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

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 allows large switching networks to be constructed. Results are presented on the performance of this switching technology within the context of future high energy physics experiments.

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 in scaling embedded multi-processor bus based systems. The prominently featured alternative in the proposed systems is large switching networks.

The performance of large switching fabrics has been studied within the context of message passing multi-processing computers[1]. However, these studies were based on telecommunications traffic patterns which differs to 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 of order 250. The traffic pattern will be 'bursty' and have different characteristics for level II and III triggers. The modeling of large switching fabrics under 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 C104 packet routing chips[5].

The serial technology used and the general communications performance of the network are briefly discussed, followed by the performance of the network under expected HEP data traffic patterns.

2 The IEEE 1355 Standard

The ESPRIT[a] OMI/HIC[b] project has developed two bi-directional link protocols which form the basis of the IEEE P1355 standard[6], these are:

- 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 figure 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, have measured a bit error rate of less than 10^{-14}.

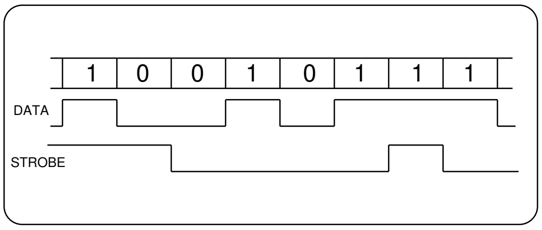

Figure 1: DS link protocol

On top of the bit level there are a further three levels of protocol: the character, exchange and packet levels. Characters are a group of consecutive bits used to represent data or control information. The exchange layer describes the exchange of characters 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 characters are lost internally due to buffer overflow. A packet is a sequence of characters 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 packet size. 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, these 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 its self, via

[a] European Strategic Program for Research and development in Information Technology
[b] Open Microprocessor Systems Initiative / Heterogeneous Inter-Connect Project

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 bi-directional bandwidth across the chip of 264 Mbytes/s.

The C104 supports 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 minimises 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 subsequently 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 C104 has been measured to be ~ 1 μs.

• Grouped Adaptive Routing

In any multi-stage network, efficient load balancing is a primary concern. The support of grouped adaptive routing by the C104 minimises 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 support 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 randomisation balances the load across the network reducing link contention and therefore minimising the formation of hot spots[c] .

3 The GP-MIMD Machine

The GP-MIMD machine has been designed as a switching fabric of 1000 DS links interconnecting up to 256 T9000 processors. In its present configuration 38 C104s provide full inter-connectivity between 48 processors. Six motherboards (see figure 2a) each carry 8 T9000s and 5 C104s. Two switch cards, each carrying 4 C104s, provide the inter-motherboard connectivity, corresponding to 25% of the full switching capability. Four independent Clos[11] networks have been implemented to efficiently use the four DS links associated with each T9000 processor.

The basic Clos network of the machine is shown in figure 2b. Each C104 on the left and right represent a single C104 on a motherboard and the switch card

[c] An initial localised bottleneck in a network, which then propagates across the network

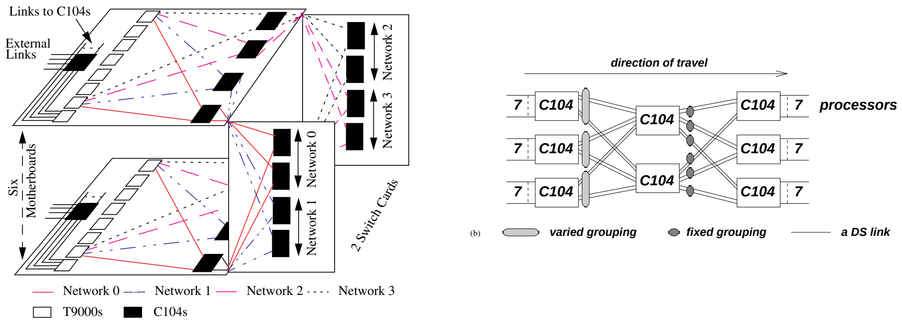

Figure 2: The *(a)* The current configuration of the GP-MIMD machine and *(b)* one of its Clos networks.

C104s are those in the center. The 4 links between each motherboard and switch card C104 may be grouped into 'bundles' of 1,2,3 or 4 links. Therefore, a single network may consist of 3-12 links between the motherboards and the switch cards. The theoretical uni-directional DS link performance of 9.26 Mbytes/s allows the bandwidth of the network to be varied between $\sim$37-111 Mbytes/s.

The measured message latency between two directly connected 20 MHz T9000s is 7.4 μs[d] and the measured uni-directional bandwidth on each of the four links is 7.0 Mbytes/s.

3.1 General Communications Performance

The T9000s were divided into 21 sources and 21 destinations and the average uni-directional cross-sectional bandwidth of the machine measured for two types of traffic pattern. These two traffic patterns give information on the general communications performance. Measurements using HEP trigger traffic patterns are presented in the next section. 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 uni-directional bandwidth allowed by the technology, network and routing algorithm are measured. The achieved uni-directional 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 figure 3a. The measured asymptotic, uni-directional bandwidth is in good agreement with the theoretically expected values to within 3%. In addition, the achieved uni-directional bandwidth scales linearly with the extent of grouped adaptive routing (3,6,9 or 12 links) between the first and middle stage of the network, showing no overheads due to its implementation. This shows that with no contention on the destination links and the maximum amount of grouped adaptive routing, each network has a uni-directional bandwidth of 108 Mbytes/s.

In the second traffic pattern, messages are sent from each of the 21 sources to the 21 destinations randomly[e] . The measured uni-directional bandwidth as a function

[d] Message latency is defined as the time to send a "zero length" message.
[e] A form of telecommunications traffic

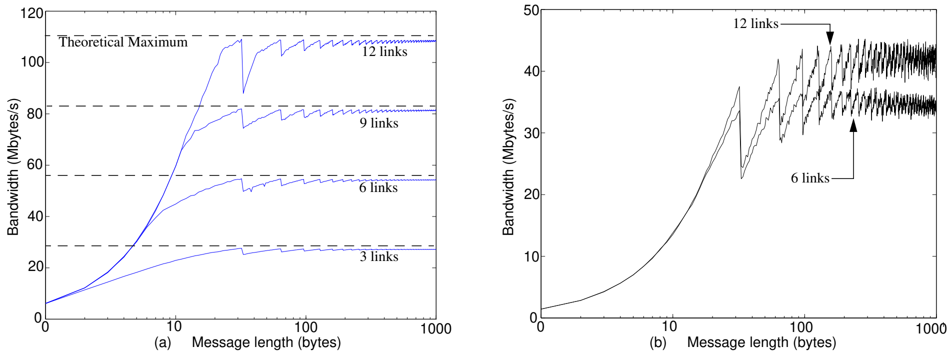

Figure 3: The uni-directional bandwidth for *(a)* fixed and *(b)* random source and destination pairings versus message length. The different curves are for different levels of grouped adaptive routing.

of message size and grouped adaptive routing is shown in figure 3b. A significant degradation in the uni-directional bandwidth compared to the paired traffic pattern can be seen. For 6 and 12 grouped links only 61% and 38% of the uni-directional theoretical bandwidth is achieved. Although the bandwidth achieved with 12 links is greater than with 6 links, there is lower utilisation of the available bandwidth. This indicates that under this traffic pattern, contention on the final destination links limits the performance.

4 HEP Traffic Patterns

In our terminology, an event consists of data blocks which are defined as the traffic going to a single destination processor. The level II data block corresponds to a subset of a sub-detectors event data and a level III data block corresponds to the total event data of a sub-detector. A data block consists of a number of fragments. Each fragment emanates from a different source and is sent as an individual message.

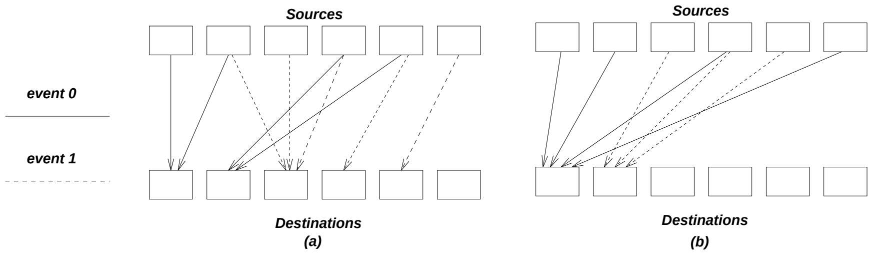

Figure 4: An example of the expected *(a)* level II and *(b)* III traffic pattern.

An example of the level II trigger traffic pattern for a sub-detector is shown in figure 4a for two events. The first event consists of 2 data blocks: each data block is distributed over two sources and routed to a separate destination. The second event consists of 3 data blocks: one data block is distributed over 3 sources and 2 data blocks originate from a single source.

The level III trigger traffic pattern is shown in figure 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 summarised in table 1.

Table 1: The traffic patterns of trigger levels II and III for a typical sub-detector.

Level	Frequency	data blocks per event	sources per data block	data block size
II	100 KHz	1-5	~ 3	0.2-1.2 Kbytes
III	1 KHz	1	~ 256	10-350 Kbytes

In our model, T9000 processors act as sources and destinations of event data. No computation is performed by the destinations. Measurements have been carried out using pre-determined event sequences, stored by the sources as look-up tables.

The performance, in terms of achievable event rates and bandwidths, of a single network under these types of traffic patterns has been measured. The total performance of the machine is 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 figures 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 3 and 1 respectively. To model the level 1 trigger a random selection is made as to which three sources participate per event. The allocation of data blocks to destination processors per event is based on a round robin selection. At the maximum achievable event frequencies, 70 Mbytes/s out of the maximum theoretical network bandwidth of 111 Mbytes/s is achieved.

The loss of bandwidth is due to the random selection of which sources participate per event. Sources may have to participate in a number of consecutive events, causing some events to be delayed as the temporarily required bandwidth at a source exceeds that available. This queuing of events in the sources results in network under utilisation: messages are queued on a busy source. The network is under utilised as idle sources may exist. This effect was demonstrated by selecting sources based on a round robin criteria, thus removing the queuing of consecutive events in the sources. The results in figure 5c and 5d for this situation show a 97% bandwidth utilisation and a corresponding increase in event rate, confirming event queuing in the sources.

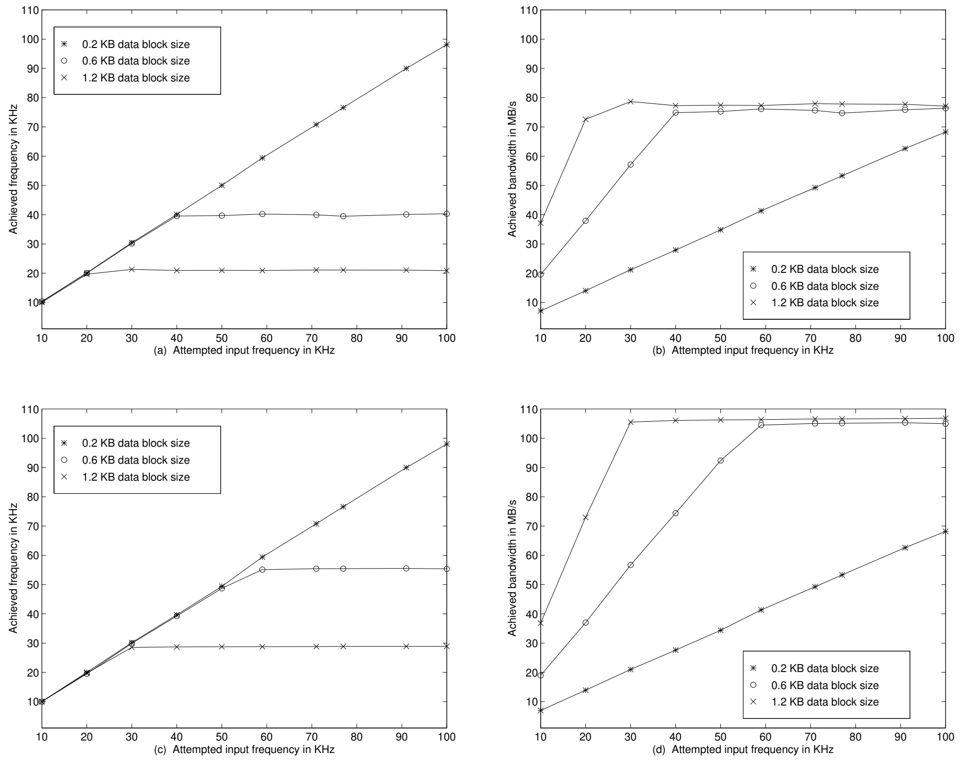

Figure 5: The achieved event frequency *(a)*,*(c)* and bandwidth *(b)*,*(d)* for a simple level II traffic pattern (see text). Which sources participate per event is based on a random selection in *(a)* and *(b)* and a round robin selection in *(c)* and *(d)*.

In figure 6 the achieved event frequency and bandwidth as a function of the number of data blocks per event is shown for different numbers of sources per data block and a fixed data block length of 1.0 Kbytes. The achieved bandwidth across the switch is independent of these parameters and the achieved event frequency decreases as the number of data blocks per event increases.

The performance of the network, as a function of the number of active[f] , sources and destinations is demonstrated in figures 7a and 7b. The achieved event frequency and bandwidth are shown as a function of the number of active sources (=destinations) for data block sizes of 0.2 and 1.2 Kbytes, 3 and 5 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 100 and 20 KHz are achieved for data block sizes of 0.2 and 1.2 Kbytes

[f] Active means sources and destinations actually using the network.

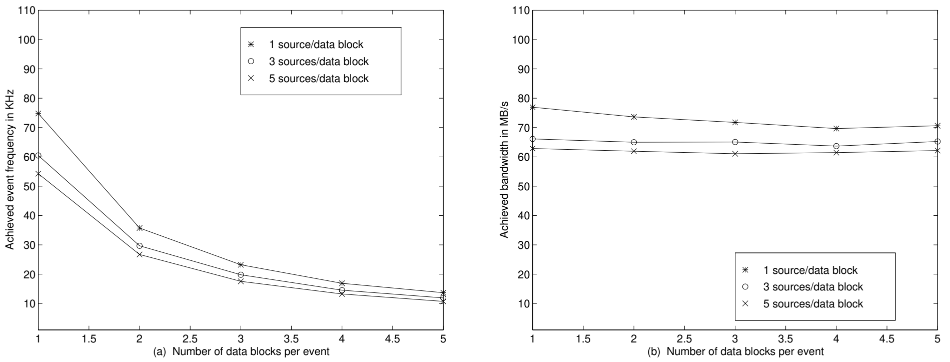

Figure 6: The *(a)* achieved event frequency and *(b)* bandwidth as a function of data blocks per event for different numbers of sources per data block.

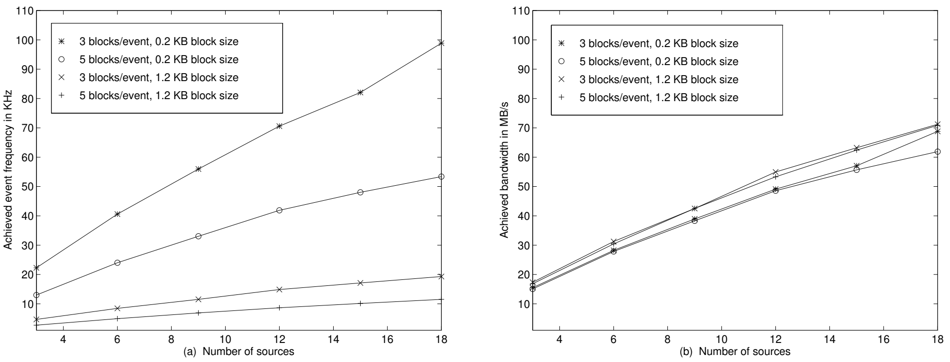

Figure 7: *(a)* The achieved event rate and *(b)* bandwidth versus the number of connected sources (destinations).

respectively. The full capacity of the machine with these parameters is therefore 400 and 80 KHz respectively.

The results presented have used the ability of the T9000 to efficiently multiplex virtual links[g] onto a single physical link. In figure 8, the bandwidth and achieved event frequencies for different numbers of virtual links are shown. It can be seen that the achieved event frequency depends on the number of virtual links and is constant for more than 5 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. In these measurements 5 virtual links have been used.

[g] A virtual link is a logical communication connection between two processes.

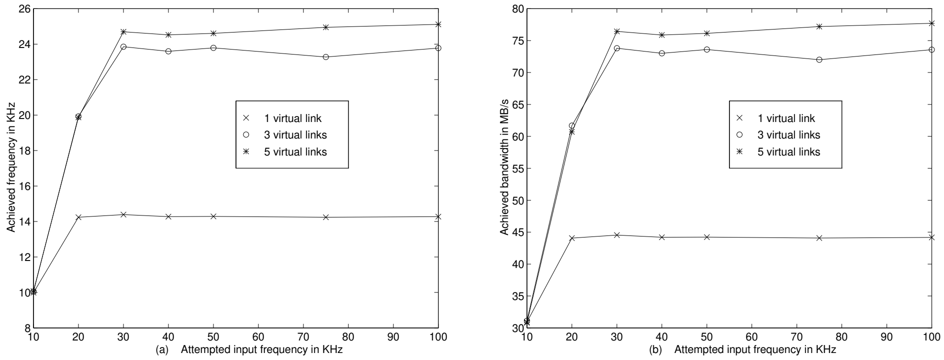

Figure 8: *(a)* The achieved event rate and *(b)* bandwidth versus the attempted frequency for different amounts of multiplexing.

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 figure 9a for 3 different data block sizes. The achieved event rate varies linearly with the number of sources per data block, due to a better utilisation of the network. At the maximum input utilisation (21 sources per data block) for the 3 data block sizes considered, ~50 Mbytes/s out of a theoretical maximum bandwidth of 111 Mbytes/s is achieved.

In figure 9b the communication latency per source as a function of 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 each source. The latency then decreases with event number, settling down to a steady state, as different sources are sending different events. The fact that 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 figure 9a. In this figure the sources have been selected, where possible, by a round robin criteria. For example, in the case of 5 sources per data block, the first five sources send data block 0 and the second five sources send data block 1. This continues up to data block 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 sources per data block.

In order to investigate the effect of communications latency on performance, the message latency in the sources has been artificially increased. The results are shown in figure 10a, for 21 sources per data block and 14 destinations. There is a very strong dependence of performance on message latency. An efficient node to network interface, together with light weight communication protocol, is vital to efficiently utilise the available network bandwidth.

The network performance scales linearly with the number of active sources (=destinations), see figure 10. This is consistent with the design specification of the GP-MIMD machines communications network; it can be scaled up to 1000 links.

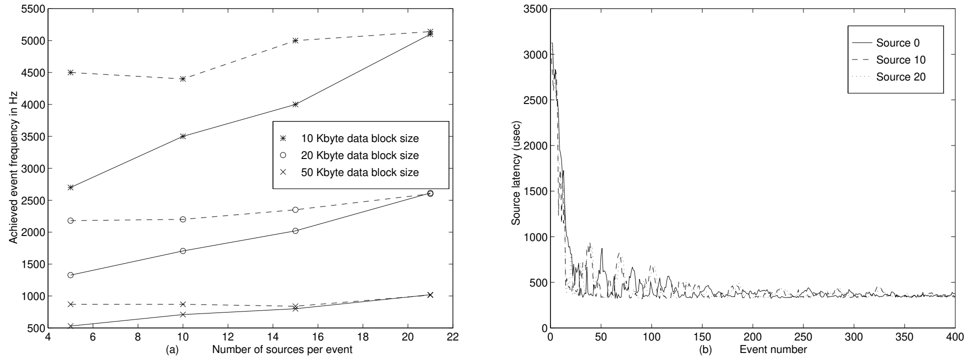

Figure 9: *(a)* The sustained frequency versus number of sources/event for level III traffic patterns. The solid line is for a random source selection while the dotted line is for a round robin selection. *(b)* The source event latency versus event number for 3 of 21 sources per event.

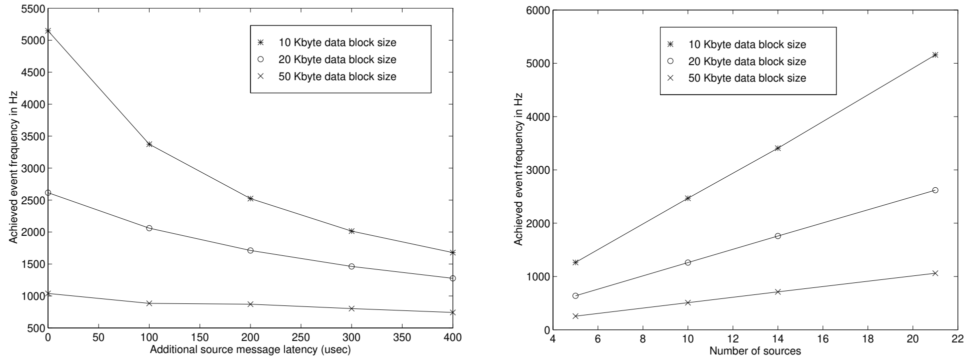

Figure 10: *(a)* The achieved frequency versus the additional source message latency. *(b)* The achieved versus attempted frequency under level III traffic patterns for different data block sizes.

5 Summary and Outlook

Measurements have been presented on the performance of a 40 node Clos switching fabric. This fabric is based on the C104 packet routing chip and the DS link protocol which is part of the IEEE 1355 standard. The DS link and associated packet routing chip, the C104, have been shown to meet their design specifications. In the present configuration of the GP-MIMD machine, only 25% the full switching capacity is used and only 20% of the maximum number of processing nodes are installed. By the end of '95 the switching fabric will be upgraded to 4 switch cards and the number of supported processing nodes increased to 64. The results presented here, will be used to calibrate models of C104 networks. These models will then be used to perform detailed simulations of the networks required for future large HEP experiments.

Under the currently expected traffic patterns of sub-detectors in the level II and III triggers of future experiments[3], bandwidths of 50-75 Mbytes/s have been achieved on a single network, corresponding to ~50-70% utilisation of the theoretical bandwidth. In table 2 the measured results for a level II traffic pattern of 3 data blocks per event and 2 sources per data block and a level III traffic pattern of 21 sources per event are summarised.

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 (KHz)	98.1	19.5	5.1	1.02
Bandwidth (Mbytes/s)	68.3	72	51	51

These results have been obtained under the following conditions:

- a single network (out of the available four) in use
- using grouped adaptive routing on the C104s
- redundant links on the C104s
- using the T9000s ability to multiplex virtual links onto a single physical link
- software written in OCCAM[12], the native language of Transputers

It appears that currently available technology can meet many of the switching and bandwidth requirements of future HEP trigger systems. This would be achieved by scaling or replicating the Clos networks used in these measurements. By adding layers of C104s, the Clos network architecture may be arbitrarily scaled. 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 simultaneously drive several networks.

Other questions remain to be addressed:

- What is the dependence of network utilisation, scalability and latency on network topology ?
- What are the cost and practical advantages of many smaller or lower speed (replicated) networks versus 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 ?
- How do the currently available serial interconnects compare ?

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 Esprit GP-MIMD project 5404 and one of us would like to thank the US National Science Foundation 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, Commission 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, DRDC/P36.
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 Micro-electronics.
6. IEEE Draft Std. P1355, Standard for Heterogeneous Inter-Connect (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.hensa.ac.uk/parallel/vendors/inmos/ieee-hic/copper.ps.gz.
8. B. Martin et. al., Realisation of a 1000 node High Speed Packet Switching Network, *International Symposium on Problems of Modular Information Computer Systems and Networks* , 1995.
9. R. Heeley et. al., The Application of the T9000 Transputer to the CPLEAR Experiment at CERN, CERN/ECP 95-8, accepted for publication in Nuclear Instruments and Methods in Physics Research.
10. L.G. Valiant, A scheme for fast parallel communications, *SIAM J. on Computing*, **11** (1982).
11. C. Clos, A Study of Non-blocking Switching Networks, Bell Systems Technical Journal 32, 1953.
12. Occam 2 reference manual, Prentice Hall. ISBN 0-13-629312-2.