics experiments.

Based on the event selection and data read-out strategies for the ATLAS detector, several possible DAQ architectures incorporating ATM switches are described. An ATM network linking several thousand front-end memories and processing elements would be required.

The expected volume of data produced under the nominal luminosity operation of LHC has been evaluated and the required system aggregate bandwidth for data traffic flow has been estimated as several tens of Gbit/s.

Building a high performance network, that will operate under the very specific traffic pattern for this application is not a trivial task. The burstiness of the traffic and continuous concentration of data flow from many front-end sources to the destination processors creates severe contention inside the network. This can induce undesirably long latencies and in extreme case can result in data losses due to buffer overflow in various parts of the switching fabrics.

Designing such a complex system requires a good understanding of its behavior and of the behavior of various main components, such as front-end memories, processors and the network.

Modelling activities are essential as they provide the main method for evaluating and understanding the performance of large DAQ architectures. The development of demonstrators is an activity which allows to gain experience with technology and validate some of the results obtained from simulations.

We have developed models of generic and existing industrial ATM switches. The switching fabrics are formed by a regular interconnection (Banyan or Omega) of elementary switching nodes. Currently modelled nodes can resolve contention in several programmable ways: output queuing (AT&T/Phoenix like), buffer sharing with output queuing (IBM/Prizma like), shared media with output queuing (FORE/Runner like). In addition the switching nodes can deploy a hardware link-level flow control to prevent data losses due to internal buffer overflow (AT&T, IBM).

We have modelled the behavior of an ATM Segmentation And Reassembly (SAR) interface chip set, available from industry. Specific features of the SAR, such as static and/or dynamic bandwidth allocation and servicing priorities, have been implemented in our model of data sources.

We have simulated two different data flow control strategies, namely “Push” and “Pull”. In the push approach the sources send their data to the allocated processing element as soon as it becomes available. In the pull approach, the processing element requests necessary data from the read-out cards when needed.

The combination of efficient bandwidth allocation with data flow control strategies reduces the contention in the switching network and increases the overall performance of the system.

We have conducted a series of simulations using the specific ATLAS traffic patterns in order to evaluate performance, understand various issues and prove feasibility. To validate the results of the studies, two independent simulation programs have been developed in different object oriented languages (Modsim-II and μC++). Good agreement has been obtained. Results of the performance assessment of the proposed architectures are presented.

In this document we have proposed an integrated architecture for the ATLAS LVL2 / LVL3 selection and data read-out system. It is based on the “Pull” principle and a single network which carries both data and protocol traffics. We have performed simulation studies for the ATLAS calorimeter subsystem to validate the proposed concepts and investigate the feasibility of using ATM as the network technology. A satisfactory system behavior has been observed at the targeted

ATLAS level 1 and level 2 trigger rates. The bandwidth allocation technique, provided by ATM technology, makes an ATM network adequate to handle efficiently our specific types of traffic. We plan to extend our simulation efforts to cover other sub-systems of the ATLAS detector. The demonstrator systems, currently under evaluation, will allow us to refine our models and evaluate performance issues on real hardware.

X. Acknowledgments

The authors wish to thank J. R. Hubbard, J. Bystricky, T. Hansl-Kozanecka, B. Thooris for the physics input; M. Nomachi from KEK for helpful discussions.

XI. 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] CMS collaboration, “The Compact Muon Solenoid Technical Proposal”, **CERN/LHCC/94-39**, December 1994.

[3] M. Costa et al, “NEBULAS - A high performance data-driven event building architecture based on an asynchronous self-routing packet-switching network”, **CERN / LHCC/95-14.**

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

[5] D. Calvet, “Some Ideas for ATLAS Level 2 Trigger Local Processors Allocation”, **RD31 Note 95-10**, August 94.

[6] J. Bystricky et al, “ATLAS Trigger Performance: Electron/Photon Triggers at Level 2”, **ATLAS Internal Note, DAQ-NO-28**, January, 1994.

[7] The International Telecommunications Union (ITU), Geneva, Switzerland; the Telecommunications Standardization Sector (ITU-T); recommendations **I.150, I.211, I.311, I.321, I.327, I.361, I.362, I.363, I.413, I.432, I.610**.

[8] The ATM Forum, c/o Interop Inc., 480 San Antonio Road, Suite 100, Mountain View CA94040-1219.

[9] Transwitch Corp., Shelton, Connecticut, USA, SARA chip set, Technical Manual, version 2.0, Oct. 1992.

[10] Fujitsu Ltd., MB86686A Adaptation Layer Controller (ALC) data sheet, version 2.0, July 1994.

[11] Integrated Device Technology Inc., IDT77201 NICStAR, 155 Mbit/s ATM SAR Controller for PCI-based Networking Applications, December 1994.

[12] Fore Systems Inc., Pittsburgh, the ASX family of ATM switches.

[13] Alcatel Data Networks, Alcatel 1100 HSS, private communication.

[14] V. P. Kumar et al, “PHOENIX: A Building Block for Fault Tolerant Broadband Packet Switches”, IEEE **Global Telecommunications conference**, December, 1991, pp. 228-233.

[15] W. E. Denzel et al, “A Highly Modular Packet Switch for Gb/s Rates”, **XIV International Switching Symposium**, Yokohama, Japan, October 1992, vol. 2, A8.3.

[16] Henrion, M. et al, “Technology, Distributed Control and Performance of a Multipath Self-Routing Switch”, in **Proceedings of the XIV International Switching Symposium**, Yokohama, Japan, October 1992, vol. 2, pp. 2-6.

[17] MODSIM II - The Language for Object-Oriented Programming, CACI Products Company, La Jolla, California, January 1993.

[18] Buhr, P.A. et al., “μC++: Concurrency in the Object-oriented Language C++”, **Software - Practice and Experience**, vol. 22(2) February 1992, pp. 137-172.

[19] C. Paillard, “An STS-OC3 SONET/ STM-1 SDH ATM Physical layer implementation and Application to an ATM Data Generator”, **RD-31 note 95-04** February 1995.

[20] L. Gustafsson et al, “A 155 Mbit/s VME to ATM interface with special features for event building applications based on ATM switching fabrics”, **RD-31 note 94-11,** November 1994.

[21] Hewlett Packard, Broadband Series Test System, 1994.

[22] S/ATM 4615 Adapter, User’s Guide, Interphase corp., June, 1994.

[23] V/ATM 5215 Adapter, User’s Guide, Interphase corp., September, 1994.