[35] Creative Electronic Systems SA, Geneva, RIO 8260 and MIO 8261 RISC I/O processors - user's manual, version 1.1 (March 1993).

[36] Transwitch Corp., Shelton, Connecticut, USA, SARA chipset, Technical Manual, version 2.0, Oct. 1992.

[37] PMC-Sierra Inc., the PMC5345 Saturn user network interface manual (May 1993).

[38] T. Lazraq, Traffic shaping hardware for event building with ATM switching fabrics, RD-31 note 94-09.

[39] K. Agehed et al., Progress report on the design and performance of a VME-ATM module using dual-ported memories, RD-31 note 94-01.

[40] Fujitsu Mikroelektronik GmbH, Germany, the MB86687 adaptation layer controller, the MB86683 network termination controller and the MB86689 address translation controller.

[41] M. Costa, ATM/VME Interface Software Description, RD-31 internal note 94-04

[15] 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.

[16] GlobeView-2000 Broadband System, System description, At&T network system, Red Bank, New Jersey, 07701

[17] I. Richer, B.B. Fuller, An Overview of the MAGIC Project, M93B0000173, The MITRE Corp., Bedford, MA, 1 Dec. 1993.

[18] For example the EXPLOIT ATM testbed, Basel, Switzerland is run under the European Commissions Research and Technology in Advanced Communications Technologies in Europe (RACE) program.

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

[20] General DataCom Inc, Middlebury, Connecticut, the APX-DV2 ATM switch.

[21] Digital Equipment Corp, the Giga-switch with ATM support.

[22] IBM Corp., Nways ATM products.

[23] I. Mandjavidze, “Software Protocols for Event Building Switching Networks”, in Proceedings of the International Data Acquisition Conference, Fermilab, Oct. 1994 (to be published).

[24] M. Letheren et al.,”An Asynchronous Data-Driven Event Building Scheme based on ATM Switching Fabrics”, IEEE Trans. on Nuclear Science, vol. 41, No 1, Feb. 1944. Also available as CERN / ECP 93-14.

[25] I. Mandjavidze, A new traffic shaping scheme: the true barrel shifter, RD-31 internal note 94-03.

[26] T. Lazraq et al., Performance evaluation of an event builder based on an ATM switching fabric with an internal link-level hardware flow control protocol, Proc. of the Open Bus Systems Conference, Munich (Nov 1993), pp. 163-169. Also available as CERN / ECP 93-24.

[27] D. Calvet, “A MODSIM Model of the AT&T Phoenix switching Fabric”, RD-31 Internal Note 94-07, Aug. 1994.

[28] W. Greiman, "A Scalable Fiber Channel Architecture for Event Building", in Proceedings of the International Data Acquisition Conference, Fermilab, Oct. 1994 (to be published).

[29] ATLAS SIMDAQ, A. Bogaerts et al., Modelling of the ATLAS data acquisition and trigger system, ATLAS Internal Note, DAQ-NO-18.

[30] Oechslin, P. et al., ALI: A Versatile Interface Chip for ATM Systems, Proceedings of the IEEE Global Telecommunications Conference ’92, Orlando, 6-9 December 1992, pp 1282-1287.

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

[32] Buhr, P.A. et al., uC++: Concurrency in the Object-oriented Language C++, Software - Practice and Experience, Vol 22(2) (February 1992), pp 137-172.

[33] Hewlett Packard, Broadband Series Test System (1994).

[34] L. Gustafsson et al., “A 155 Mbit/s VME to ATM interface with special features for event building applications based on ATM switching fabrics”. In Proceedings of the International Data Acquisition Conference, Fermilab, Oct. 1994 (to be published).

this intermediate layer would be to segment the event fragments into several AAL5 packets and to recombine them in the destination.

Inevitably software and hardware layers of the protocol are not totally independent of each other; but we are attempting to design the software in a way that identifies general functions, independent of a particular hardware implementation, and isolates code that is dependant on the particular chip sets used for the lower layers of the protocol. We are developing the software [41] in parallel with the hardware of the source/ destination modules and with the assembly of the test system. Regular contacts with RD13 are maintained in order to design a protocol stack that is compliant with their recommendations.

References

[1] J. Christiansen et al., NEBULAS - A high performance data-driven event building architecture based on an asynchronous self-routing packet-switching network, CERN / DRDC 92-14 and CERN / DRDC 92-47.

[2] J. Christiansen et al., RD-31 status report, CERN / DRDC 93-55.

[3] J-Y. Le Boudec, The asynchronous transfer mode: a tutorial, Computer Networks and ISDN Systems 24 (1992) 279-309.

[4] C. Partridge, Gigabit Networking, Addison-Wesley Professional Computing Series, 1994.

[5] A series of recommendations concerning ATM developments has been issued by the ATM Forum, 303 Vintage Park, Foster City, CA 94404 USA.

[6] Relevant recommendations, drawn up by the International Telecommunications Union, Geneva, Switzerland are:

G.707, Synchronous Digital Hierarchy Bit Rates
G.708, Network Node Interface for Synchronous Digital Hierarchy,
G.709, Synchronous Multiplexing Structure.

See also G. Pellegrini and P.H.K. Wery, Synchronous digital hierarchy, Telecommunication Journal Vol 58 - Nov. 1991, pp 815-824.

[7] ANSI T1.105-1991, Digital hierarchy - Optical interface rates and formats specifications (SONET).

[8] I. Mandjavidze, “Modelling and performance evaluation for event builders based on ATM switches”, RD-31 internal note 93-06 (December 1993).

[9] L. Goldberg, Electronic Design, 51-65, April 4, 1994.

[10] “The ATM report guide to ATM network interface cards” in The ATM Report, August 31, 1994, Broadband Publishing Corporation, Rockville, MD, USA.

[11] V.P. Kumar et al., Phoenix: A building block for fault tolerant broadband packet switches, Proceedings of the IEEE Global Telecommunications Conference, December 1991, Phoenix, USA.

[12] Fujitsu Mikroelektronik GmbH, Germany, the MB86680 4-by-4 self-routing switch.

[13] W.E. Denzel, A.P.J. Engbersen, I.Illiadis, “A flexible Shared-Buffer Switch for ATM at Gb/s Rates”, accepted by Computer Networks and ISDN Systems, Dec 1993.

[14] Henrion, M. and Boettle, D., “Alcatel ATM Switch Fabric and its Properties”, in Electrical Communications, Vol. 64, No. 2/3, (Alcatel Paris HQ, 33 rue Emeriau, 75725 Paris Cedex 15, France, 1990) pp.156-165.

This development has proved to be very useful to familiarize ourselves with the implementation of ATM technology and has shown us which are the critical points requiring improvement in order to be able to sustain traffic at the full bandwidth offered by the 155 Mbit/s bit-rate.

ATM Data Generator

Modules to generate ATM traffic for the demonstrator are also under development. In one approach [39] an alternative commercial ATM segmentation and reassembly chip set [40] is used. A second approach provides a simple, cheap and flexible way to "load" many inputs of the switch with any desired traffic pattern by generating sequences of ATM cells from a "library" of pre-defined cells stored in a memory. It will be used to generate traffic through the switch, in connection with full function ATM adapters.

12. Hardware and software implementation of the data acquisition protocol stack

In the data-driven event builder, the sources have the task of sending event fragments to the destination. Each source must collect the data (or poll for their arrival) in a memory and identify the virtual connection (VC) over which the data will be sent to the destination. In the destination, event fragments from different sources have to be linked together to form a built event. Some method must be applied to recognize when all fragments of an event have been received, and missing fragments must be flagged. This global scheme of assembling the fragments constitutes the top layer of the data acquisition (DAQ) protocol stack [41].

The underlying layers of the DAQ protocol stack are dependant on the switching fabric architecture and technology. We consider here only the case of an ATM self-routing packet switching architecture. We have selected the standard *ATM adaptation layer (AAL)* protocol called AAL5 (one of several standardized AAL protocols) to implement the method by which variable length data packets, with a maximum length of 64 kByte, are segmented into (and reassembled from) sequences of fixed-length ATM cells. The next lower level of the protocol stack, namely the *ATM layer*, handles the operation of routing the cells through the switching fabric. The *physical layer* specifies how the cells are to be framed and transported over some physical medium (in our case fibre optic links). For high performance we have chosen to implement the AAL and lower layers using commercially available protocol chip-sets, and to implement only the upper layers in software.

An intermediate software layer is necessary to adapt the requirements of the top DAQ layer described above to the services provided by any AAL hardware that may be selected. The implementation of this layer depends on whether traffic shaping is needed or not. Thus we distinguish two cases:

(1) for **fabrics without link-level hardware flow control**, traffic shaping is required, and every source module must maintain one logical FIFO queue per destination. The segmentation of packets is done by scanning the queues in a round-robin fashion, either extracting one cell at a time (cell based traffic shaping [24]) or extracting cells continuously from the same queue until a timing signal synchronizing all the sources is received (barrel shifter traffic shaping [25]). The segmentation chip that we use provides all the functionality needed to create and manage the queues and to organize the round-robin extraction. The DAQ software, apart from declaring the new fragments to be transferred, only needs to intervene, in the case of the barrel shifter, on receipt of the timing signal in order to move the segmentation process to the next queue.

(2) for **fabrics with link-level hardware flow control**, the switch can be operated without traffic shaping under certain conditions as described above. In that case, the fragments from successive events are simply queued in a single FIFO queue and the role of the DAQ software is to initiate the transmission by specifying to the segmentation hardware the VC identifier to be used and the location in memory of the event fragment. In the case where the flow-control mechanism propagates the congestion "back-pressure" as far back as the sources, the AAL5 layer must react by momentarily stopping the segmentation process.

If the event fragments can be larger than 64 kByte (expected for calibration data), an additional task of

The source modules in the demonstrators can be either commercial interfaces, or the interface developed by RD31 and described below, or they can be simpler “data generators” which are also described below They will, in a first phase, send data pre-loaded in a local memory.

11. ATM Interface and data generator developments

ATM Interface

We are developing a full duplex ATM network interface [34]. Fig 12 shows a block diagram of this

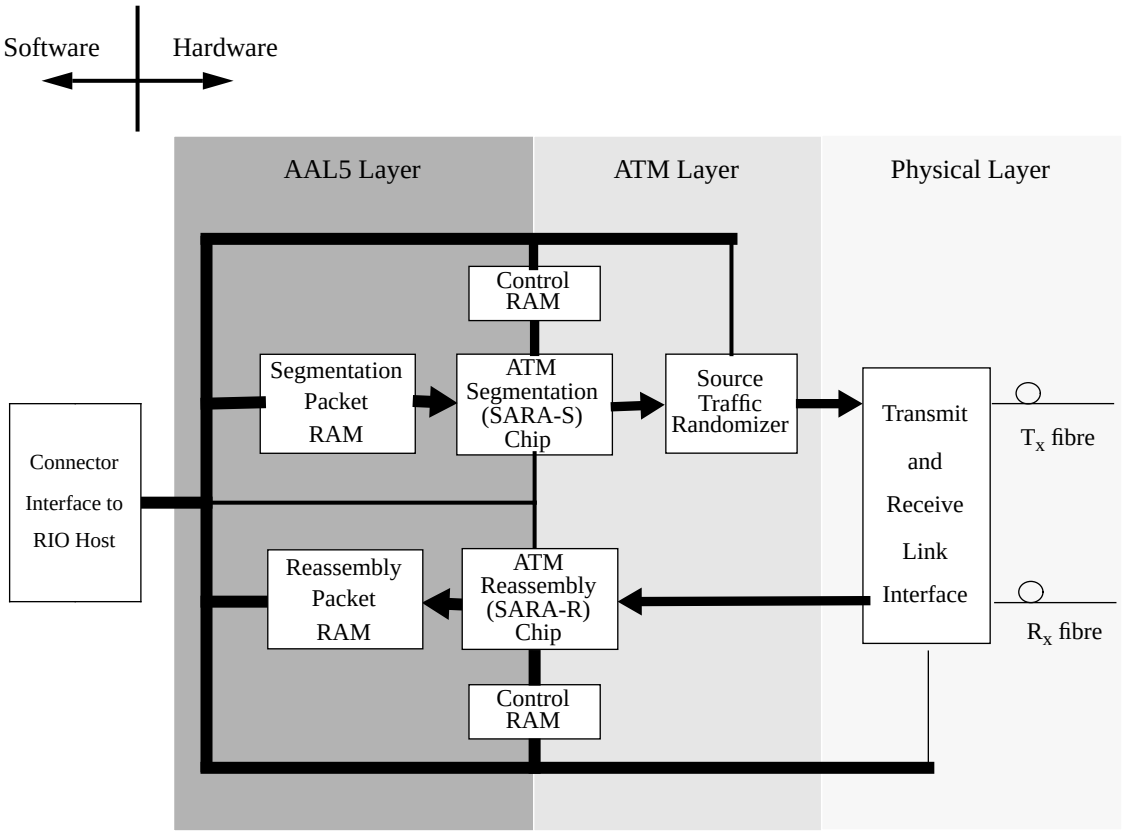


Fig 12: Block diagram of the VME - ATM interface

interface. It is implemented as a daughter board on a commercial VME module [35]. This module includes a processor which will run the software implementing the higher layers of the data acquisition protocol stack. A commercial chip set [36] performs in hardware the segmentation and reassembly of data packets (up to 64 kByte long) into/from ATM cells. Another commercial chip [37] is used for framing the ATM cells for transport over optical fibres using the 155 Mbit/s bit-rate option of the SONET [7] standard. The interface includes special hardware [38] to perform the traffic shaping required for event building over telecommunications switches which do not have internal hardware flow control. We intend to build several of these interfaces and to use them as source and destination modules in the demonstrator system.

The interface has been fully tested with SONET/SDH framing and optical fibre, across the Alcatel switch. It correctly receives at the input port variable length data packets transmitted from the output port. Further optimization of the design is required in order to sustain the full bandwidth offered by the 155 Mbit/s bit-rate of the fibre optic transmission standard.

Part 3 - Interfaces and Event builder demonstrator systems

10. Event builder demonstrators

RD31 is implementing event builder demonstrators using ATM switches. The aims are to validate, to some extent, the results from modelling, to test traffic shaping methods and to develop and test the software high level DAQ protocols. A demonstrator can also be used as a real time, small scale simulator of a Level 2 trigger system part, if real or simulated physics data are stored in the source memories and injected repeatedly in the switch.

Figure 11 shows one of the demonstrator systems currently being assembled. The switching fabric is a prototype 8 x 8 multi-path self-routing architecture [14, 15] provided by Alcatel Bell Telephone.

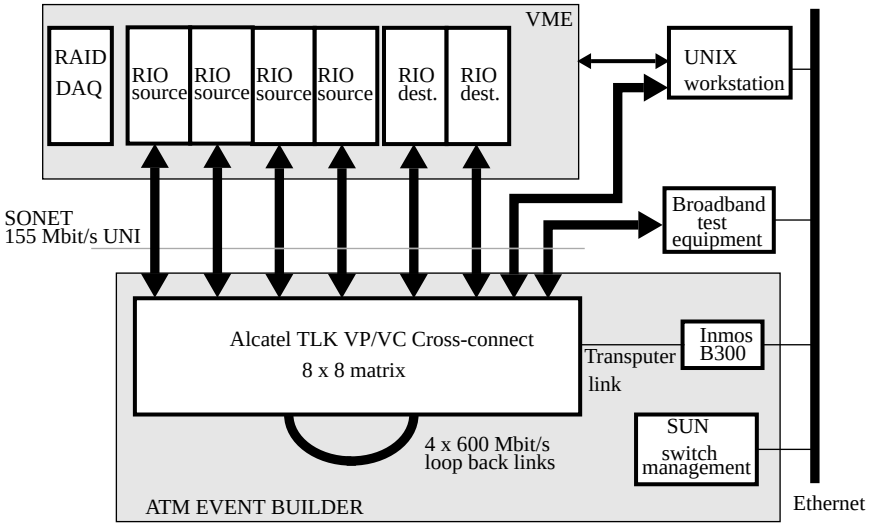


Fig 11: The lay-out of the event builder demonstrator system

The switch has been delivered together with embedded operations and management software (transputer based), and an operator interface which runs on a SUN workstation and communicates with the embedded software via an ethernet to transputer-link bridge. A Hewlett Packard broadband test system [33] is used for ATM protocol validation at the physical and ATM layers and also allows performance measurements and comprehensive error stressing. The functioning of the 8 x 8 Alcatel switch has been successfully validated using the HP test system.

The Alcatel switch supports the 155 Mbit/s SONET User-Network-Interface (UNI) standard, and will be used, together with the ATM source and destination modules described below, to test data acquisition protocols and traffic shaping techniques. Commercial workstations supporting the SONET standard can be incorporated into the demonstrator event builder, and alternative SONET-compliant switching architectures can be evaluated.

A different demonstrator, based on an AT&T Phoenix ATM switch, and using the internal flow control strategy is planned to be set-up at Saclay in 1995. It will allow the verification of some of the behavior patterns observed in the modelling of this type of switch.

The simulation results presented above are preliminary. Based on the developed models and their future versions we are going to perform extensive studies of the relative merits and disadvantages of the "Push" and "Pull" data flow control strategies. Also we would like to investigate a DAQ architecture which is based on a unique switching network for Level 2 and Level 3 traffic. We plan to feed our simulations with more realistic input parameters derived from the physics simulations. The simulation code can be used for similar studies for other detector types, like the Muon detector , the TRT, etc.

extraction (FEX) algorithm duration, though used in the model, has been chosen to be constant. We plan to introduce more realistic FEX algorithm time distribution at a later stage.

The results from the calorimeter local processors have to be combined with the results from the other subdetectors in a global processor. For this we assume that the same network interface and the same fabric can be utilized. Therefore, we use bidirectional ATM connections. We consider that sources are also equipped with bidirectional links, which allows to implement both "Push" and "Pull" architectures.

In our model a generic ATM Multistage Interconnection Network (MIN), based on switching elements of various sizes, provides a data path between the read-out crates and the farms of local processors. For example, 2x2 switching elements can form a MIN with a connection scheme that can be either of Banyan or of Omega type. Contention resolution inside the switching elements can follow either central queueing discipline (like AT&T/Phoenix) or output queueing discipline (like IBM/Prizma). Although each link to the network operates at 622Mbit/s rate, the fabric can provide bandwidth expansion internally in order to reduce contention. External to internal bandwidth matching is done in the network adapters like AT&T/ALI [30]. The relatively simple model of the fabric can be refined in the future as more detailed information will be available from the industry.

Two completely independent simulation codes have been developed in concurrent object oriented languages: Modsim [31] and μC++ [32]. The results derived from the programs have been compared. For this, special runs have been performed with the same configuration and in the same conditions. We have obtained good agreement between the two simulation models.

Figure 10 shows part of the simulation results. The "Crate Occupancy" histogram shows the probability that, for a given event, a part of at least one RoI will fall in a crate. In our model, crates numbered from 0 to 7 and from 24 to 31 perform data read-out from the calorimeter end-caps. The observed lower occupancy for the end-cap crates is due to our assumptions on RoI distribution in η direction. We assume flat distribution of RoI's in the ϕ direction. The average crate occupancy amounts to 27%. For 100 KHz Level 1 trigger rate this means that each crate must be able to collect, format and send Level 2 data at a rate of 27 KHz. Up to four sources may contain data for a given RoI. Therefore one, two or four packets have to be collected in the destination to reconstruct a Region of Interest from the calorimeter. The time distribution necessary for this operation is shown on the "RoI Collection Latency" histogram on figure 10. The average RoI collection latency is ~50μs. The different peaks observed on the time distribution correspond to the different types of RoI's and their distribution among the crates.

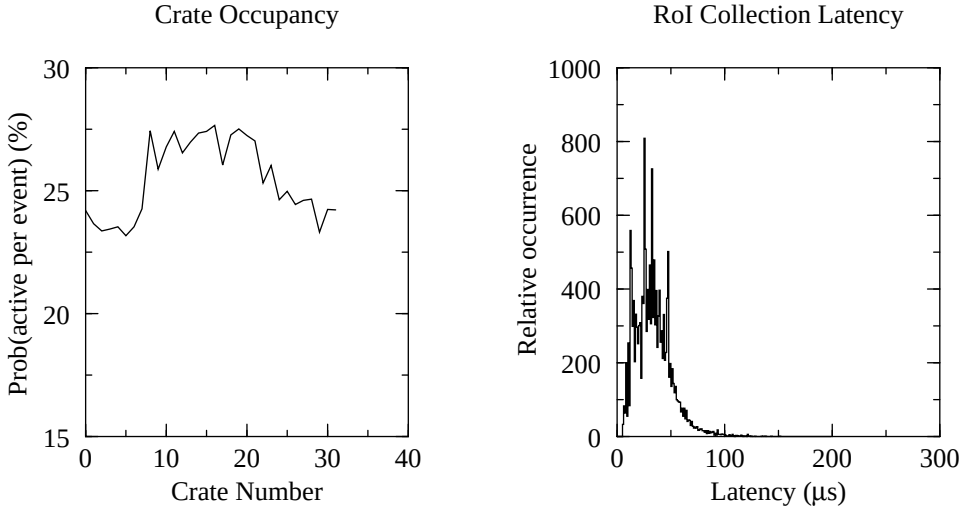


Fig 10: Calorimeter source crate occupancy and RoI data collection latency

Table 3: Average number of ROIs

ROI type	%	ROI μ	ROI e/γ	ROI Jet	Total
jet	3%	1	1	3	5
2 em	12%	2	3	1	6
2 μ	3%	3	2	1	6
missing E_t	3%				

The data presented in the above tables are derived from physics simulation. A simple numerical evaluation, assuming a 100 KHz trigger rate after Level 1, shows that an aggregate bandwidth of 5Gbit/s is required for collecting the data for the processing of the Level 2 selection. Assuming that the event-building traffic for the next selection level will use the same links, a total aggregate data bandwidth of approximately 7Gbit/s is required. This results in a data throughput of 225Mbit/s per crate. Assuming that standard 622Mbit/s links are used to inter-connect the read-out crates with the switching network, the links will be utilized at ~40% (including the overhead due to the ATM protocols), which is a reasonable value. The simulated architecture is shown on figure 9.

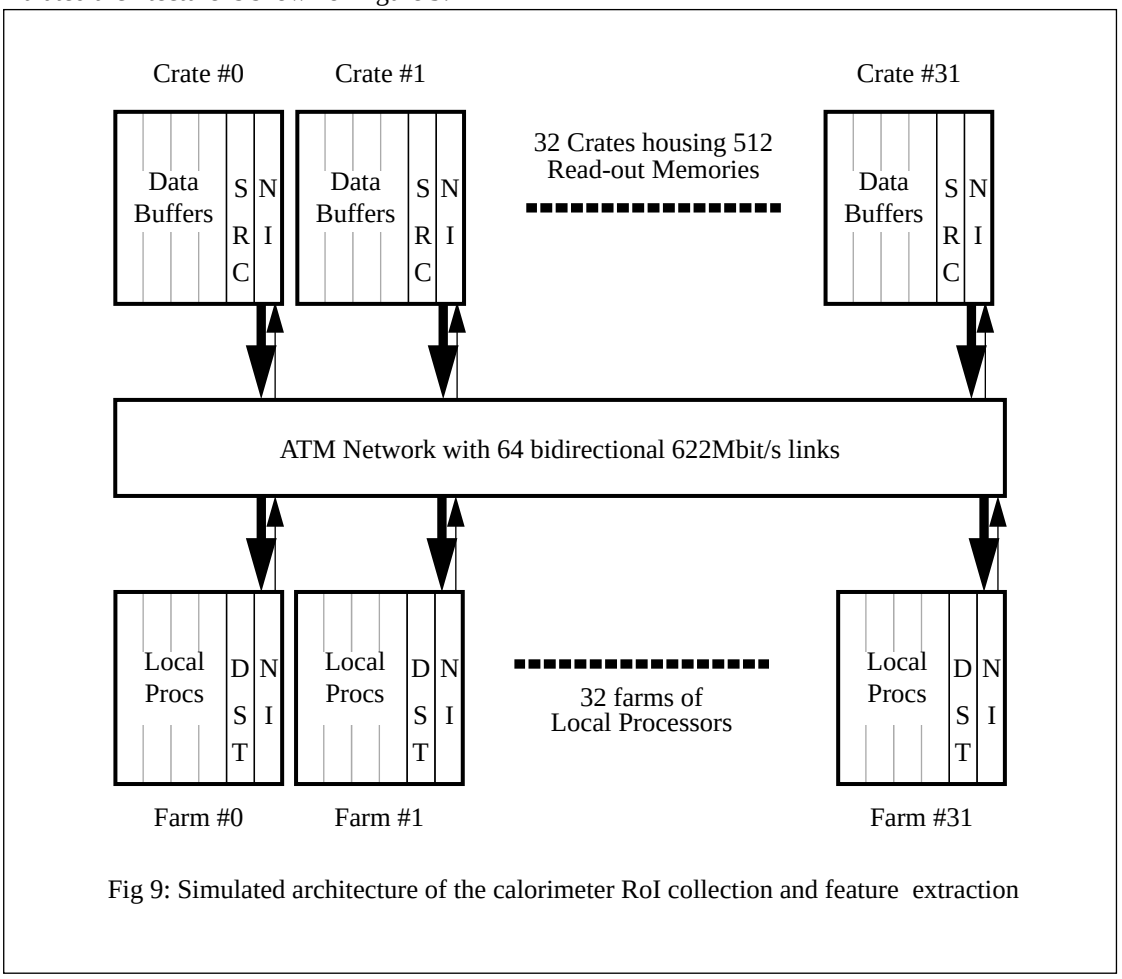


Fig 9: Simulated architecture of the calorimeter RoI collection and feature extraction

The local processors are grouped in farms. At the output of the switch, the Level 2 data are delivered to each farm through a single link. 32 processing farms have been used in our model. In this first modelling study we are just interested in the latency due to the network (RoI collection latency). Therefore the Feature

d) The networks

Data collected from each source data buffer are sent towards the different processing, control or storage elements through one or several data networks. The nature and topological arrangement of these networks depend on different architectural considerations that are under study using modelling.

e) The event selection and collection system

Data selection is subdivided in a series of processing steps: at least 2 before the final transfer to the permanent mass storage media. The synchronization of Trigger/DAQ control messages as well as Accept/ Reject signals including ROI information broadcasting is performed under the control of a trigger supervisor system.

9. Application to the calorimeter L2 sub-system

According to the read-out scheme presented in section 8, for each L1 accepted event data are transmitted from the calorimeter (PS, EM, HAC) front-end boards to the Intelligent Read-out Memories. The read-out of the calorimeter is organized in towers of 0.1 x 0.1 in the η, ϕ space. The detector consists of 64 ϕ by 60 η towers. 512 Intelligent Read-out Memories will be used to store the data during Level 2 and Level 3 decision latencies. The memories will be housed in 32 crates. For example, 16 crates will map the barrel part of the detector, each crate covering 1.4 x 0.8 in the η, ϕ space. One link per crate can be used to transmit the data from the Regions of Interest, required for the Level 2 decision, into the feature extraction (local) processors via a switching network. The types of RoI-s and the corresponding amount of data to be transferred are given in table 2.

Table 2: ROI Size and Data Volume

RoI type	RoI Size	Data Volume Em +PS	Data Volume HAC
μ	0.6x0.6	6 x 6 samples x x (1 layer PS + 2 layers EM) x x 3bytes/sample = 324 bytes	6 x 6 samples x x 3 layers x x 3 bytes/sample = 324 bytes
e/γ	0.3x0.3	3 x 3 samples x x (32 strips PS + 16 in EM1 +8 in EM 2) x x 3 bytes/sample = 1512 bytes	3 x 3 samples x x 3 layers x x 3 bytes/sample = 81 bytes
Jet	0.9x0.9	9 x 9 samples x x (1 layer PS + 2 layers EM) x x 3 bytes/sample = 729 bytes	9 x 9 samples x x 3 layers x x 3 bytes/sample = 729 bytes

The simulated trigger types, their relative occurrence and the average number of RoI-s are given in table 3.

Table 3: Average number of ROIs

ROI type	%	ROI μ	ROI e/γ	ROI Jet	Total
μ	12%	2	1	1	4
em cluster	70%	1	3	1	5

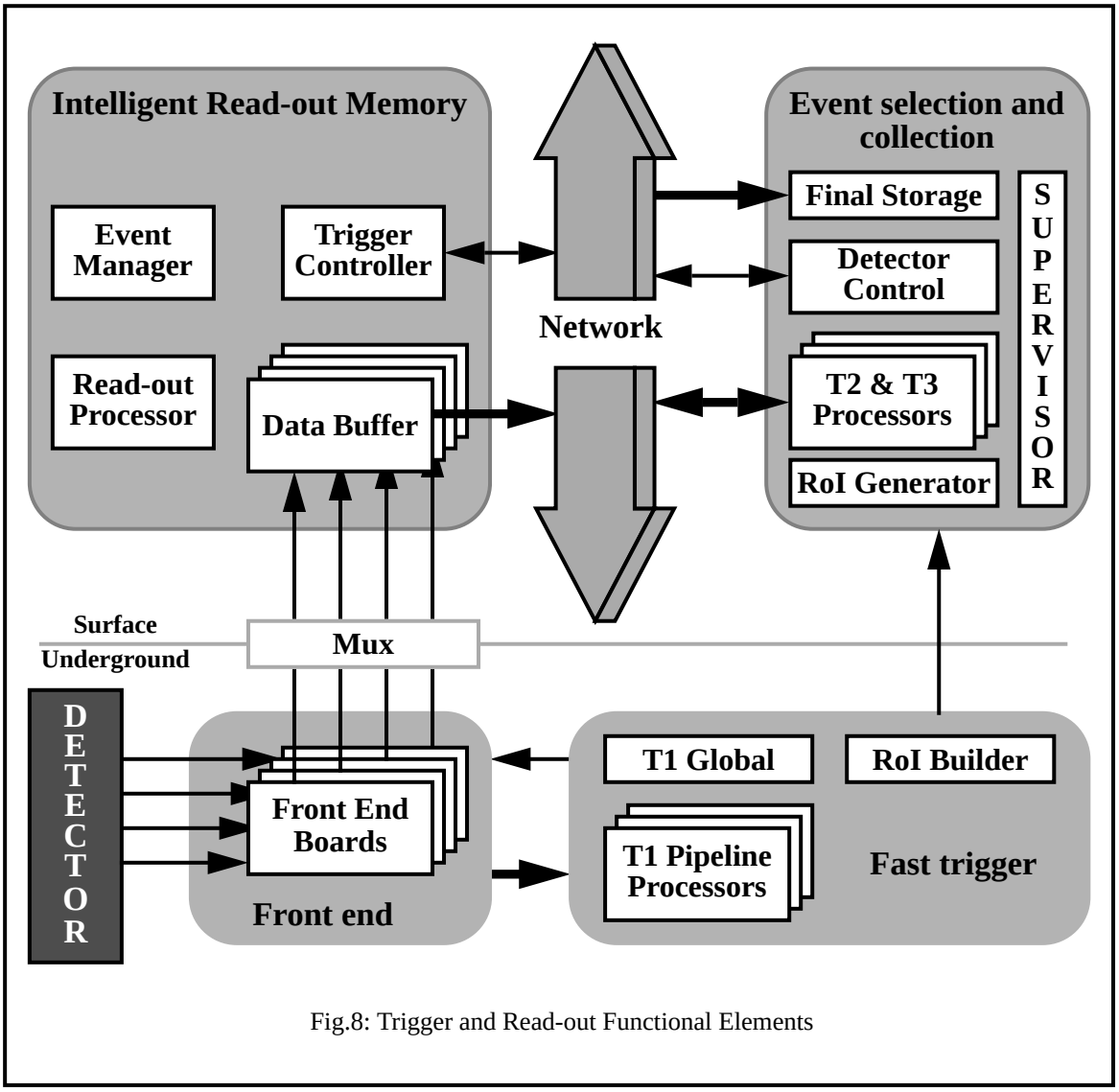


Fig.8: Trigger and Read-out Functional Elements

The final Accept/Reject decision is made by the "Global" T1 processor An associated "Region Of Interest" builder extracts the geographical position of each possible interesting part of the event i.e. muon candidate and large energy cluster above a minimum threshold. Further processing will be carried out only with the data collected from these areas in order to minimize the data transfer bandwidth.

c) The Intelligent data buffer and server

After a Level 1 accept, the digitized raw data are transferred from the front end boards to an external data buffer. Prior to that transfer, an intermediate functional stage is foreseen to multiplex, regroup and order the electronics channels, including fast digital signal processing such as energy and time filters for the calorimeters. The complexity and efficiency of the data extraction for the next trigger levels depends on the topological organization of the read-out. As an example, the Level 1 trigger tower segmentation grid has been chosen as the baseline for the calorimeters and the muon chambers. This device has 2 basic functions: data storage and memory management during the Level 2 latency and may be also during the Level 3 processing time. Additional tasks like data extraction from a Region of Interest and preprocessing, Level 3 data formatting and compression, error detection and recovery necessitate some local fast signal processing.

Table 1: Data Volume Summary

Subsystem	**Channel count**	**Event Size** *KBytes*	**FE Bandwidth** *Gbit/s*	**L2 Bandwidth** *Gbit/s*	**L3 Bandwidth** *Gbit/s*
Pixel	137e+06	55	44	0.18	0.44
SI SIT, GaAs, MSGC	5.16e+06	265	212	0.48	2.12
TRT	421120	337	674	13.16	2.70
PS em preshower	28672	9	69	0.24	0.07
Calorimeter EM, HAC, IFC	197504	147	474	5.60	1.17
Muon Trigger RPC, TGC	890 000	55	44	0.13	0.44
Muon Chamber	360 000	180	144	0.41	1.44
Trigger	4000	32	26	0.04	0.26
Total	-	1 079	1686	20.23	8.63

8. Read-out block diagram

Figure 8 describes the present organization of the read-out. It is tentatively divided into a series of uniform functional blocks that every detector subsystem should implement. Each block has one or more dedicated functions with well defined input and output interfaces. This modular concept allows, in principle, to build and test each block independently with the best technique available at the construction time. For such a large detector, the main technical constraint is not the processing power itself which can be carried out by very efficient hardwired or programmable processors, but rather the topological organization and extraction of signals and data at each step. The efficiency, complexity and cost of the read-out system depend strongly on such an organization.

One can identify 5 groups of elements:

a) The Front End electronics

This first part is located inside or near the detector itself. It contains the digitizer chain including the transducer, the preamplifier, the analog or digital 2 microseconds pipeline and a local de-randomizer RAM. This is also the place where specific data are extracted for the detectors participating in the L1 trigger

b) The fast trigger (Level 1)

Dedicated sets of macro granular information are extracted from some detector front end electronics through point to point links into “local trigger” pipeline processors. Trigger components that must operate at the full 40 MHz beam-crossing rate require fast electronics. The constraint of minimal latency, to match the size of the front end buffers, dictates that the algorithms used at 40MHz speeds must be essentially hardwired and of limited sophistication.

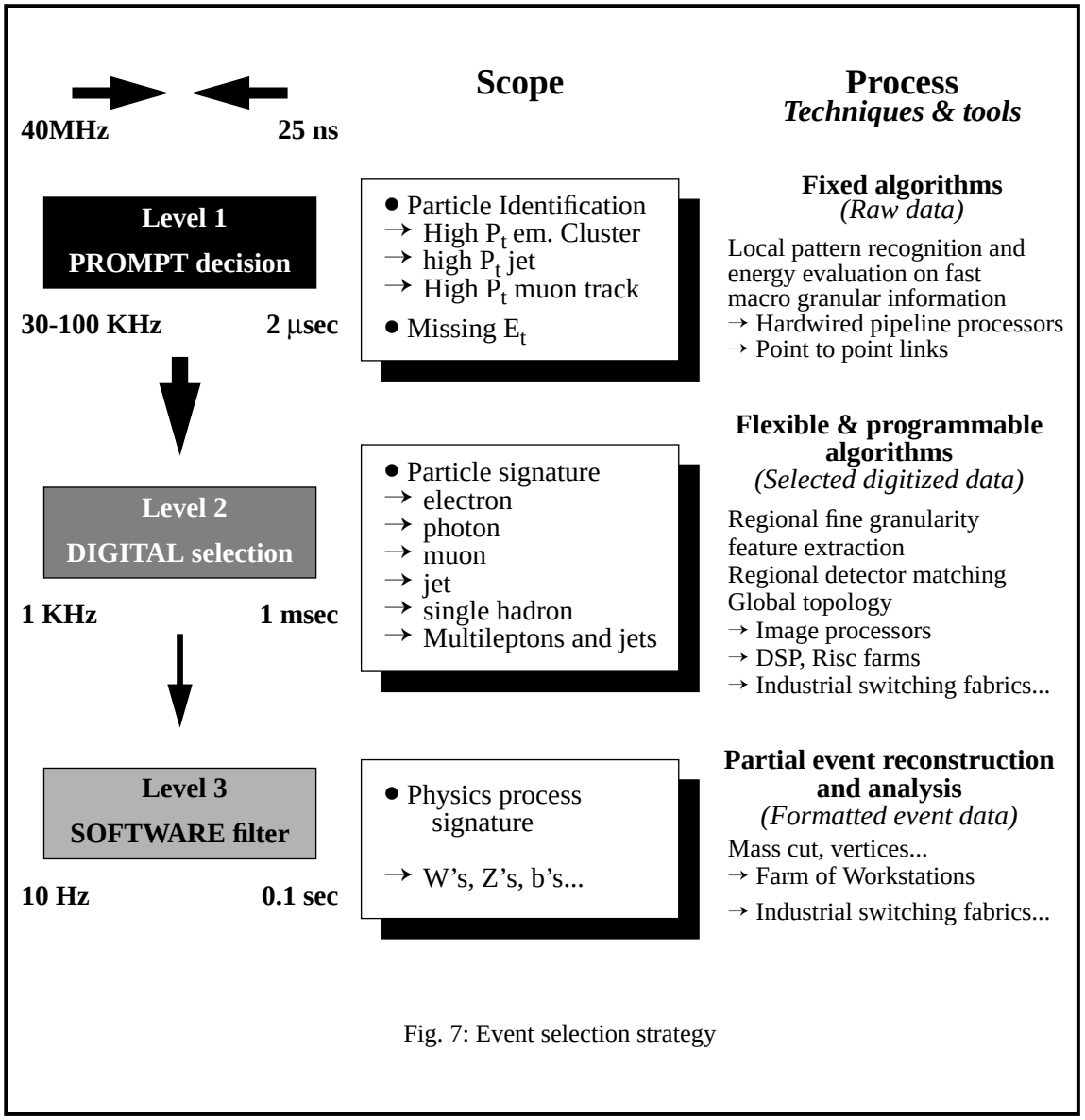


Fig. 7: Event selection strategy

- Matching inner and muon trackers.
- Matching trackers and energy cluster detectors.
- Forming triggers involving pairs of particles and other combination of physics signatures (e.g. Missing Et+ Electron, Electron + Muon....etc.).

The goal of Level 3 is to select and classify the physics processes using the most precise data sample available. Level 3 uses high purity data. On-line corrections and calibrations may be used at that level. For example, the following filter algorithms acting on final partial or full event data might be implemented:

- Precise mass reconstruction (W's, Z's..) and cut.
- Search for secondary vertices and B identification.

A summary of the amount of data produced by the different subsystems is presented below in table 1, assuming Level 2 and Level 3 input rates of respectively 100 KHz and 1 KHz.

Part 2 - ATM Modelling for ATLAS

6. A Model of an ATM switch for the ATLAS DAQ modelling project.

A model of an ATM switch, based on the PHOENIX AT&T [11] switching element has been developed [27], using the MODSIM language [31]. It can be plugged into the ATLAS DAQ model [29]. This work had also as a goal to compare the performances of switches with and without internal flow control. It has shown that rather high loads (75%) can be applied to the Phoenix based switching fabric up to a size of 512 X 512, with good characteristics of latency and buffering space. Event building latencies are significantly lower than in the case of a switch with traffic shaping (fig. 6).

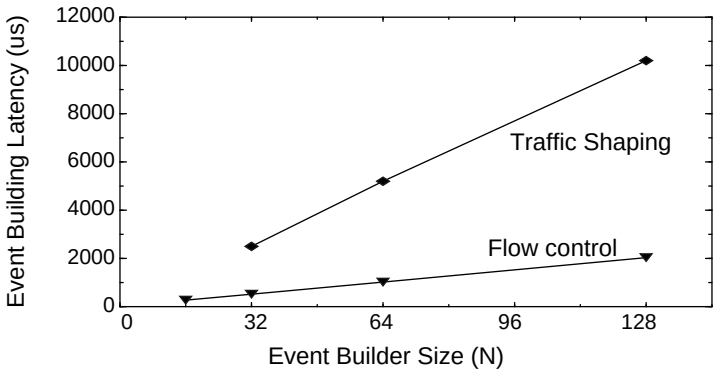


Fig 6: Event-building latency vs. size for different types of switches (75% load)

7. Input parameters, rates and data volumes

The Atlas trigger consists of three logical levels, shown schematically in figure 7. Beam crossing interactions occur at a rate of 40 MHz. At the nominal luminosity of 10^{34}, the input rate coming from the level 1 threshold cuts is estimated to be approximately 30-40 KHz. A value of 100 KHz has been adopted as a standard conservative value. Level 1 is deadtimeless, with all the data pipelined during the 2 microseconds allowed to decide whether to accept or reject the event candidates. The physical rate to be eventually recorded on tape is estimated to be in the range 10-100 Hz. Thus a further reduction of the order of 10^4 is necessary.

Qualitatively this important rejection factor can be achieved by using 2 more logical sequential steps. Level 2 differs from Level 1 in that it does not have to finish before the data comes out of the pipeline. Level 2 decision consequently has a flexibility that Level 1 does not have as it can take longer than a millisecond. Level 2 can implement trigger algorithms based on information that could not be available in time for Level 1. Two areas, for example, that may require Level 2 to be handled are the inner tracker system and the muon precision chambers. Moreover, Level 2 can have access to the full granularity of the calorimeter information as well as to new information from other subsystems that cannot be available at Level 1. In order to reduce the input bandwidth, 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 (see table 1). A trigger reduction of a factor of 100 is expected by executing the following tasks:

- Adding tracking information.
- Refining muon P_t cuts.

(Fig 5 b)). The limiting load is a decreasing function of the switch size, in contrast to switches where traffic

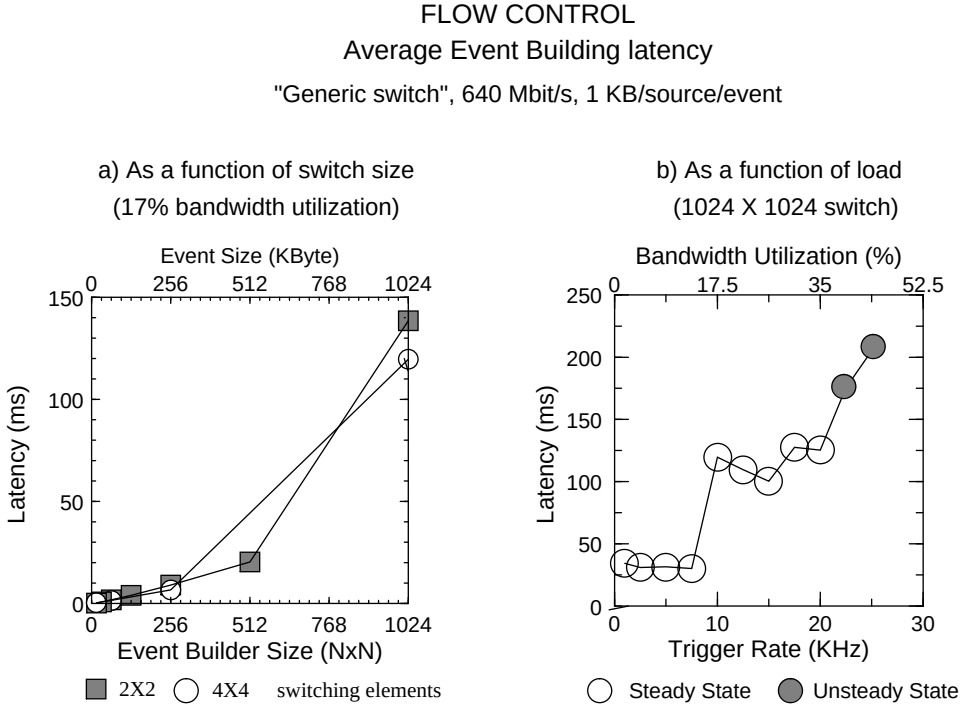


Fig 5: Latency in a switch with internal flow control

shaping is applied and which can be operated up to 80%, independently of the size. The highest achievable load depends also on the internal architecture of the switch, e.g. on the value of the internal bandwidth. Reference [27] shows that high values of load (75%) can be sustained, for switches up to 512 x 512, thanks to the 400 Mb/s bandwidth of the internal paths of the Phoenix switch.

Our simulations also suggest that a switch with internal flow control, used for a particular event building statistical traffic with sufficiently stable characteristics, after a phase of adaptation, tends to organize itself internally in such a way as to make the best use of its resources. The introduction of a small contamination by traffic with widely different characteristics destroys this self organization and results in a significant reduction of the maximum usable load of a switch. The irregular shape of the latency values in Fig 5 b) can be interpreted in terms of different organization patterns (and resulting latencies) that establish themselves with changes in the traffic characteristics that occur when the number of concurrently assembled events in a destination jumps from 1 to 2 to 3, etc.

It should be noted that other packet-switched, virtually non-blocking, switching technologies that implement flow control in order to guarantee data delivery (e.g. Fiber Channel, class 2) would be expected to show similar behaviours of limited usable load. Even using the circuit-switched class 1 mode of Fibre Channel results in limited usable load due to contention for connection establishment [28].

5.3 Combination of both techniques.

It may be interesting to combine both techniques: traffic shaping allows to reach higher loads on the switch while the flow control guarantees a zero cell-loss probability inside the fabric. The propagation of the flow control to the UNI's can be avoided by the traffic shaping and with a careful dimensioning of the buffers in the destinations.

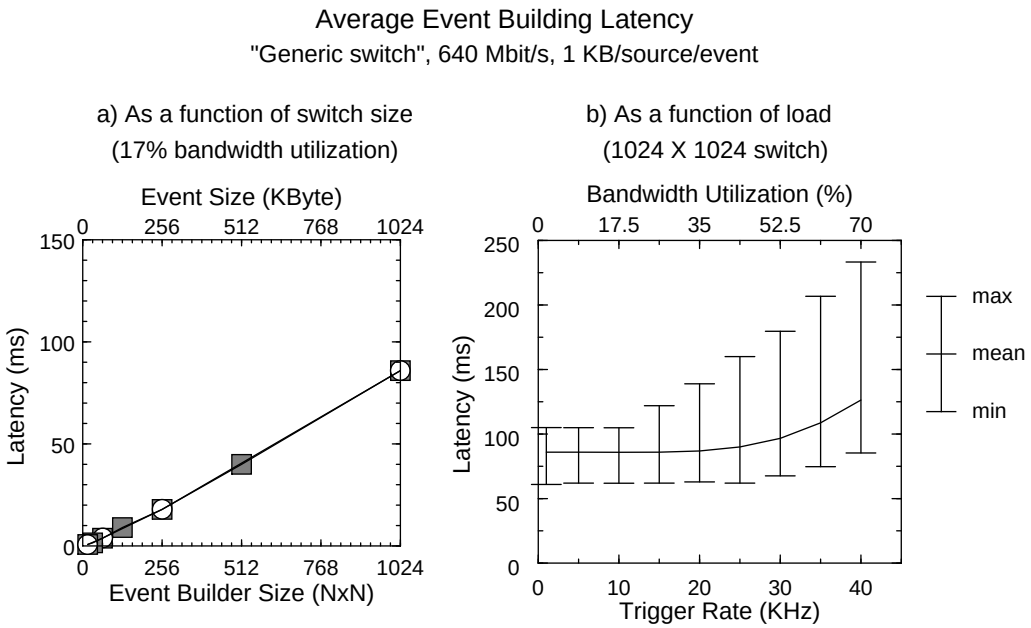


Fig 3: Latency diagrams for a switch with traffic shaping

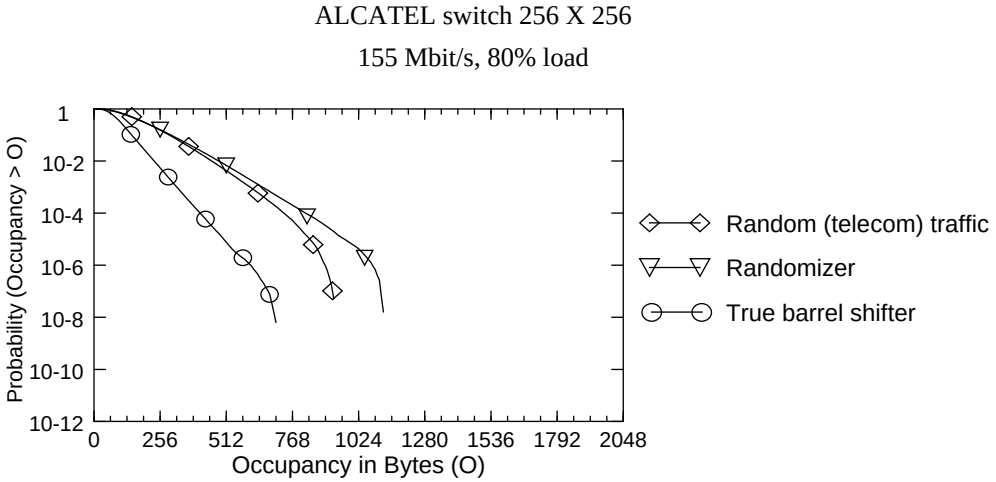


Fig 4: Probability of buffer occupancy.
(Cell loss probability can be deduced by extrapolation)

a switching network. This higher level signalling technique is used in telecommunications applications to modulate traffic so as to reduce any detected congestion areas within the switching network, but it may not be suited for event building applications if the feed-back is too slow.

Our simulations [26, 27] show that an event builder can be implemented on a switch with internal flow control without traffic shaping. Fig 5 a) shows that in this case the latency does not scale linearly at constant load, with the size of the switch. The load on the switch cannot exceed a maximum value beyond which the event building latency grows very rapidly and, consequently also the buffer space required in the sources

sum of the individual cell latencies through the switch as it depends on the size of the queues in the sources and on the bandwidth reduction applied if traffic shaping is used (see below).

Several techniques are envisaged to determine when the operation of assembling an event is completed. The simplest one requires that a source sends at least a control packet if it has no physics data to transmit. A review of different possible techniques has been given in [23].

The event building traffic is characterized by a bursty concentration pattern: for each event, all sources will normally try to send data at the same time towards one destination. Switches are not designed for this type of traffic which would require, even momentarily, a bandwidth at the destination which is N times the nominal bandwidth at the sources. There are two ways to cope with this problem: either to modify the natural traffic pattern of the application before submitting it to the switching network, a technique called *traffic shaping*, or to rely on internal flow control. A combination of both is also possible. We now describe those options in more detail.

5.1 Traffic shaping.

The general principle of traffic shaping is to:

- Limit the average bandwidth on every virtual connection in such a way that the aggregate bandwidth at the destinations does not exceed the nominal bandwidth of the output port. This means assigning approximately 1/N of the nominal bandwidth to each connection.
- Eliminate the burstiness, due to the trigger firing all sources simultaneously by skewing in time the emission of cells from the different sources towards the same destination.

Our simulation of event builders has shown that traffic shaping is necessary with telecom type switches (because they do not implement internal flow control). As this technique is very efficient to organize the traffic in the switch, it allows the use of the available bandwidth in an efficient way, up to 80%, while keeping the probability of data loss at a very low level. In addition, traffic shaping allows a linear growth of latency and required source and destination buffer size as a function of the size of the switch (or size of the experiment). We have investigated several methods of traffic shaping and shown that it is possible to avoid congestion in this way and to produce traffic patterns that are as well or even better adapted to a switch than the so-called random telecom traffic patterns. References [24, 25] give details on the techniques and their results.

Fig 3 (a) shows that latency (and consequently buffer occupancies) scales linearly with the switch size at constant load. Fig 3 b) shows that a switch with traffic shaping can be used at high loads, in fact up to 80% according to simulation results. Fig 4 shows a "tail distribution" of the occupancy of an internal buffer of the Alcatel telecom switch [14, 15]. This tail distribution gives the probability that an internal switching element's buffer occupancy exceeds the value in the abscissa. An extrapolation of the curves to the available buffer size gives the probability of cell loss. The figure shows curves for two traffic shaping methods (the "randomizer" and the "true barrel shifter" [25]) and for the random telecom traffic pattern assumption. The available buffer space being 2048 bytes for each switching element, it can be seen that with proper traffic shaping, values of cell loss probabilities well below 10^{-12} can be achieved.

Traffic shaping can of course also be applied to switches with internal flow control.

5.2 Use of switches with internal flow control.

As already mentioned, switches with internal flow control can prevent overflow of their internal buffers. One flow-control method is the *back pressure* technique, where the transmission of a cell over an internal connection has to be acknowledged by the receiving node. One example, that we have modelled in detail, is the Phoenix switch [11]. In general, switches for LAN applications implement some sort of flow control. We distinguish *internal* flow control used at the link level from flow control protocols that are implemented between subscribers at the level of the ATM layer and make use of ATM cells to signal congested regions in

5. ATM based event builders.

A generic ATM based event builder is shown in Fig 2. It consists of N sources that are connected to the Front-End data acquisition part, M destinations which will receive assembled events and a switch to interconnect the sources to the destinations. Virtual connections are opened permanently, at system start-up, between all N sources and all M destinations. When a trigger occurs, a destination is assigned by a supervisor and its identifier is broadcast to all the sources. Every source sends, through the switch, the part of the event data that it has received from the front-end (an *event fragment*), towards the designated destination using the appropriate virtual connection. This is repeated for each trigger, while the destination is changed. The supervisor determines the destination by some assignment scheme that aims at optimizing the use of the switch.

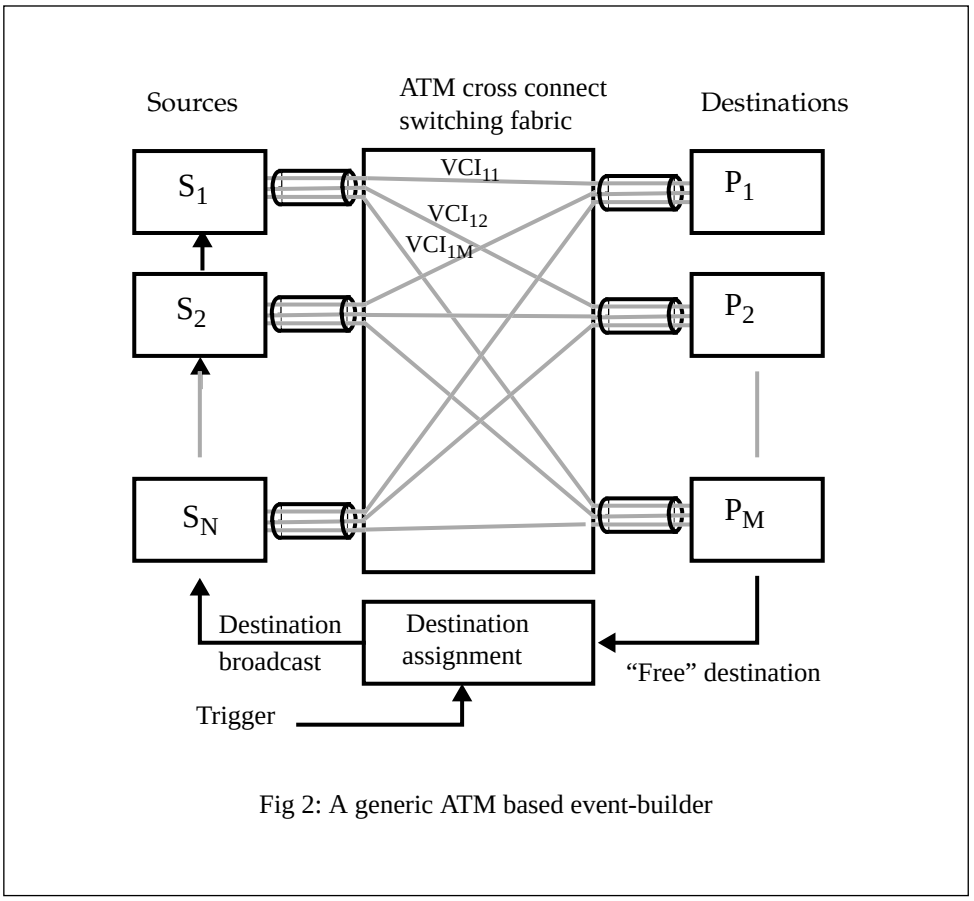


Fig 2: A generic ATM based event-builder

At the destination, all the event fragments are re-assembled into an event. The order in which the cells of an AAL5 packet arrive is the same in which they were transmitted, but the order in which completion of the reassembly of event fragments from different sources occurs is not predictable and will depend on the traffic in the event builder and on the queuing in the sources. However, each event fragment is reassembled in a dedicated reassembly buffer associated with the virtual connection used to transmit it. The ATM label of the incoming cells is used by the reassembly hardware to route the cell to the appropriate reassembly buffer. When the event fragments are reassembled they must be moved from the interface's reassembly buffer into user memory space and linked together to form the complete event record. The event building software must maintain a data structure that links the event fragments into an event. Depending on the trigger frequency and the statistical distribution of event fragment sizes there can be 0, 1 or more events being assembled concurrently in a destination.

The *event building latency* (or "latency" in short) is the time elapsed between the event trigger time and the arrival time at the destination of the last event fragment of that event. It may be much longer than the

A number of ATM interfaces for various bus types (EISA, S-bus, VME, PCI) are, or will soon be, available on the market at 155 Mb/s. Higher bit-rates (622 Mb/s) are not yet common, although indications are that they are under development in the labs. A review of ATM adapters is given in [10].

Integrated switching elements which can be used to build switching fabrics are available from several sources; e.g. AT&T Microelectronic's 2 x 2 Phoenix switch [11], a 4 x 4 switch from Fujitsu [12], IBM's 16 x 16 Prizma switching chip [13], etc.

Telecom switches are available from several sources, e.g. the A1000 multi-path self-routing switch from Alcatel [14,15], switches from Ericsson, or the GlobeView-2000 switch from AT&T [16]. Telecom switches are now under test in various B-ISDN pilot projects in the national telecom companies. Experimental ATM networks already exist in the US [17] and in Europe [18]. Many ATM switches for high-speed LANs are available on the market, including those from Fore Systems [19] , General DataCom [20], DEC [21], IBM [22], AT&T etc.

The main classes of ATM switches.

The switch fabric is built of a number of switching elements interconnected in a topology that can provide at least one path from every input to every output**.** Each switching element contains a local buffer for temporarily storing cells.

There are 3 main classes of ATM switches:

- those designed for the telecommunications industry, where expandability to large dimensions, low-latency and non-blocking characteristics are important. Delivery of data by the switch fabric is not guaranteed, but under the "random" traffic pattern resulting from the aggregation of the traffic of a large number of independent subscribers, the probability of data loss is acceptably small (of the same order as the loss probability in a long distance link),
- switches which implement an internal flow control in order to guarantee transfers with no loss of data. These are more likely to be used for LAN applications.
- switches based on a shared medium (e.g. bus)

Cell loss occurs in a switching fabric when an internal buffer overflows, as a consequence of traffic congestion. Even if the network control system ensures that the connection characteristics do not exceed, on average, the resources of the switch, traffic burstiness and particular traffic patterns (e.g. concentration of traffic) can produce overflow in some of the internal buffers. Usually the telecom switches implement some mechanisms to detect congestion and slow down traffic at input. However the reaction time of these controls is slow. For random traffic, the switch's internal buffers are dimensioned such as to give a very low probability of loss, typically of the order of 10^{-10} or lower at 80% load on the switch (see for example [14]).

When there is any correlation between the traffic flowing on different virtual connections, the traffic patterns are no longer random and the probability of internal buffer overflow increases. In this case it is the task of the user interface to regulate the traffic in such a way as to avoid congestion. This technique is called *traffic shaping* and can be used for event building over a telecom switch.

The technique of *internal flow control* can be used to prevent buffer overflows in the switch by holding up the traffic flowing towards a nearly full buffer until sufficient buffer space becomes available. In this way, no cells are lost in the switch. However one must consider the case where the buffers of the 1st stage overflow and the case where the destination user buffers overflow. The ATM standard does not specify an action in those cases, except for a higher-level flow-control protocol, which might not react fast enough to avoid data loss.

The shared media switches are mostly used in the LAN applications and they are not expandable.

In all cases there are techniques to limit the data losses to acceptable values. A careful evaluation of the switching network by means of simulation is necessary to properly dimension the network and the interface buffers.

transmission over optical fibers. SONET defines transmission bit-rates which are multiples of a base value: 51.84 Mb/s, (for instance, SONET STS-OC3 at 3 times the base frequency, namely 155.52 Mb/s). The hierarchy of SDH bit-rates forms a geometric progression obtained by multiplying the SDH base bit-rate of 155.52 Mb/s (STM-1) by factors of four. The sequence of bit-rates (155.52 Mb/s, 622.08 Mb/s, 2.488 Gb/s) allows a simple 4:1 multiplexing of streams into the higher bandwidth links used for cost-efficient long-haul trunk lines. In what follows, we shall abbreviate the exact bit-rates to 155 Mb/s, 622 Mb/s etc.

The SDH and SONET standards define a data "frame", the payload of which can carry several ATM cells. The frame also contains control and status information to support link-level error detection, operation and management functions. The overhead, due to the SDH frame control and status information reduces the effective payload throughput to 26/27, thus at 155.52 Mb/s the effective bit-rate for ATM cells is 149.76 Mb/s. ATM cells themselves have a 10% overhead in the form of the 5-byte cell header.

SDH also specifies standards for the physical medium sub-layer. We distinguish three, namely one for long-haul optical links using lasers over monomode fibers, one for cheaper, short-haul optical links based on LEDs (light emitting diodes) over multimode fibers, and for short-haul electrical connections, the Unshielded Twisted Pairs (UTP) at 155 Mb/s.

The ***Adaptation layer***, defines how to adapt the ATM layer to the requirements of specific services. Several adaptation layer standards exist for different applications. For data transmission the so-called AAL5 protocol is used; it specifies that data can be transferred in variable-length blocks of up to 64 kBytes. There is no header, but it uses a trailer of 8 (or more) bytes terminated by a CRC. The trailer may include padding bytes to make the total length a multiple of the ATM cell payload length (48 bytes). The task of the adaptation layer is, in the transmission direction, to build the complete AAL5 structure and to segment it into ATM cells which are passed to the ATM layer. On the receiving side, the AAL5 packet is reassembled from the incoming ATM cell stream to deliver the user data to the higher levels of the application protocol. Obviously, the protocol overhead in software would limit performance, and for that reason commercial chipsets have been developed to perform the functions of the adaptation layer by hardware at full speed.

There is no concept of "guaranteed delivery" in the ATM standard. Guaranteed cell-delivery has been deliberately sacrificed in order to allow the design of very large, low-latency, non-blocking switches. Cell loss detection operates by using the AAL5 CRC and "length" fields. If needed, cell-loss recovery has to be implemented in higher levels of the protocol stack. However, we have shown in [8] that one can set up an event builder in which the probability of cell loss can be kept very low so that retransmission of data is not needed (see also section 5.1, Traffic shaping).

4. Status of industry deployment of ATM

In telecommunication applications, bit rates of 155 Mb/s, 622 Mb/s and above are used, or will be used in the future. For computer networks, although many standards have been implemented for LAN applications, the market now seems to favor the SDH/SONET standard at 155 Mb/s. A growing choice of products are becoming available for this standard and commercial chipset are available which implement the various protocol layer functions:

- AAL5 segmentation and reassembly,
- the various physical layer standards,
- the combination of both in a single chip.

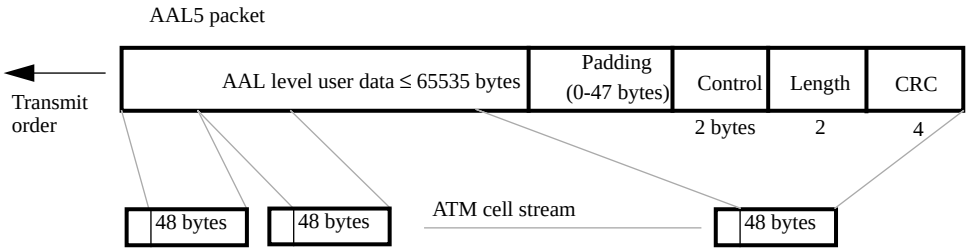
A review of currently available ATM chipsets is given in [9]. A practical difficulty encountered when assembling systems with chipsets from different vendors is the logical and physical compatibility between the products for the AAL5/ATM layer and for the physical layer. Interworking between these layers should be simplified by the recent introduction of an ATM Forum standard (known as "UTOPIA") which defines the interface between the two layers (signals and their timing, board and connector geometry, position and pin-assignments, etc.).

- The *Physical Medium Sublayer* which defines the physical support for bit transfer (e.g. the optical fiber specifications) and the timings.

Numerous standards have been defined at this level, for different applications, by the ITU on one hand and by the ATM Forum on the other hand, with the aim of allowing for early, cost-effective deployment of the technology for computer networking using existing components.

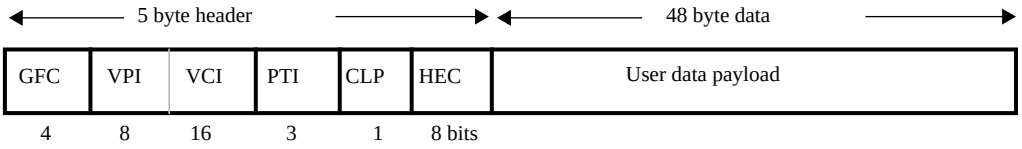
The telecommunication standard for the physical layer is the SDH (Synchronous Digital Hierarchy) [6], a norm based on SONET (Synchronous Optical Network) [7], which was a pre-existing US standard for serial

ATM Adaptation Layer



AAL5 for data communication service
Variable packet size: up to 64 KByte
Segmentation, Re-assembly

ATM Layer



Fixed size cells (48 bytes + 5 bytes header)
Virtual connections (can be kept open --> no overhead)
Self routing of cells (VCI + VPI id in header)

Physical Layer

SONET frame generation:

- 155 Mbit/s
- 622 Mbit/s
- 2440 Mbit/s

(non SONET implementations for LAN's)

Physical media:

- optical fibers (mono/multi mode)
- UTP, STP (155 Mbit/s)

Fig 1: The main features of the 3-layer ATM Standard

3. The ATM standard.

The ATM standard is described in numerous places. Reference [3] is an early and concise description of ATM, while reference [4] provides a thorough and recent view of ATM in the general context of high bandwidth networking. We give here a summary of the main aspects that are relevant to event building applications.

ATM is a standard that has been developed for telecommunications applications by the ITU (International Union of Telecommunications, formerly known as the CCITT) to form the basis for the future B-ISDN (Broadband Integrated Services Digital Network). It is also actively promoted within the ATM Forum [5] by the computer industry as a future standard technology for very high bandwidth LANs (Local Area Networks) and WANs (Wide Area Networks), which will support real-time multi-media and distributed computing applications (c.f. the future Information Superhighway). It is expected to have a long life cycle because of its origin and expected large scale deployment in the telecommunication field.

The ATM technology has been designed to support a massive, low latency, non-blocking switching capacity that is suitable for carrying, on a common infrastructure, the traffic generated by a wide range of services, each with its own specific performance requirements. In order to achieve this flexibility it is based on the principle of packet-switching, in which information is carried between "subscribers" in packets (called "cells" in ATM jargon), each of which carries routing information in its header. The network is composed of one or more switching fabrics, each of which uses this routing information to forward the cells through the network towards their destination.

The ATM standard specifies the connections between an end-station and the network (UNI = User Network Interface) and between sub-networks (NNI = Network to Network Interface). To simplify, the standard is sub-divided into 3 layers (Fig 1):

- The Physical Layer
- The ATM Layer
- The ATM Adaptation Layer (AALn)

We start by describing the core of the standard, the ***ATM Layer,*** which is common to all services. It defines that the information (voice, image or data) is to be transported by means of small, fixed length, cells containing 53 bytes (48 bytes of "pay-load" and 5 bytes of "header"). The cell header includes 3 bytes that carry a label identifying a connection between a particular source and a particular destination. This label is used by the switching fabric hardware to route the cell to its destination ("self-routing"). Connections can either be set-up permanently at hardware initialization, or can be established and broken dynamically by a "signalling protocol" (a procedure which is analogous to establishing a telephone connection by dialing the subscriber code).

Connections need not be assigned with a constant available bandwidth, but instead average and peak characteristics of the connection's traffic can be declared and bandwidth can be used on demand. Several connections can be mixed on the same physical path, provided that their aggregate characteristics do not exceed the physical characteristics of the support. In the network, cells belonging to various connections can be interleaved asynchronously. The connections are said to be *virtual* because the cell multiplexing and switching technique obviates the need to reserve dedicated hardware paths between subscribers (the so-called circuit switched technique). Establishing a virtual connection consists in loading, at the various nodes along the route between subscribers, the characteristics of each connection so that the cells can be routed according to their labels. It is possible to have several concurrent virtual connections between a source and a destination as well as from or towards a given station. It is important to note that the ATM layer guarantees that, on any given virtual connection, cells will be received in the same order as they are transmitted.

The ***Physical Layer*** specifies how ATM cells are physically transmitted over a link. It consists of 2 sub-layers:

- The *Transmission Convergence Sublayer* which defines the bit rates and the framing patterns (grouping of cells into larger packets)

Part 1 - Generalities

1. Goals and layout of this document

This document supplements the ATLAS Technical Proposal. It presents in more detail the concept of using commercial switching fabrics developed for the Asynchronous Transfer Mode (ATM) to implement parallel event builders for the ATLAS data acquisition (DAQ) system. In order to ease the introduction of the concept to the reader, it has been written as a self-contained document that can be understood without reference to many other papers. However, it also includes an extensive bibliography to assist more in-depth investigations.

In the first part, the ATM standard is summarized, as well as the current level of support by industry. A presentation of on-going generic R&D on the applicability of ATM to event building is given, with an introduction to the most important issues.

The second part describes a specific R&D programme targeted at the application of ATM event builders in the ATLAS experiment, including some simulation studies for the application of ATM to the ATLAS Level 2 system.

The third part describes the status and future plans for implementation of event builder demonstrators. This part describes the development of ATM interfaces, data generators and event builder software protocols undertaken by the RD31 project [1, 2].

2. Organization of ATM research for event building.

ATM [3] is one of the possible standardized and commercially available technologies that are candidates to implement the switching networks required for data collection for the L2 and/or L3 ATLAS trigger sub-systems. The RD31 project is carrying out generic research on the applicability of ATM to high rate and/or high bandwidth data acquisition systems. The RD31 project is also studying the application of the technology to specific experiments, in particular for ATLAS. The main research directions in RD31 are:

- Detailed study of the behavior of commercial switches and the development of simulation models for some of them.
- Development of simulation models of event builders for Level-2 and Level-3 triggers and study of the performance of various architectures for the expected trigger rates and data distributions.
- Development of ATM interfaces, data generators and their use to implement small demonstrator event builders, where DAQ protocols can be evaluated and the results scaled to predict the performance of a full scale (ATLAS) system.

Modelling activities are essential as they provide the main method for evaluating the performance of event builder architectures. An accurate model of existing industrial switches is necessary to prove the feasibility, but generic models are also useful to understand the general behavior of the switch-based event builder and the trade-offs to be made in selecting various switch architectures.

The development of demonstrators and interfaces is a complementary activity which, to some extent, allows the validation of the results obtained from simulation. This hands-on experience with ATM technology leads to the identification and understanding of issues that are critical factors in judging the pros and cons of using various commercial products in the event building application.

ATLAS Internal Note
DAQ-NO-024
1st December 1994

ATM-based Event Building

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

D. Calvet, K. Djidi, P. Ledu, I. Mandjavidze.
CEA DSM/DAPNIA, Saclay

L. Gustafsson
Institute of Radiation Sciences, University of Uppsala, Uppsala

T. Lazrak, Th. Lindblad, H. Tenhunen
The Royal Institute of Technology, Stockholm

(RD31 Collaboration)

A. Manabe, M. Nomachi.
National Laboratory for High Energy Physics (KEK), Japan