Performance Measurements of Mixed Data Acquisition and LAN Traffic on a Credit-based Flow-controlled ATM Network

M. Nomachi, Y. Sugaya, H. Togawa, K. Yasuda and I. Mandjavidze[2]

Research Center for Nuclear Physics, Osaka University,
Mihogaoka 10-1, Ibaraki, OSAKA 567, Japan
[2] CERN, CH - 1211 Geneva 23, Switzerland

Abstract

The high speed network is a key component in networked data acquisition systems. An ATM switch is a candidate for the network system in DAQ (data acquisition system). We have studied the DAQ performance of the ATM network at RCNP (Research Center for Nuclear Physics) Osaka University. Data traffic on DAQ system has a very much different traffic pattern from the other network traffic. It may slow down the network performance. We have studied the network performance on several traffic patterns.

I. Introduction

It has been considered to use a general purpose network for the data collection path in large scale DAQ. However, it is not obvious that we can use them. Most general purpose network systems are designed to handle random traffic. In contrast, the data transfer in DAQ system is a continuous flow that may cause heavy congestion in the network. There have been simulation studies to investigate such effects [1,2]. It is simulated and analyzed that coherent burst of data traffic may cause severe problem on the system [2,3]. Possible ways to avoid such congestion have been studied by the RD31 project [4,5]. However, so far there were few measurements made on real systems.

RCNP, Osaka University introduced a new computer and network system for the analysis of experimental data and for theoretical calculations. The system is also designed so that real-time data from experimental apparatus can be handled. Several performance measurements have been made to find optimal set-up parameters for mixed traffic resulting from office LAN traffic and DAQ traffic.

The results of performance tests made at RCNP are relevant for networked DAQ R&D for large scale experiments such as ATLAS [6]. The performance tests of switch type event builders reported here are carried out in a collaboration with Saclay and CERN. Some results are presented using demonstrator event building software developed within the ATLAS collaboration.

II. RCNP Computer and Network System

The new central computer system at RCNP comprises 6 SMP (Symmetric Multi Processor) servers, 20 workstations and 5 tape library systems with the total capacity of 12.9TB. Two of 6 SMP servers are DEC Alpha server 8400 5/440. Each of them has 12 Alpha CPUs and a 155 Mbps ATM interface. They are used for general data processing and computing. The other four smaller servers are DEC Alpha server 4100 5/400. Each of them has 4 Alpha CPUs and a 155 Mbps ATM interface. There are 20 workstations, which are DEC Alpha station 500/333. Each of them has 155 Mbps ATM interface.

The server computers in the central computer system and workstations are linked with an ATM network. Four GIGAswitch/ATM switches are located at the computer room, the cyclotron laboratory, the experiment area and the main office building. These switches are connected with 155 Mbps optical fiber links. Those links were replaced by 622 Mbps optical fiber links in the later measurements in this document. The GIGAswitch/ATM system provides a 10.4-Gb/s aggregate bandwidth in a 13 x 13 non-blocking crossbar switch. It integrates many computer systems as a single powerful computing facility.

One of the server machines is used for real time data acquisition. Data from single board computers in the experimental area is transferred to the server. Reflective memory modules with dedicated optical data link are temporarily used. It might be replaced by an ATM network, whenever an ATM interface becomes available on the single board computers. Data transfer from the real-time server to the data storage server and computing server already uses the general purpose ATM network.

III. Networked DAQ R&D

There are many items we have to study to develop networked data acquisition system. In RCNP, we have studied ATM data transfer over DEC's GIGAswitch ATM. The performance of a demonstrator's DAQ system[7], which is developed at Saclay for ATLAS DAQ, was measured.

A. *ATM performance test*

Point to point communication between two workstations has been tested. A client node and a server node are connected via 155 Mbps links to GIGAswitch ATM. The client requests a certain amount of data to the server. The server node responds with the requested amount of data. The requests are pipelined. The maximum frequency of the sequence was measured for several layers of protocol. Figure 1 shows the results. TCP and UDP protocols give about 10 kHz maximum frequency for short response lengths. The data rate reaches its maximum value if the response length is bigger than 8 k bytes. Direct access of AAL5 protocol gives about 16

0018-9499/98$10.00 © 1998 IEEE

kHz maximum frequency. It is a simpler protocol. It reaches maximum speed if the response length is larger than 2 k byte.

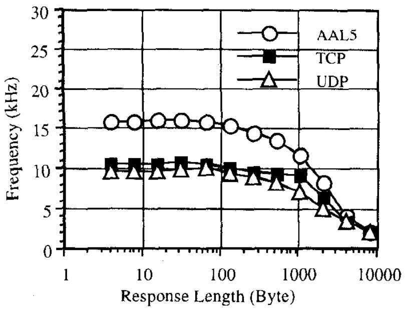

Figure 1: Benchmark results for Point-to Point communication.

The data is taken on DigitalUNIX which is not a real-time operating system. The frequency of more than 10 kHz is fast enough for data acquisition system to use as *request* link. One can achieve maximum data rate at a few k bytes. It is enough efficient to use as *data* link in most of DAQ applications. One can conclude that networked DAQ systems can be developed on the GIGAswitch ATM with DigitalUNIX.

Note that for all tests reported in this paper we use credit-based data flow control provided by DEC[8,9]. Use of flow control avoids possible data loss.

B. *Event builder performance*

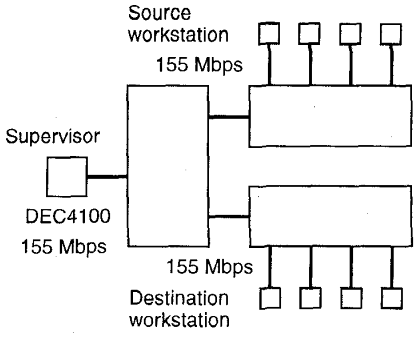

Figure 2: Configuration of an event builder emulator.

It is very much interesting to emulate an event builder on the ATM switch cascade at RCNP. The cascade connection is necessary for a large scale switch. It is also interesting to measure the performance of an event builder on a credit-based flow control system. An ATLAS event builder demonstrator program [7] based on ATM network with AAL5 protocol has been ported for the emulation. It is developed at Saclay. Three GigaSwitches, which are shown as large boxes in figure 2, connected with 155 Mbps link in cascade are used to emulate the event builder. Four workstations emulating sources are connected at one GigaSwitch ATM. Four workstations emulating destinations are connected at another GigaSwitch ATM. A DEC 4100 server emulating a supervisor is connected at the last GigaSwitch ATM, which sits in between.

The supervisor assigns a destination for each event. The destination sequentially requests the event data from the sources. The maximum aggregates event rate that the total system can handle is measured. The results are shown in Figure 3. When the destination can accept only one event at a time, the event builder emulator operates from about 3 kHz to a little bit less than 4 kHz rate, depending on the event size. For small size events, the maximum event rate can be increased significantly by assigning more than one event to a destination. One can achieve 11-12 kHz event rate. There was no data loss even at the maximum event rate, due to the DEC proprietary Flow control system. [8,9]

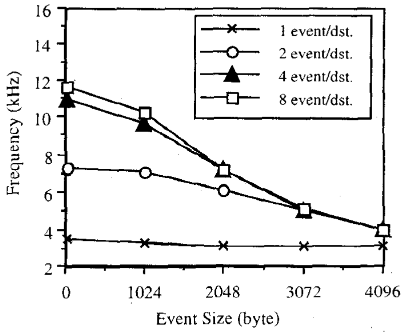

Figure 3: Maximum event rate of the event builder emulator.

We will extend the emulation to a larger system by adding more machines to the network at RCNP. We may be able to study external congestion avoidance mechanisms by switching off the DEC proprietary Flow control system. A study of congestion in large systems could be made. Simulation studies indicate that the effect of congestion becomes more important in larger switches [3].

IV. DATA FLOW ANALYSIS

The data sources of the DAQ system send data continuously, which is very much different from the traffic pattern on an office LAN that is used for communication with for example X-protocol. The congestion generated by random traffic appears statistically and will disappear. On the other hand, the congestion generated by continuous traffic stays at the same place and causes a hot spot in the switch. The problems generated by this hot spot may extend to the other parts of the switch and may cause a performance reduction.

We have studied these effects by using traffic patterns to generate hot spots and then measuring the performance of the switch. In stead of using the ATM Forum's standard rate-based flow control mechanism the GigaSwitch ATM uses the DEC proprietary credit-based flow control scheme, which transmits cells over a virtual connection only as long as there is buffer space available to receive them, thereby avoiding cell loss in a congested network. The credit-based flow control scheme regulated the rate of data transmission to use the available bandwidth. When the data flows of different virtual connections compete for the resources, the flow control scheme should ensure the fair sharing of bandwidth. The implementation of GigaSwitch ATM flow control mechanism was naturally designed to handle typical LAN traffic patterns. The purpose of our measurements was to investigate how the credit-based flow control mechanism performs for the continuous data flows that are typical of event building and generate permanent, localized hot spots.

We measured the performance of data transfer on AAL5 protocol of 8 K byte frame size. All sources sent data as much as possible till the Credit-based Flow-control system slows down their data transfer. Measured data rates were generally a fraction of the maximum possible data rate. They are shown in the figures.

A. *Resource sharing*

A four-to-four data transfer has been tested. Four source workstations are connected to one ATM switch with 155 Mbps link. Four destination workstations are connected to another ATM switch also with 155 Mbps link. Those ATM switches are interconnected with a 622 Mbps link, which should provide sufficient bandwidth to avoid any bottle-neck on data-links connecting the switches. Figure 4 shows the set-up of a four-to-four independent transfer test. All virtual connections work with full speed of 155 Mbps link, which is almost equivalent to 128 Mbps.

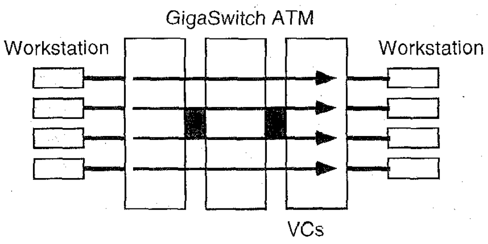

Figure 4: Four-to-four independent connection.

Figure 5 shows the set-up for a four-to-four full mesh connection test. All four source workstations keep sending as much as possible. Four virtual connections at one workstation are equally served with 32 Mbps. Total payload data rate at one workstation is 128 Mbps, which is the maximum that can be achieved over a 155 Mbps link. There is no performance reduction since there is no hot spot.

Next, we measured the effect of traffic patterns resulting in a non-uniform distribution of hot spots. Two to one merging data flow generates continuous congestion in a hot spot.

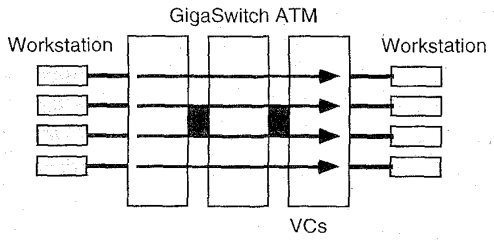

Figure 5: Four-to-four, all-to-all connection.

We found that the two merging virtual connections are equally served as shown in Figure 6.

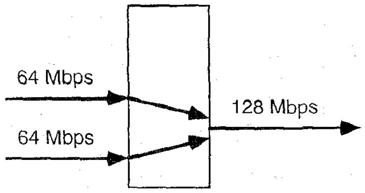

Figure 6 : Two merging data stream.

Two sources send as much data as they can till the hot spot slows down the sources. In this case, the sources have ability to send 128 Mbps payload data rate. However, they can send up to 64 Mbps since the hot-spot in the destination slow them down.

We measured the throughput of a virtual connection passing near-by the hot spot. The hot spot should not affect the other data paths. However, we observed unexpected performance reduction, as shown for the upper link in Figure 7. There is no hot spot on the upper link. Therefore it should be able to send data at the 128 Mbps payload rate. However, it is slowed down to half the expected value. The explanation is because there is some resource sharing between the hot spot and the data path.

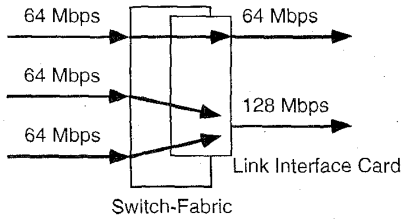

Figure 7: Two merging data stream and the data path sharing the same interface card.

The switch-fabric (the high speed back-plane) in the GIGAswitch ATM is shared by the link interface cards. However, its sharing is well managed so as not to have any problem. The problem occurs only if a data path shares the four-port interface card with the link that has the hot spot. Figure 8 shows that the data path that does not share the interface card with the link that has the hot spot is not affected by the hot spot. Therefore, the problem is in the interface card. It is because there is only one processing unit in the card

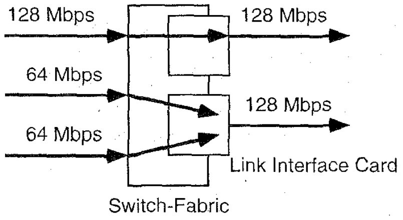

Figure 8: Two merging data stream and the data path going to another interface card.

The above problem is not severe for random traffic. This *unexpected* phenomenon happens because of the hot spot generated by continuous data flows. It is an avoidable problem. However, we should remember that under continuous data flow unexpected performance reduction may occur due to effects that are not seen by the other LAN users.

B. Flow control

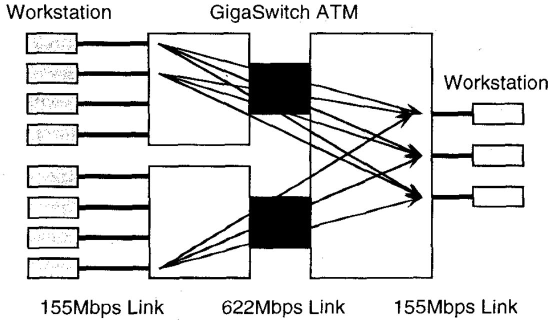

Figure 9: A schematic view of all-to-all test configuration.

We have tested all-to-all connections. Figure 9 shows the schematic set-up of the measurement. A source workstation send data sequentially to three destination workstations. Source workstations generate data independently of each other and send as much as possible. Since there are fewer workstations in destination than those in source, congestion occurs.

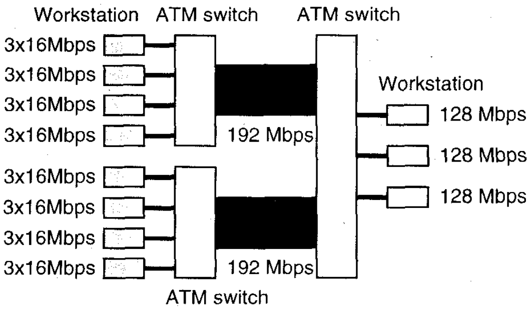

Figure 10: All-to-all test with balanced data flow.

It causes long queues to grow until source workstations are forced to slow down the data transfer.

The measured data rate on a balanced configuration is shown in Figure 10. Eight source nodes are connected to two ATM switches. The number of workstations connected to each ATM switch is four. All source nodes sent data as much as they can till hot spots slow them down. Total data rate from the each source nodes was slowed down to 48 Mbps in a well balanced pattern. Data flows between switches were 192 Mbps. The data flows merge to a maximum speed of 128 Mbps at the destination.

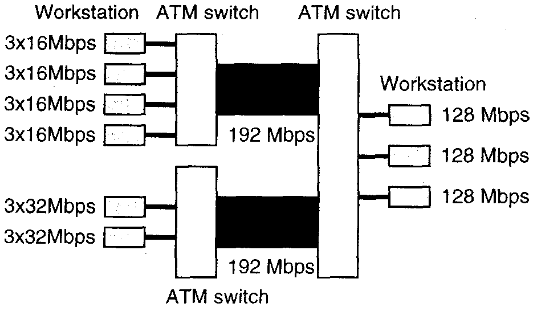

Figure 11: All-to-all test with unbalanced data flow.

The measurement on an unbalanced configuration is shown in Figure 11. Six source nodes are connected to two ATM switches. The number of workstations connected to one ATM switch is four. The number of workstations connected to the other ATM switch is two. The data flows from each source node were slowed down by the hot-spots. The data rates slowed down were not balanced in this case. Data flows between switches are balanced at 192 Mbps. The ATM switch managed to assign equal band width to each 622 Mbps inter-switch link, but, it did not manage to assign equal band width to all source nodes.

We did not anticipate this result. We expected all sources should be served equally. However, it is not achieved for the following reason. A switch does not know the bandwidth that we want to assign. At first, the switch assigns arbitrary band width to the virtual connections. If the assignment is not suitable, then a hot spot will be generated. Then, the switch knows the data flow to the hot spot should be reduced. It will balance the data flow. However, in the case of our test, all queues are full. The tails of the queues to the hot spot are not in the switch. The switch has no way to know the queue length. The only thing the switch can do is to give back-pressure to the source nodes to slow down the data transfer. In the test set-up, the back-pressure slows down the individual data transfer rate. As a consequence, the system is stabilized at that point, which has unbalanced data flow.

This result tells us that in an event builder the data flow control cannot be provided by the internal flow control mechanism of the switch alone. An external control system, which can be a part of trigger system, will be required to operate an event builder.

V. CONCLUSION

The performance of ATM switches and high speed workstations is promising for use in DAQ systems. However, we need careful control for mixed data traffic.

Permanent hot spots by "continuous" data flow may cause unexpected resource sharing. Statistical congestion does not cause such problem because sooner or later the congestion disappears and reappears at a different point. The congestion by continuous data flow stays in the same place for long time. The effects of the congestion may extend upstream.

The system may be trapped in an unexpected stable point by the built-in flow control algorithm in the switch. It is not an incorrect operation of the switch. We can not rely on the internal flow control system even of an ideal switch. If we need particular flow control, an external control system is required.

In DAQ system we know the traffic shape we want. It is possible to implement it with an external flow control system. We may use the trigger system for such purpose since it already exists in the DAQ system. The event builder architecture we have tested can provide such external flow shaping system using the ATM bit-rate flow control system [4,5].

Mixed data flows, constituted from office LAN and DAQ data traffic, may unexpectedly slow down the system. However, flow shaping with bit-rate control may solve the problem. Further study will be continued.

VI. Acknowledgments

The authors would like to thank JSPS (Japan Society for Promotion of Science) for supporting the collaboration between RCNP and Saclay. We would like to thank RCNP for allowing us to use its computer and network system for these research studies. We also thank DEC engineers, Mr. Yakoshi, Mr. Hayakawa and their colleagues, for technical discussion and set-up of the test environment.

VII. References

[1] 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. 1994.

[2] I. Mandjavidze et al., "Review of ATM, Fibre Channel and Conical Network Simulation Results," International Data Acquisition Conference '94, Fermi National lab., (Batavia) (1994),

[3] Y.Nagasaka et al., "Stability of coherent data traffic in a switching network", *IEEE Trans. on Nucl. Sci.*, Vol.43, No.1, 1996, pp.85-89

[4] D. Calvet et al., "Evaluation of a Congestion Avoidance Scheme and Implementation on ATM Network based Event Builder", Proceedings of the 2nd International Data Acquisition Workshop on Networked Data Acquisition Systems, Edited by M. Nomachi and S. Ajimura, World Scientific (1997) pp 96.

[5] D. Calvet et al., "A study of Performance Issues of the ATLAS Event Selection System based on an ATM Switching Network", *IEEE Trans. on Nucl. Sci.*, Vol.43, No.1, 1996, pp.90-98

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

[7] J. Bystricky, "A Sequential Processing Strategy for the ATLAS Event Selection", *IEEE Trans. on Nucl. Sci.*, Vol.44, No.3, 1997, pp.342-347

[8] K. Suruga and K. Hayakawa, "No Cell Loss: DIGITAL's ATM Flow Control", Proceedings of the 2nd International Data Acquisition Workshop on Networked Data Acquisition Systems, Edited by M. Nomachi and S. Ajimura, World Scientific (1997) pp 81.

[9] B. Simcoe et al., "FLOWmaster Flow-Control", DEC white paper 1995, in http://www.networks.digital.com/dr/techart/.