

# EVENT BUILDING IN FUTURE DAQ ARCHITECTURES USING ATM NETWORKS

DAVID C. DOUGHTY JR., DAVID GAME, STEPHANIE HOLT, LISA MITCHELL

*Christopher Newport University, Newport News, VA, USA*

PAUL BANTA, GRAHAM HEYES, THEODORE PUTNAM, W. A. WATSON III

*Continuous Electron Beam Accelerator Facility, Newport News, VA, USA*

ATM switches and links have been investigated for use as the event building fabric for future data acquisition architectures and will be used in CEBAF's CLAS detector. To avoid contention problems and cell-loss, a linked dual token passing algorithm has been devised, with two different types of tokens being passed through the switch. This algorithm leads to a 'barrel shifter' type of parallel data transfer. We describe the hardware architecture and the dual token algorithm, and present simulation and test results.

## 1 INTRODUCTION

Data acquisition architectures in the next millennium will face the challenge of moving hundreds to thousands of megabytes of data each second from front end electronics to on-line processor farms. Event building in such environments requires an approach that is both scaleable and parallel, which minimizes contention for links and processor resources. Switched architectures provide many advantages in event building, most notably the potential for parallel data movement into processor farms.

Asynchronous Transfer Mode (ATM) is one of the newest switching architectures and is rapidly gaining wide acceptance. In ATM all data is transferred in 53 byte cells, comprised of a 5 byte header and 48 bytes of data. The header controls the routing of the cell through network switches. This small and fixed cell size is designed to provide the low latency switching required by integrated video and data networks. ATM does not provide guaranteed delivery of cells, and ATM switches typically have limited buffering, so contention for output bandwidth may cause cell loss. This is not a serious problem in video applications, but for event builder applications this type of random data loss is unacceptable, so a way to eliminate it must be found in order to use ATM.

We have investigated ATM switches and links for use as the event building fabric for future DAQ architectures, and plan to employ them in the CLAS detector at CEBAF. In this paper we first detail our method for using ATM as an event building fabric, then discuss a MODSIM based simulation of an event builder for the CLAS detector. Actual tests of an ATM event builder are discussed next. Finally, the extension of this architecture to large arrays of readout controllers and farm processors is discussed.

## 2 EVENT BUILDING USING ATM

### 2.1 Hardware Architecture

Figure 1 shows a block diagram of one of the event building architectures which has been studied for use in the CLAS detector at CEBAF. The ATM switch has 16 fiber optic OC-

3 (155.52 Mbit/s) ports, and serves as the vehicle to route both data and control information between on-line farm processors (OLFPS), readout controllers (ROCs), and the tape processor. The ROCs collect data from the electronics crates while the OLFPS are used for triggering, reformatting, partial analysis, and monitoring. They also send data to be written to tape through the switch to the tape processor.

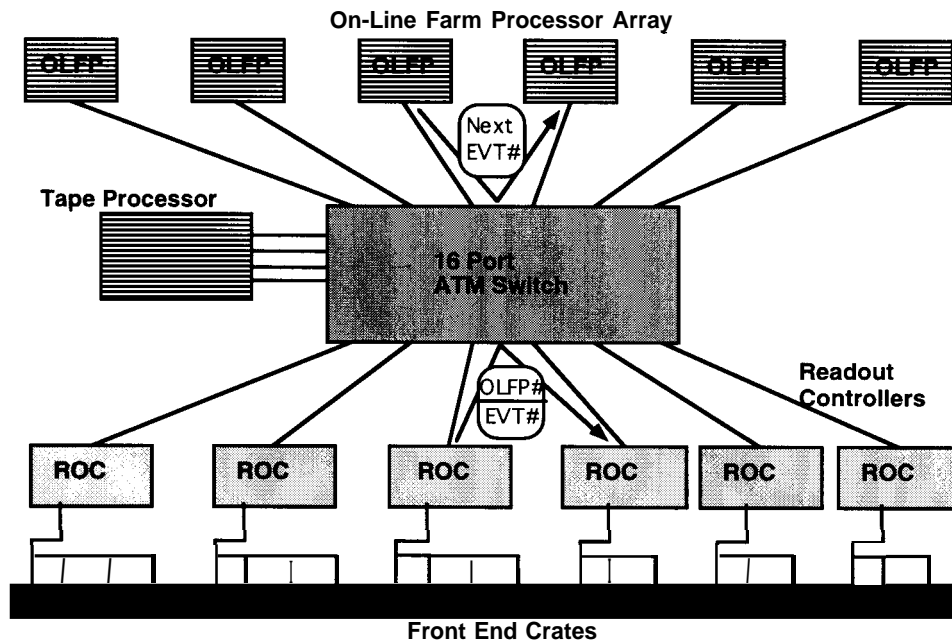


Figure 1. An ATM based event builder. The ATM switch serves all communication needs, including token passing for control as well as data movement. The farm processor token is circulating among the OLFPS, while an event request token is being passed from one ROC to the next.

## 2.2 Tokens and Transfer Control

As previously mentioned, one of the major problems in using an ATM switching network for event building is contention on the outbound links to the OLFPS, causing cell loss. This is handled in two ways. First, all communication is done using the TCP/IP sockets protocol, which guarantees reliability without the need to rewrite applications to use a native ATM application programming interface (API). If cells are lost the TCP/IP layer will retransmit them. If retransmission occurs frequently, due to contention and accompanying cell loss, it may cause a drastic reduction in throughput. We greatly reduce the contention by shaping the traffic using a linked dual-token passing algorithm.

The dual-token algorithm not only minimizes the output link contention problem, but also handles the arbitration among OLFPS for events and communicates this information to the ROCs. **To reduce the amount of control information passed, all arbitration and data transfer is done using blocks of events. A 'farm processor token' circulates among the OLFPS and controls which of them will receive a specific block of events. As each OLFPS takes an event block it sends a second type of token, an 'event request token' down to the first ROC, identifying itself as the recipient of data from that event block.**

The first ROC enqueues these tokens in event order; when it has collected its fragments of the events specified by the token on top of the queue, it sends that event fragment block to the requesting OLFP. It then sends the token through the switch to the next ROC in the chain, which does the same. This continues until the last ROC has sent its fragments to the OLFP, after which it sends the token back. Additional details of the token passing algorithm are given in reference [1].

Since multiple event request tokens may be active simultaneously, with each ROC sending its data to a different OLFP, a 'barrel shifter' type of parallel data transfer occurs. This architecture is also scaleable since larger numbers of ROCs or OLFPs can be accommodated either by using larger switches, or by cascading smaller ones.

### 2.3 Queue Size in the ROCs

Because of the delay in passing the event request token down the ROCs, a backlog of events builds up, causing larger queues of unsent events in those farther down the token passing chain. This queue size may be modeled in the simple case where each ROC has the same amount of data. If we let R be the event rate, N the number of events per event block, B the number of bytes per event in each ROC, T the actual data transmission rate from the ROCs through the switch to the OLFPs, and A the time to pass the token between two ROCs, then the queue size in bytes at the Ith ROC is given by :

$$\text{Queue}[I] = NB*(1 + I * B * R/T + I * A * R / N) \quad (1)$$

This equation indicates that to minimize the queue size the number of events per block should be small. However, to minimize the amount of arbitration by the OLFPs the number should be as large as possible. The correct choice of N in any application will be a tradeoff of the cost of memories vs. the cost of processor performance.

The delta term in this equation depends on the latency of the switch and the interrupt service time of the ROCs, but in most applications may be neglected. For example, with B=1Kbytes, T=4 Mbytes/s, N=64 events, and  $\Delta=1$ ms the delta term is **16 times smaller than the preceding one**. Table 1 shows the queue length at the last ROC for several different experiment sizes, neglecting the contribution of the delta term.

Table 1. Queue size at the last ROC for various experiment sizes.

# of ROCs	N evts/blk	R evts/s	B bytes/evt/ROC	T bytes/s	Queue size at last ROC
6	64	2 K	2 K	4 M	896 KB
20	64	2 K	2 K	4 M	2.63 MB
1 K	64	1 K	1 K	8 M	16.1 MB

## 3 SIMULATION AND HARDWARE TESTS

### 3.1 Simulation

We have developed a simulation of this architecture using the MODSIM2 language. The simulation uses tables to set the parameters of the model, which include the event rate, event size in each ROC, number of ROCs and OLFPs, the MIP ratings of the ROCs and OLFPs, the computational power required by the link protocol, and the amount of analysis required of the ROCs and OLFPs. More details of the model are given in Ref [2].

Initial simulations focused on the 6 x 6 event builder shown in figure 1. The first three ROCs had 2.5 Kbytes of compressed data per event while the last three had 1

Kbyte. With ROCs of 125 MIPS and OLFPs of 250 MIPS the simulation indicated that up to 2000 events/s could be handled before the ROCs were unable to keep up with both the processing and data movement demands.

The asymptotic performance was tested using ROCs and OLFPs of infinite MIPS, to gauge the maximum possible performance of the architecture. Under these conditions the architecture supported event rates up to 8 kHz, but was unable to keep up at 10 kHz.

These simulations indicate that ATM is viable for use in the CLAS detector with the expected event rates of less than 2 kHz.

### 3.2 Hardware Tests

A small switch and several ATM adapters were acquired which allowed a small test system to be set up. An HP 735 and 755 were used as the OLFPs with two HP 743 VME boards and two HP 715s used as the ROCs. All machines were running HP-UX, using supplied or developed ATM drivers. Initial tests used 48 Kbyte packets passed from one machine to another as fast as possible, directly and through the switch. Then a 1 x 1 (one OLFP and one ROC) version of the dual token event builder was tested. The results, shown in Table 2, indicate that the switch has no noticeable effect on the transfer rate while the event building algorithm has an overhead of about 10% on the EISA bus computers. The low utilization and large decrease in transfer rate of the VME processor when running the dual token algorithm is puzzling, but may be due to problems in the ATM driver or VME hardware.

Table 2. Data transfer rates in Mbit/s for various ATM communication links and protocols. Processor utilization was measured while running the dual-token algorithm. The destination processor is an HP-755 in all cases.

Source Processor	Direct	Through Switch	Dual-Token Algorithm	Processor Utilization
715/75	53	53	49	43%
715/50	43	43	38	64%
743166	49	49	32	15%

The rates for the 1 x 1 event builder can be used to predict the performance of a 1 x 4 event builder. The predicted value of 37 Mbit/s is the same as the actual measured value of 37 Mbit/s. If another OLFP is added to make a 2 x 4 event builder, the data rate on both outbound links is 37 Mbit/s, indicating that this architecture scales linearly as additional OLFPs are added.

## 4 LARGER SYSTEMS

This architecture can be extended to very large systems. As an example we will consider a system composed of 1 K ROCs and 1 K OLFPs with an event rate of 1 kHz and an event size (after zero suppression and compression by the ROCs) of 1 Mbyte evenly distributed among the ROCs.

To connect the ROCs to the OLFPs we use sixty-four 64-port ATM switches arranged in a two stage Clos network as shown in figure 2. This connection between the ROC layer switches and the OLFP layer switches minimizes contention and allows fully parallel data transfer to occur if the path of the farm processor token is modified. Instead of passing to the adjacent OLFP the token must be passed to an OLFP in the next switch as shown in the OLFP sequence numbers on the right. This event builder has an

aggregate data rate of 1 GB/s off the detector and into the processor farm. The queue size at the last ROC (shown in Table 1) is a reasonable 16 M bytes.

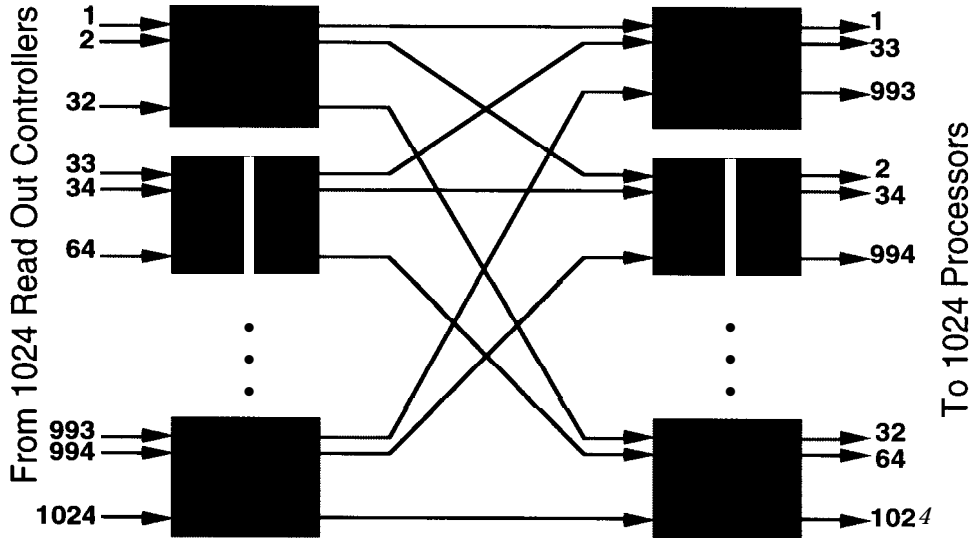


Figure 2. The event building architecture for a 1K x 1K event builder. Sixty-four 64-port ATM switches are arranged in a two stage Clos network. The farm processor token skips processors following the sequence numbers on the right.

## 5 CONCLUSIONS

The event building architecture described here will be used for CEBAF'S CLAS detector. A 32-48 port ATM switch will be used to build an 8-16 x 20 event builder which will be in use by the fall of 1996. The ATM links are fast enough to handle the data rates expected, and the TCP protocol handling leaves enough processing power so the OLFPs can perform some analysis. The use of the 'dual token barrel shifting' algorithm almost completely eliminates contention and cell loss in the ATM environment. Using the switch to pass the tokens simplifies the design by eliminating a separate control network. The maximum queue size in the ROCs is reasonable.

This design is also very scaleable. The use of larger switches arranged in a two stage Clos network allows the data from large arrays of ROCs to be moved to large arrays of OLFPs with minimal contention at extremely high bandwidths.

## 6 ACKNOWLEDGMENTS

This work was supported in part by Department of Energy contract DE-AC05-84ER40150 and by National Science Foundation grant PHY-9512705.

## REFERENCES

- 1 Lisa Mitchell, "ATM in Event Building Architectures", MS Thesis, CNU, 1995.
- 2 D. Doughty et al "Event Building Using an ATM Switching Network in the CLAS Detector at CEBAF," Proceedings of the International Data Acquisition Conference on Event Building, Fermilab, 1992.