Results from the Macramé 1024 Node Switching Network

S. Haas [a,b], D.A. Thornley [a,c], M. Zhu [a,b], R.W. Dobinson [a,b,d], R. Heeley [a], N.A.H. Madsen [a,d], B. Martin [a]

[a] *CERN, 1211 Geneva 23, Switzerland*

[b] *University of Liverpool, Liverpool, UK*

[c] *University of Kent, Canterbury, UK*

[d] *RHBNC, University of London, London, UK*

The work reported involves the construction of a large modular testbed using IEEE 1355 DS link technology. A thousand nodes will be interconnected by a switching fabric based on the STC104 packet switch. The system has been designed and constructed in a modular way in order to allow a variety of different network topologies to be investigated. Network throughput and latency have been studied for different network topologies under various traffic conditions. We also present results of studies carried out with traffic patterns expected for the ATLAS second level trigger system.

Key words: IEEE1355; DS-Links; network performance measurements; packet switching networks

1 Introduction

Data acquisition and trigger systems for LHC experiments will depend critically on the use of high speed point to point links and packet switches [1,2]. A range of alternative technologies are presently under study including, Fibre Channel [3], ATM [4], SCI [5] and IEEE 1355 [6] (DS and HS links).

To date, practical experience with packet switching networks in High Energy Physics (HEP) has been confined to relatively small systems and there are no experimental results on how the performance of such systems will scale up to several hundred or even several thousand nodes.

Theoretical studies [7,8] have been carried out for large IEEE 1355 networks of up to one thousand nodes connected by different switching topologies. However, the traffic patterns used are not those relevant to HEP experiments, but rather those found in telecommunications systems or massively parallel computers.

We present results obtained on a large modular testbed using DS link and switch technology. One thousand nodes will be interconnected by a switching fabric based on the 32 way STC104 packet switch [9]. The system has been designed and constructed in a modular way to allow a variety of different network topologies to be investigated (Clos [10], hypercube, grid, torus, etc.).

The Macramé testbed has been constructed using three basic building blocks:

- Traffic nodes, which can send and receive data at the nominal link speed of 100 Mbits/s with a programmable traffic profile.
- Packet switch units based on the STC104.
- timing nodes, which transmit and analyse time stamped packet in order to perform latency measurements.

Figure 1 shows a picture of the testbed architecture, the topology shown is a two dimensional grid (2D grid). A detailed description of the testbed architecture can be found in [11], preliminary results have been reported at the DAQ'96 workshop [12].

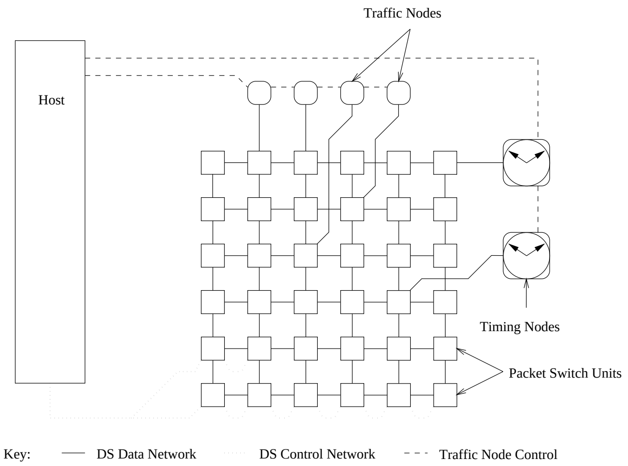

Fig. 1. Architecture of the Macramé testbed

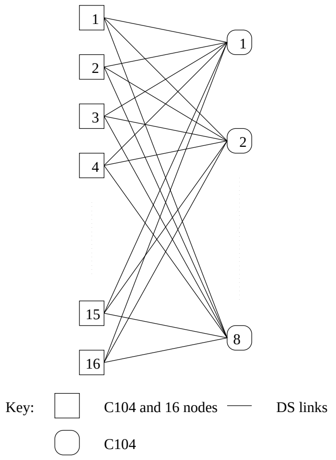

Fig. 2. 256 node Clos network

Network throughput and latency are being studied for various traffic conditions as a function of topology and network size. Results have been obtained using a 400 node 2D grid and a 256 node multistage Clos network, as show in Figure 2.

This work is carried out within the framework of the European Union's Esprit program as part of the Macramé project (Esprit project 8603).

2 Results

2.1 Network latency for Clos networks

Figure 3 shows the latency as a function of the aggregate network throughput for three different size Clos networks. The traffic pattern is random, i.e. transmitting nodes send 64 byte packets to a destination chosen from a uniform distribution. The results are produced by varying the network load and measuring the corresponding throughput and latency values. It can be seen that the average latency increases exponentially as the network throughput approaches saturation. To achieve low average latencies the applied network load must therefore be below the saturation throughput.

Figure 4 shows the probability that the latency of a packet will be greater than a given value for various network loads. The traffic pattern is random, with a packet length of 64 bytes. For 10% load the latency distribution is very narrow. Near the saturation throughput (about 60% load) a significant percentage of the packets experience a latency many times the average value of 18 μs. To reduce the probability of large network latencies the network must be lightly loaded.

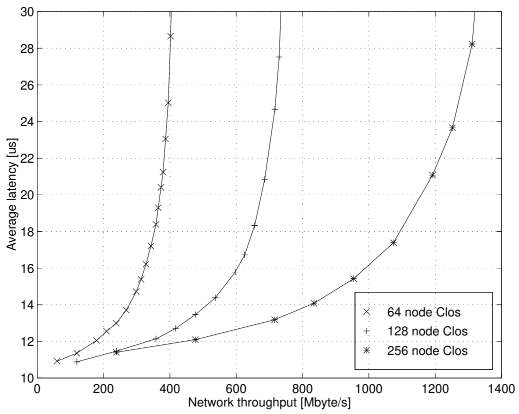

Fig. 3. Latency versus throughput for 64, 128 and 256 node Clos networks

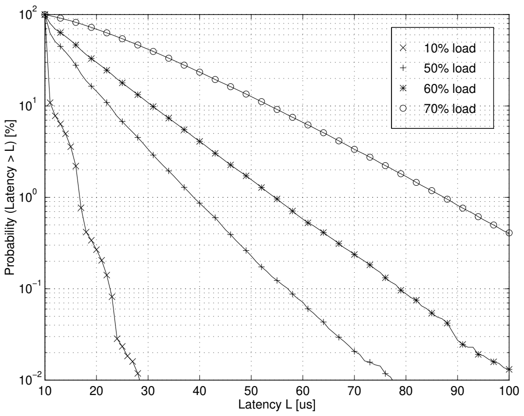

Fig. 4. Probability that a packet will have greater than a given latency value for a 64 node Clos network

2.2 Comparison of network topologies

Figure 5 shows the per node throughput for different size 2D grid and Clos networks under random traffic as a function of the packet length. The Clos networks show better performance because of the higher cross-sectional bandwidth of this topology.

The effect of packet length on throughput can also be seen, for small packets the throughput is reduced due to fixed packet overheads. Medium sized packets give the best performance because of the buffering present in the STC104. Each switch can buffer 32 bytes in both the link input and output ports. Long packets fill the entire path from source to destination, and therefore throughput is reduced by head-of-line blocking.

2.3 Scalability of Clos and grid networks

Figure 5 also shows that the throughput of Clos and 2D grid networks does not scale linearly with network size under random traffic, the per node throughput is reduced as the network size increases.

Figure 6 shows saturation network throughput for different sizes of Clos and 2D grid networks under random and systematic traffic. The packet length is 64 bytes. Systematic traffic involves fixed pairs of nodes sending to each other. For the grid, systematic traffic corresponds to communication between nodes attached to nearest neighbour switches. The performance of the Clos under systematic traffic is independent of the choice of pairs.

For random traffic, contention at the destinations and internally to the network reduces the network throughput compared to that obtained for systematic traffic, where there is no destination contention. The fall off in performance from systematic to random traffic is more pronounced for the grid than the Clos.

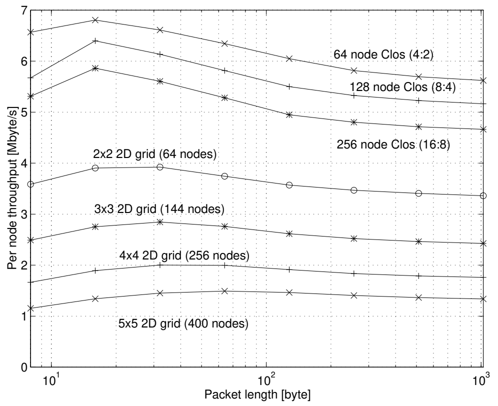

Fig. 5. Per node throughput for 2D grid and Clos networks under random traffic

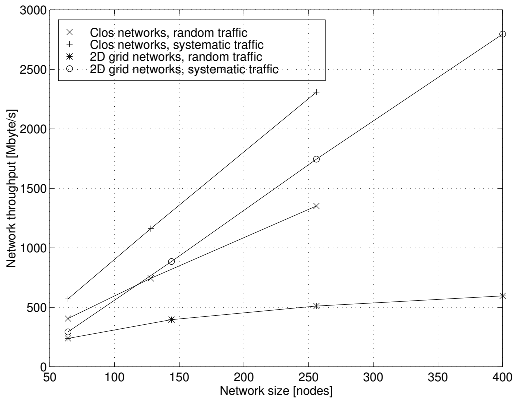

Fig. 6. Throughput versus network size for Clos and grid networks

2.4 Packet transmission overhead

The overhead in dispatching packets in the traffic nodes is determined by hardware and is approximately 650 ns. This will not in general be the case when interfacing links to a microprocessor. To demonstrate the effect of the packet overhead the dispatching delay has been artificially increased. Figure 7 shows the dependence of the total network packet rate on the individual packet overhead for a 128 node Clos under random traffic. The fall off in performance is particularly marked for short packets; the packet rate drops by nearly an order of magnitude when the overhead is increased from 10 to 100 μs. This underlines the importance of an efficient processor to link interface.

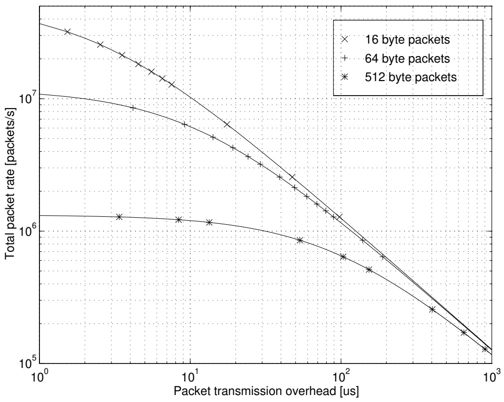

Fig. 7. Packet rate versus packet transmission overhead for a 128 node Clos network

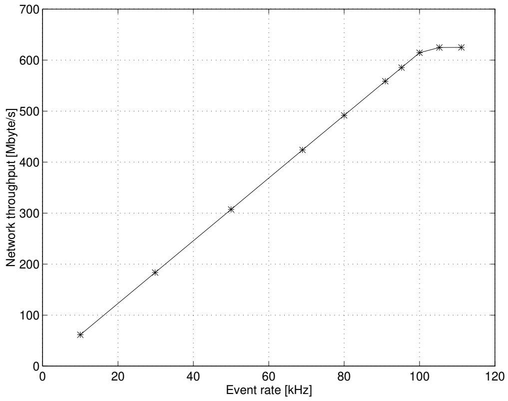

Fig. 8. Network throughput versus attempted event rate for a 256 node Clos network under HEP traffic

2.5 HEP trigger traffic

Network performance under traffic patterns as expected in the ATLAS second level trigger system have also been investigated. Figure 8 shows the total network throughput versus the event rate for a 256 node Clos network.

The traffic shows a fan-in pattern, i.e. several sources (level-2 buffers) are sending to the same destination (trigger processor). Several destinations are selected per event. The traffic parameters are based on the barrel of the Silicon Tracker (SCT) subdetector (see table below) and references [1,13].

The maximum sustainable event rate is about 100 kHz, which corresponds to the rate envisaged for the ATLAS level two trigger.

	Barrel of SCT subdetector
No. of Buffers	128
No. of Processors	128
No. of Processors selected per event	3
No. of active Buffers per event	12
Data size per event [Kbytes]	6
Packet size [bytes]	512

3 Conclusions

We have demonstrated a system running a significant number of nodes, and we expect to have a 1024 node network by April. An extensive program of systematic performance studies will be carried out. The results presented suggest that Clos networks have superior performance compared to 2D grids, other topologies (e.g. hypercube, torus) will also be investigated. We have shown that the event rate achieved with a 256 node Clos network corresponds to the rate expected within a subsystem of the ATLAS level two trigger.

Acknowledgements

We are very grateful for the support of the European Union through the Macramé project (Esprit project 8603). We acknowledge useful discussions with members of the ATLAS collaboration. Encouragement throughout was given by Professor Erwin Gabathuler (University of Liverpool), Professor John Strong (Royal Holloway) and Professor Peter Welch (University of Kent).

References

[1] The ATLAS Technical Proposal, CERN/LHCC/94-43, LHCC/P2, ISBN: 92-9083-067-0.

[2] The CMS Technical Proposal, CERN/LHCC/94-38, LHCC/P1.

[3] ANSI X3T11 Fibre Channel Standards and draft Standards, http://www.cern.ch/HSI/fcs/spec.htm#ANSI_FC_specs

[4] Marc Boisseau, Michel Demange and Jean-Marie Munier, An introduction to ATM technology, International Thomson Publishing, London, 1995. ISBN 1-850-32140-X.

[5] ISO/ANSI/IEEE Std 1596-1992, The Scalable Coherent Interface (SCI).

[6] IEEE Std. 1355, Standard for Heterogeneous Inter-Connect (HIC). Low Cost Low Latency Scalable Serial Interconnect for Parallel System Construction. IEEE Inc., 1995.

[7] 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.

[8] Networks, Routers and Transputers, edited by M.D. May, P.W. Thompson, P.H. Welch, ISBN 90 5199 129 0, http://www.hensa.ac.uk/parallel/www/nrat.html

[9] The STC104 Asynchronous Packet Switch, Data sheet, April 1995. SGS THOMSON Microelectronics.

[10] C. Clos, A Study of Non-blocking Switching Networks, Bell Systems Technical Journal 32, 1953.

[11] R.W. Dobinson, B. Martin, S. Haas, R. Heeley, M. Zhu, J.R. Hansen, Realization of a 1000-node high-speed packet switching network, ICS-NET'95, St. Petersburg, Russia.

[12] R.W. Dobinson, S. Haas, B. Martin, D.A. Thornley, M. Zhu, The Macrame 1024 Node Switching Network: Status and Future Prospects, submitted to 2nd International Data Acquisition Workshop (DAQ96), Osaka, JAPAN 13-15 November 1996.

[13] R. Bock and P. LeDu, Detector and readout specifications, and buffer-RoI relations, for the level-2 trigger demonstrator program, ATLAS DAQ note 62, 27 Jan 1997.