EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN–ECP/95–20
13 November 1995

TRIGGERING AND EVENT BUILDING RESULTS USING THE C104 PACKET ROUTING CHIP

R.W. Dobinson
CERN, Geneva, Switzerland

D. Francis
Department of Physics, Boston University,
Boston, MA 02215, USA

R. Heeley
CERN, Geneva, Switzerland and
University of Liverpool, Liverpool, England

J.R. Moonen
Eindhoven University of Technology, Eindhoven, Netherlands

Abstract

The C104 is an asynchronous 32-way dynamic packet routing chip. It has a 264 Mbytes/s bi-directional bandwidth and a 1 μsec switching latency. It offers high-density cost-effective commodity communications, which allow large switching networks to be constructed. Results are presented on the performance of this switching technology within the context of future High Energy Physics level II and level III trigger data traffic patterns.

(To be submitted to Nuclear Instruments and Methods)

Presented at the International Conference on Computing in High Energy Physics, Rio de Janeiro, 18–22 September 1995 and OBS'95, 11–13 October 1995, Swiss Institute of Technology ETH, Zürich

1 Introduction

The bandwidth and connectivity requirements of trigger and data acquisition systems for future HEP experiments at the CERN LHC and HERA-B at DESY, highlight the limitations when scaling embedded multiprocessor bus-based systems. The most prominently featured alternative in the proposed systems are large switching networks.

The performance of large switching fabrics has been studied within the context of message passing multiprocessing computers [1]. However, these studies were based on telecommunication traffic patterns which differ from those expected in the trigger systems of HEP experiments. In the latter, switching fabrics will be required to connect N sources to M destinations, where M and N are $\mathcal{O}(500)$. The traffic pattern will be 'bursty' and have different characteristics for level II and III triggers. The modelling of large switching fabrics using such traffic patterns is being performed by several groups [2, 3, 4]. We report here on the actual performance measurements of a switching fabric built from ST C104 packet routing chips [5].

The serial technology used and the general communication's performance of the network are briefly discussed, followed by the performance of the network with expected HEP trigger data traffic patterns.

2 The IEEE P1355 Standard

The ESPRIT[1)] OMI/HIC[2)] project has developed two bi-directional link protocols which form the basis of the IEEE P1355 standard [6]:

- a 100 Mbit/s Data-Strobe (DS) link,
- a 1–3 Gbit/s High-Speed (HS) link.

The work reported here is based on the the DS link protocol depicted in Fig. 1. The data line carries the data, and the strobe line only changes state when the data remains constant. In this protocol the clock is encoded, enabling autobauding at the receiver and asynchronous links. Studies on the reliability of DS links [7], up to distances of 20 m, show good reliability, with a bit error rate of less than 10^{-14}.

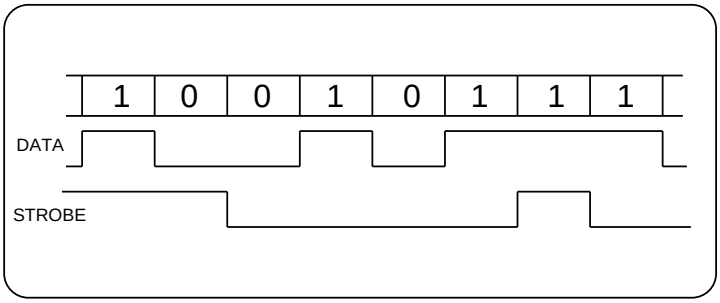

Figure 1: DS link protocol

On top of the bit level there are a further three protocol levels: the character, exchange, and packet levels. Characters are a group of consecutive bits used to represent data or control. The exchange layer describes the exchange of characters needed to ensure the proper function of a link. In particular, flow control characters are used to enable traffic flow from the link sender. This ensures that the switching fabric is lossless: no packets are lost internally due to buffer overflow. A packet is a sequence of characters

[1)] European Strategic Programme for Research and development in Information Technology.

[2)] Open Microprocessor Systems Initiative/Heterogeneous InterConnect Project.

with a specific order and format: a header, which contains routing information, a payload containing zero or more data bytes, and an end of packet marker. The protocol does not specify a specific or maximum size for a packet. Messages are sent through a network as a sequence of packets.

A family of communication devices has been developed to support the DS link protocol, which include a parallel-DS link converter and an asynchronous packet routing chip. The DS link protocol is also used by the T9000 Transputer. These components have been discussed in detail elsewhere [8, 9] and only the packet routing chip is described here.

2.1 The C104 Packet Routing Chip

The C104 is a low-latency asynchronous packet routing chip for the DS link protocol. It can be used to connect up to 32 devices, including other instances of itself, via a 32×32 non-blocking crossbar switch. The 32 links operate asynchronously, enabling packets of any length to be routed between a link pair without affecting the packet routing between any other link pair. The link speed of 100 Mbits/s allows a maximum, user data, bi-directional bandwidth across the chip of 264 Mbytes/s.

The C104 implements several advanced routing techniques to address the issues of latency, buffering and link contention. These techniques are:

• Worm-hole Routing

The implementation of worm-hole routing minimizes the latency and buffering requirements of the C104 compared to switches using message buffering with store and forward techniques. The routing decision is made as soon as the packet header arrives, the header is then sent to the chosen output link and the rest of the packet follows without being internally buffered. This implies that packets may be passing through several layers of C104s at the same time. The header, when passing through each device, creates a temporary circuit (the worm hole) through which the rest of the packet flows. As the end of the packet passes through each device the circuit closes. The packet latency across a chip has been measured to be $\sim 1\ \mu$s.

• Grouped Adaptive Routing

In any multi-stage network, efficient load balancing is a primary concern. The possibility to implement grouped adaptive routing on the C104 minimizes the effects of load imbalance [1]. If consecutive links of a C104 are used to access a common destination, they may be logically 'bundled' together, allowing a higher bandwidth to the common destination. Any packet destined for the final common destination is routed through the first available link in the 'bundle'.

• Universal Routing

It is not practical or in some cases possible to increase the number of links in a 'bundle' to match the bandwidth requirements to a destination. As a complementary method of load balancing, the C104 has been designed to perform universal routing [10]. In this technique, packets are first sent by a C104 to a random device (another C104) in a multi-stage network. At this device, the packets are then routed to their final destination. The initial randomization balances the load across the network reducing link contention, thus minimizing the formation of hot spots[3)].

[3)] An initial localized bottleneck in a network, which then propagates across the network.

3 The GPMIMD Machine

The GPMIMD machine has been designed as a switching fabric made up of 1000 DS links interconnecting 256 T9000 processors. In its present configuration 38 C104s, 25% of the full switching capability, provide full inter-connectivity between 48 processors. Six motherboards (see Fig. 2a) each carry eight T9000s and five C104s. Two switch cards, each carrying four C104s, provide the inter-motherboard connectivity. Four independent Clos networks [11] have been implemented to efficiently use the four independent DS links associated with each T9000 processor. The message latency measured between two T9000s is 7.4 μs and the unidirectional bandwidth on each of the four links is 7.0 Mbytes/s.

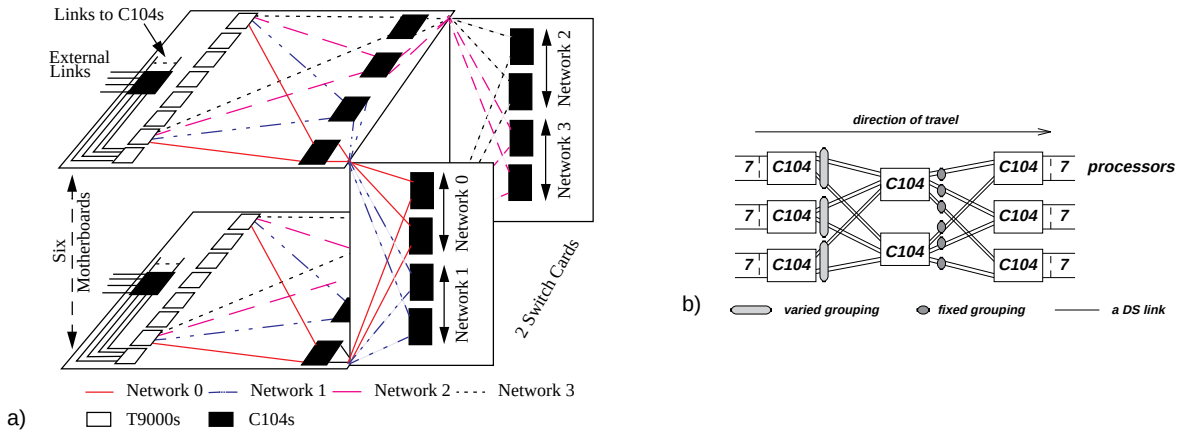

Figure 2: *(a)* The current configuration of the GPMIMD machine, and *(b)* one of its Clos networks.

The basic Clos network of the machine is shown in Fig. 2b. Each of the C104s on the left and right represent a single C104 on a motherboard, and the switch card C104s are those in the centre. The four links between each motherboard and switch card C104 may be grouped into 'bundles' of 1, 2, 3 or 4 links[4)]. Therefore, a single network may consist of 3–12 links between the motherboards and the switch cards. The single unidirectional DS link performance of 9.26 Mbytes/s means that the bandwidth of the network may be varied between $\sim$ 37–111 Mbytes/s.

3.1 General Communications Performance

The T9000s were divided into 21 sources and 21 destinations, and the average unidirectional cross-sectional bandwidth of the machine was measured for two types of traffic pattern. The T9000s were only used as data producers and consumers; they performed no processing.

In the first of these traffic patterns, sources and destinations are formed into fixed pairs; no two sources send data to the same destination. In this situation, the maximum unidirectional bandwidth allowed by the technology, network and routing algorithm are measured. The achieved unidirectional bandwidth as a function of message size, and the extent of grouped adaptive routing (3, 6, 9 or 12 links) for a single network is shown in Fig. 3a. The measured asymptotic, unidirectional bandwidth is in good agreement with the theoretically expected values to within 3%. In addition, the achieved unidirectional

[4)] Only bundles of two and four links maintain the Clos network architecture.

bandwidth scales linearly with the extent of grouped adaptive routing (3, 6, 9 or 12 link 'bundles') between the first and middle stage of the network, showing no overheads caused by its implementation. This shows that with no contention on the destination links and the maximum amount of grouped adaptive routing (12 link 'bundles'), each network has a unidirectional bandwidth of 108 Mbytes/s.

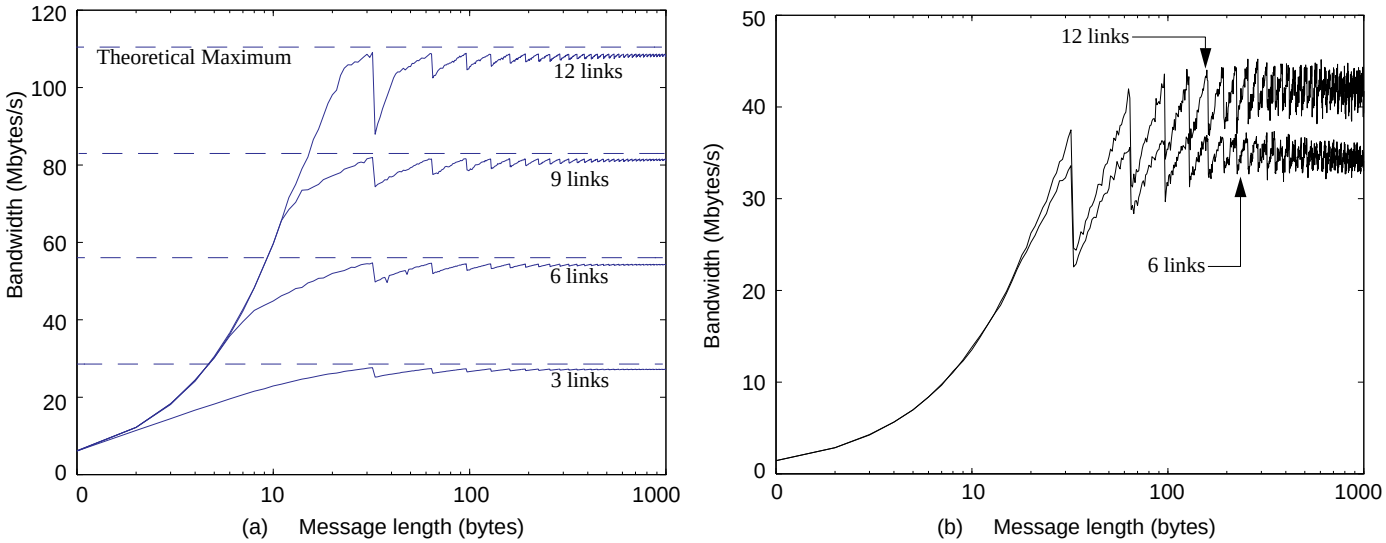

Figure 3: The unidirectional bandwidth for *(a)* fixed and *(b)* random source and destination pairings versus message length. The different curves are for different levels of grouped adaptive routing.

In the second traffic pattern, messages are sent from each of the 21 sources to the 21 destinations randomly[5]. The measured unidirectional bandwidth as a function of message size and the extent of grouped adaptive routing is shown in Fig. 3b. A significant degradation in the unidirectional bandwidth compared to the paired traffic pattern can be seen. For link 'bundles' of 6 and 12 only 61% and 38% of the unidirectional theoretical bandwidth is achieved. Although the bandwidth achieved with 12 links is greater than with six links, there is lower utilization of the available bandwidth. This indicates that under this traffic pattern, contention for the final destination links limits the performance.

4 HEP Traffic Patterns

An event is a set of data blocks, where a data block is defined as the traffic going to a single destination processor. Each source sends a message to this destination comprising of one or more packets. For example, a level II data block could correspond to a region of interest in a detector. A level III data block is defined as a complete event.

T9000 processors act as sources and destinations of event data. No computation is performed by the destination. Measurements have been carried out using predetermined event sequences, stored by the sources as look-up tables. An overall supervisor function has not been implemented.

An example of the level II trigger traffic pattern for a sub-detector is shown in Fig. 4a for two events. The first event consists of two data blocks: each data block is distributed over two sources and routed to a separate destination. The second event consists of three data blocks: one data block distributed over three sources and two data blocks originating from a single source.

[5] Known as telecommunications traffic.

The level III trigger traffic pattern is shown in Fig. 4b. Event 1(2) consists of a single data block which is distributed over 4(3) sources and routed to a single destination. The currently expected level II and III traffic patterns of a sub-detector in the ATLAS experiment [3] at CERN are summarized in Table 1.

Table 1
Trigger level II and III traffic patterns for a single sub-detector

Level	Frequency	Data blocks per event	Sources per data block	Data block size
II	100 KHz	$\mathcal{O}(3)$	$\mathcal{O}(3)$	0.2–1.2 Kbytes
III	1 KHz	1	$\mathcal{O}(300)$	$\mathcal{O}(1$ Kbyte)

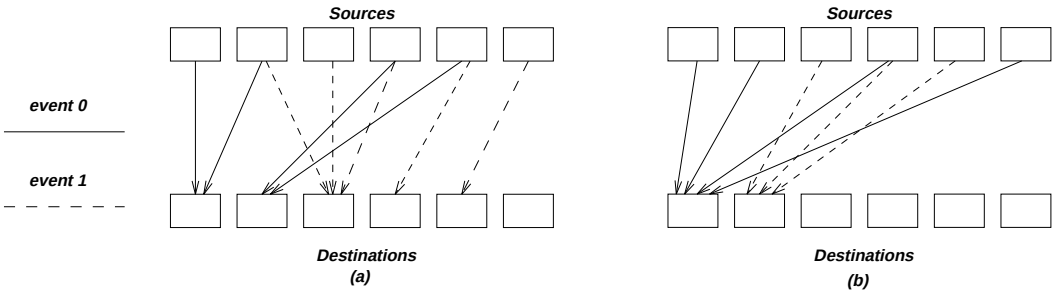

Figure 4: An example of the expected *(a)* level II and *(b)* level III traffic patterns.

The performance, in terms of achievable event rates and bandwidths, of a single network with these types of traffic patterns has been measured. The total performance of the machine is simply four times the single network result. The number of source and destination processors used is typically 18, and the maximum amount of grouped adaptive routing (12 links) is applied between the first and second C104 layers. Data blocks are distributed equally over the number of participating sources. No studies have been performed with varying packet size as only packets of up to a maximum length of 32 bytes are supported by the T9000.

4.1 Level II Triggering Performance

The achieved event rate and bandwidth, for a configuration of 18 sources and 18 destinations, as a function of the attempted event rate is shown in Figs. 5a and 5b, for different data block sizes. In these measurements the number of data blocks per event and sources per data block has been fixed to three and one, respectively. A random selection is made as to which three sources participate per event, and the choice of which three destinations per event is now based on a round robin selection. At the maximum achievable event frequencies for the different data block sizes, 70 Mbytes/s out of the maximum network bandwidth of 111 Mbytes/s is achieved.

The loss of bandwidth is due to the random selection of which sources participate per event. This criteria leads to queuing in the sources and link contention at the destinations. Figures 5c and 5d are equivalent to Figs. 5a and 5b except that the choice of participating sources per event is now based on a round robin selection. In this situation 97% of the maximum bandwidth is achieved.

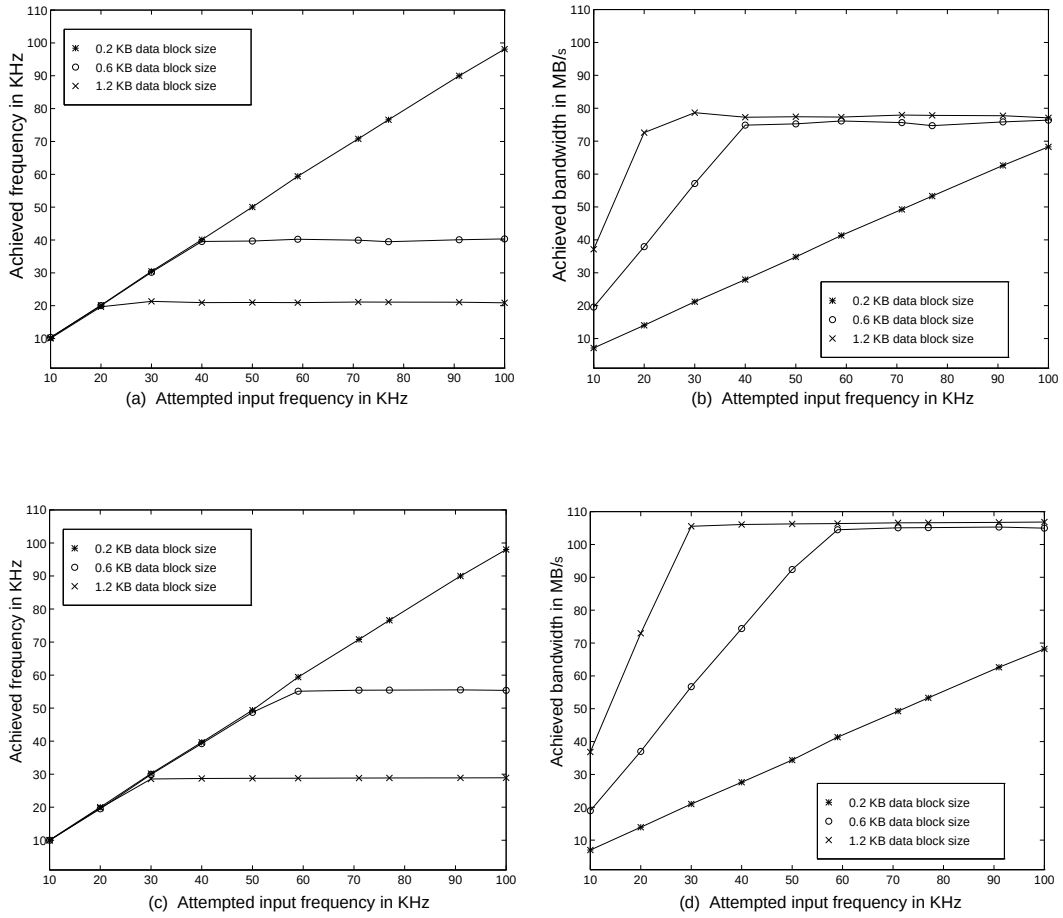

Figure 5: The achieved event frequency *(a)* and *(c)* and bandwidth *(b)* and *(d)* for a simple level II traffic pattern (see text). Which sources participate per event is based on a random selection in *(a)* and *(b)* and a round robin selection in *(c)* and *(d)*.

In Fig. 6 the achieved event frequency and bandwidth as a function of the number of data blocks per event is shown for the different number of sources per data block and a fixed data block length of 1.0 Kbytes. For more than two data blocks per event the achieved frequency does not vary significantly with the number of sources per data block.

The performance of the network, as a function of the number of sources and destinations it connects, is demonstrated in Figs. 7a and b. The achieved event frequency and bandwidth are shown as a function of the number of connected sources (= destinations) for data block sizes of 0.2 and 1.2 Kbytes, three and five data blocks per event and two sources per data block. The performance scales linearly from 3 to 18 sources and the achieved event rates depend on the the number of data blocks per event. Event rates of 45 and 18 KHz are achieved for data block sizes of 0.2 and 1.2 Kbytes, respectively. The full capacity of the machine for these data block sizes is therefore 180 and 72 KHz, respectively.

The results presented have used the ability of the T9000 to efficiently multiplex data on to multiple virtual links[6], in this case five destinations on to a single physical link. In Fig. 8, the bandwidth and achieved event frequencies for different numbers of virtual links

[6] A virtual link is a logical communication path between two processes.

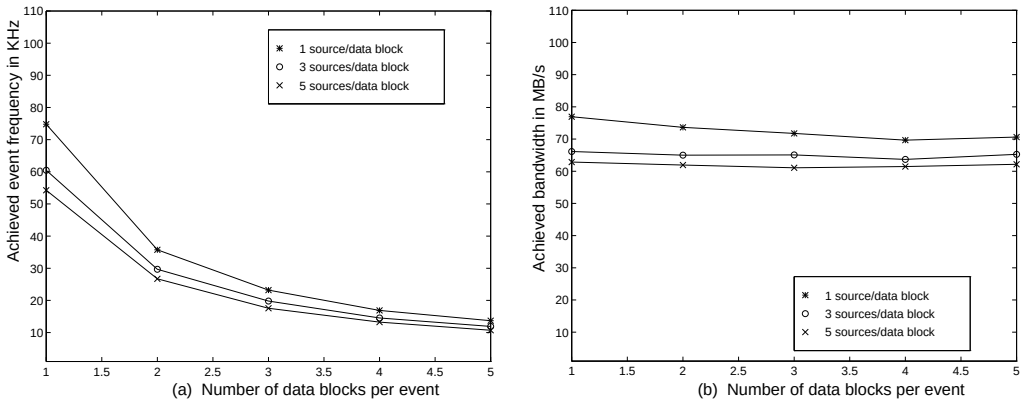

Figure 6: *(a)* The achieved event frequency and *(b)* bandwidth as a function of data blocks per event for the different number of sources per data block.

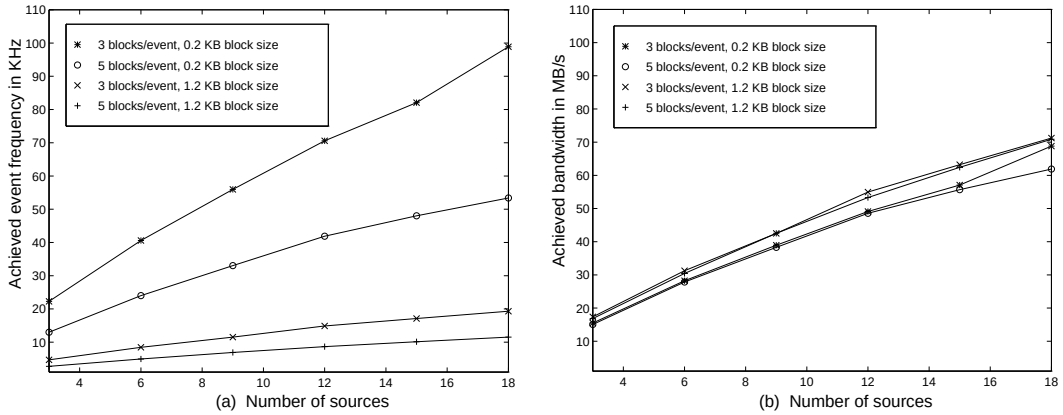

Figure 7: *(a)* The achieved event rate and *(b)* bandwidth plotted against the number of connected sources (destinations).

are shown. It can be seen that the achieved event frequency depends on the number of virtual links and is constant for more than five virtual links per physical link. In the case of a single virtual link the end-to-end packet flow control limits the achievable bandwidth, as a source must wait for a packet acknowledgment from a destination. This effect is reduced by using multiple virtual links. While waiting for a packet acknowledgment on one virtual link, a source may send a packet on another virtual link to another destination.

4.2 Level III Trigger Performance

The achieved event frequency as a function of the number of sources per data block, in a network of 21×14, is shown in Fig. 9a for three different data block sizes. The achieved event rate varies linearly with the number of sources per data block, showing a better utilization of the network. At the maximum input utilization (21 sources per data block) for the three data block sizes considered, $\sim$ 50 Mbytes/s out of a maximum available bandwidth of 111 Mbytes/s is used.

In Fig. 9b the communication latency per source as a function of the event number is shown for three of the sources in the above configuration. For the first event, all sources are competing for the same destination link resulting in a high latency for every source. The latency then decreases with the event number, settling down to a steady state, as different

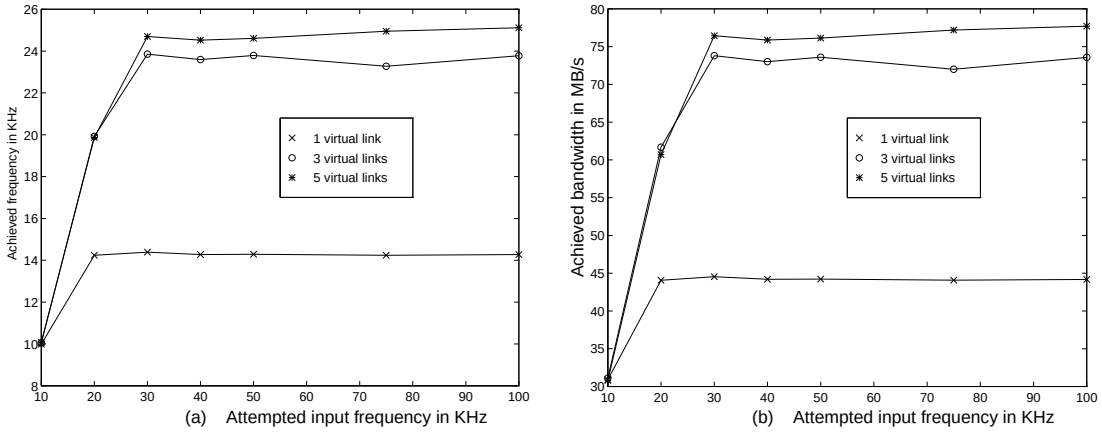

Figure 8: *(a)* The achieved event rate and *(b)* the bandwidth plotted against the attempted frequency for different amounts of multiplexing.

sources work on different events. The fact that the sources are working on different events implies different destinations are being used in parallel, and therefore less contention occurs on the destination links. This effect is also shown by the dotted curves in Fig. 9a. In this figure, the sources have been selected, where possible, by a round robin criteria. For example, in the case of five active sources, the first five sources send data block 0 and the second five sources send data block 1. This continues up to message 3, and then the first five sources send data block number 4. This artificially spreads the load across the number of sources on an event basis and results in an improved event rate for 5 and 10 active sources.

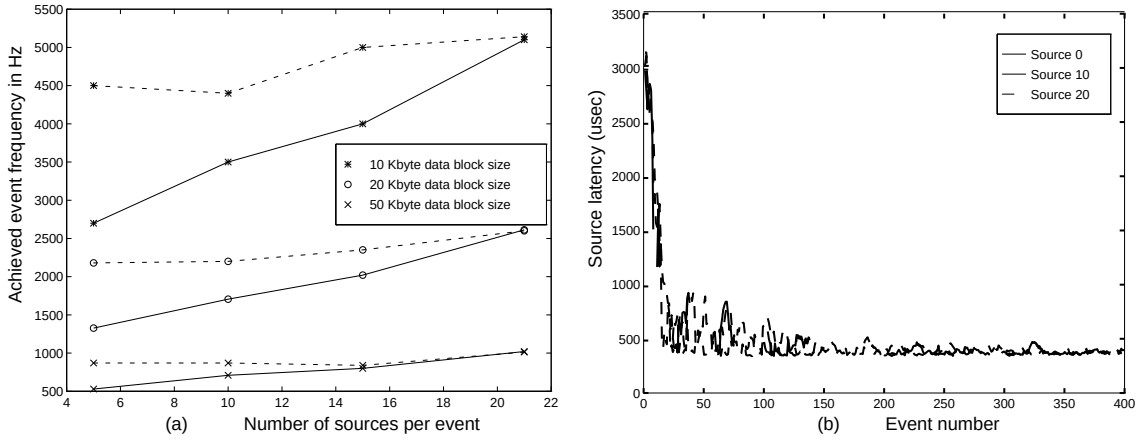

Figure 9: *(a)* The sustained frequency plotted against the number of sources/event for level III traffic patterns. The solid line is for a random source selection while the dotted line is for a round robin selection. *(b)* The source event latency plotted against the event number for 3 of 21 sources per event.

In order to investigate the effect of communication's latency on performance the message latency in the sources has been artificially increased. The results are shown in Fig. 10a, for 21 active sources and 14 destinations. It is clear that there is a very strong dependence of performance on message latency and that an efficient node to network interface, together with lightweight communication protocols, is vital.

In Fig. 10b the performance of the network, in terms of the number of connected sources (=destinations), is shown. The network performance scales linearly with the number of connected sources. This is a design feature of the GPMIMD machine communications network which can be scaled up to 1000 links.

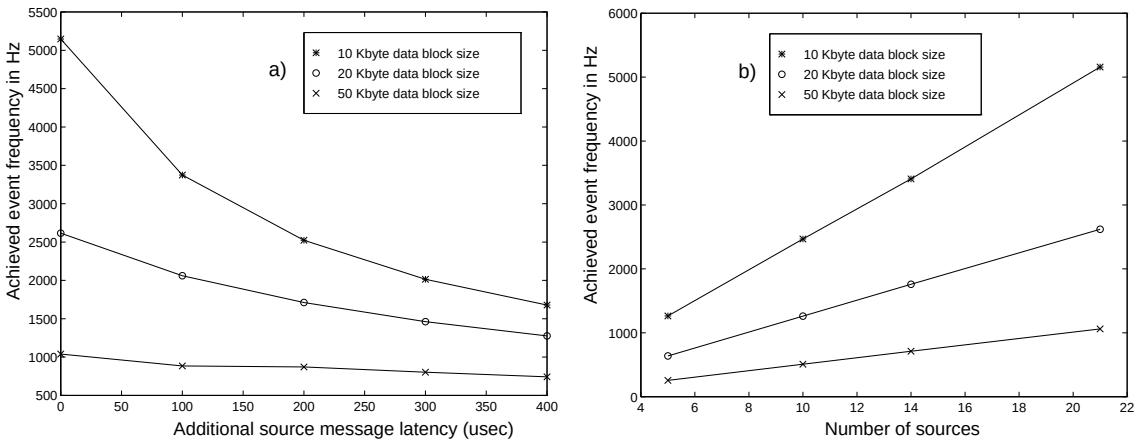

Figure 10: *(a)* The achieved frequency plotted against the additional source message latency. *(b)* The achieved frequency plotted against the attempted frequency under level III traffic patterns for different data block sizes.

5 Summary and Outlook

Measurements on the performance of a 40 node Clos switching fabric have been presented. This fabric is based on the ST C104 packet routing chip, and the DS link protocol which is being standardized by the IEEE P1355 Committee. The DS link and associated packet routing chip, the C104, meet their design specifications. In the present configuration of the GPMIMD machine, only 25% of the full switching capacity is used and only 20% of the maximum number of processing nodes have been installed. By the end of 1995 the switching fabric will be upgraded to four switch cards and the number of supported processing nodes increased to 64.

Under the currently expected traffic patterns of sub-detectors in the level II and III triggers of future experiments [3], bandwidths of 50–70 Mbytes/s have been achieved on a single network, corresponding to $\sim$ 50–60% utilization of the theoretical bandwidth. In Table 2 the measured results for a level II traffic pattern of three data blocks per event and two sources per data block, and a level III traffic pattern of 21 sources per event are summarized.

Table 2
Measured results using level II and III traffic patterns

	Data block size			
	Level II		Level III	
	0.2 Kbytes	1.2 Kbytes	10 Kbytes	50 Kbytes
Event frequency	A	B	C	D
Bandwidth	A	B	C	D
Latency	A	B	C	D

These results were obtained when:

- using the ability of the T9000 to multiplex virtual links onto a single physical link,
- writing the software in OCCAM [12], the native language of Transputers.
- using grouped adaptive routing on the C104.

It appears that currently available technology can meet many, if not all, of the switching and bandwidth requirements of future HEP trigger systems. This could be achieved by scaling or replicating the Clos networks used in these measurements. Simulations have shown that by adding layers of C104s, the Clos network architecture used may be arbitrarily scaled [1]. However, networks that allow the bandwidth to be scaled with the number of nodes require an increase in the number of switches which is more than proportional. Replication of the network is potentially a more efficient use of resources but may require that data sources be able to drive several networks simultaneously.

Other questions remain to be addressed:

- What is the dependence of network utilization, scalabilty and latency on network topology?
- What are the cost and pratical advantages of many smaller or lower speed (replicated) networks compared with fewer large or higher speed networks?
- What is the effect of different packet sizes on network performance?
- What are the issues involved in network management, monitoring, error handling and fault tolerance?

To address these questions a large 1000 node variable topology switching fabric [8] is being built at CERN as part of the MACRAMÉ project and should start to provide results in the summer of 1996.

Acknowledgments

We gratefully acknowledge the support provided by the European Union via the GPMIMD project 5404, and one of us would like to thank the US National Science Fondation for funding. We would also like to thank members of the ATLAS level II trigger group for valuable discussions.

References

[1] A. Klein, Interconnection Networks for Universal Message-Passing Systems, *Proc. ESPRIT Conference '91*, pp. 336–351, Commision for the European Communities, Nov. 1991, ISBN 92–826–2905–8.

[2] NEBULAS A High Performance Data-Driven Event Building Architecture based on an Asynchronous Self-Routing Packet-Switching Network, CERN/DRDC/92–14, CERN/DRDC/92–47.

[3] The ATLAS Technical Proposal, CERN/LHCC/94–43, LHCC/P2, ISBN 92–9083–067–0.

[4] The CMS Technical Proposal, CERN/LHCC/94–38, LHCC/P1.

[5] The ST C104 Asynchronous Packet Switch, Preliminary Data sheet, June 1994. SGS THOMSON Microelectronics.

[6] IEEE Draft Std. P1355, Standard for Heteregeneous InterConnect (HIC). Low Cost Low Latency Scalable Serial Interconnect for Parallel System Construction. IEEE Inc., 1995.

[7] S. Haas, X. Liu and B. Martin, Long Distance Differential Transmission of DS Links over Copper Cable (CERN). http://www.ac.uk/parallel/vendors/inmos/ieeehic/copper.ps.gz.

[8] B. Martin et. al., Realisation of a 1000 node High Speed Packet Switching Network, Proc. *International Symposium on Problems of Modular Information Computer Systems and Networks*, St. Petersburg, 1995.

[9] R. Heeley et. al., The Application of the T9000 Transputer to the CPLEAR Experiment at CERN, CERN/ECP 95–8, to be published in Nucl. Instrum. Methods.

[10] L.G. Valiant, A scheme for fast parallel communications, *SIAM J. Comput.* **11** (1982) 350–361.

[11] C. Clos, A Study of Non-blocking Switching Networks, Bell Syst. Tech. J. **32** (1953) 406–424.

[12] Occam 2 reference manual (Prentice Hall, U.K., 1988) ISBN 0–13–629312–2.