The Macramé 1024 Node Switching Network

S. Haas [1,2] D.A. Thornley [3] M. Zhu [1,2]
R.W. Dobinson [1,2] B. Martin [1]

[1] *CERN, 1211 Geneva 23, Switzerland*
[2] *University of Liverpool, Liverpool, UK*
[3] *University of Kent, Canterbury, UK*

e-mail: `Stefan.Haas@cern.ch`

March 27, 1997

Abstract.
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.

1 Introduction

To date, practical experience in constructing switching networks using IEEE 1355 technology [1] 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. Some theoretical studies [2, 3] have been carried out for large networks of up to one thousand nodes for different topologies.

We present results obtained on a large modular testbed using 100 Mbits/s point to point DS links. Up to 1024 nodes will be interconnected by a switching fabric based on the 32 way STC104 packet switch [4]. The system has been designed and constructed in a modular way to allow a variety of different network topologies to be investigated (Clos, grid, torus, etc.). Network throughput and latency are being studied for various traffic conditions as a function of the topology and network size.

The work presented has been carried out within the framework of the European Union's Esprit[1] program as part of the Macramé [5] project (Esprit project 8603).

[1]European Strategic Program for Research and development in Information Technology

2 The IEEE 1355 standard

The Esprit OMI/HIC[2] project has developed two bi-directional serial link technologies which form the basis of the IEEE 1355 standard :

- 100 Mbits/s Data-Strobe (DS) Link
- 1 Gbit/s High Speed (HS) Link

The work reported here is based on the DS link technology. The bit-level encoding is shown in Figure 1. The data line carries the binary data values and the strobe line only changes state when a data bit has the same value as the previous one. This scheme allows for a full bit period of skew tolerance. It also means that the links are asynchronous, as the data/strobe signal pair carries an encoded clock. Studies on the reliability of DS links [6], up to distances of 20 meters, have measured a bit error rate of better than 10^{-14}.

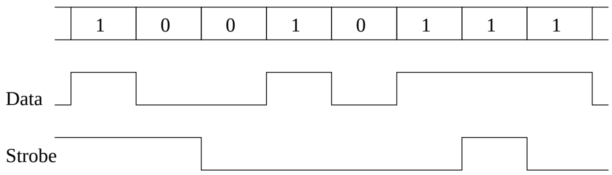

Figure 1: DS-Link bit level protocol

The IEEE 1355 protocol stack defines another three protocol layers above the bit layer: the character, exchange and packet layers. Characters are a groups of consecutive bits which represent data or control information. The exchange layer controls the exchange of characters in order to ensure the proper functioning of a link. It includes functions such as link flow control and the link startup mechanism. DS links use a credit based flow control scheme which operates on a per link basis. This ensures that the switching fabric is lossless : no characters will be lost due to internal buffer overflow.

Information in a DS link network is transferred in packets. A packet consists of a header, which contains the routing information, a payload of zero or more data bytes and an end of packet marker. The protocol allows arbitrary length packets to be sent.

3 The Macramé network testbed

3.1 Hardware

The requirement to study different topologies for a large number of nodes, imposes a system design and implementation which is highly modular and flexible while maintaining a very low per node cost. This has been achieved by building the testbed from three basic modules, which are packaged in VME mechanics :

- Traffic node modules
- Switch units
- Timing and spy nodes

[2]Open Microprocessor Systems Initiative / Heterogeneous Inter-Connect Project

A traffic node can simultaneously send and receive data at the full link speed of 100 Mbits/s. A series of packet descriptors is used to define the traffic pattern. The packet destination address, the packet length and the time to wait before dispatching the next packet is programmable. Each traffic node has memory for up to 8k such packet descriptors. The dispatch algorithm is implemented in an FPGA[3] which can be reconfigured under host control. A control processor is used to supervise the operation of a group of 4 traffic nodes.

To reduce the number of external connections, sixteen traffic nodes are connected directly to an on board STC104 packet switch. The remaining 16 ports of the switch are brought out to the front panel for inter module connections. Boards can be interconnected either directly, or via switch units which contain one STC104 with all 32 ports brought out to the front panel.

To measure latency, the timing nodes transmit and analyse time stamped packets which cross the network between chosen points. The same modules can also be inserted into any cabled connection to provide a snapshot of the traffic passing through that point. This can provide debugging information and additional data on congestion "hot spots".

A VME crate contains 128 traffic nodes and the entire 1024 node system can be housed within eight crates. All crates have an Ethernet port which drives an OS-link[4] daisy chain connection to the control processors. The STC104 packet switches are configured via a separate DS control network which is independent from the main data path. Further details on the design of the testbed have been presented in [8].

3.2 Software

A set of files are prepared off line containing: the packet descriptors, the configuration information for every traffic node and the routing tables for the packet switches. Prior to loading this data, the control networks for the traffic nodes and packet switches are used to verify that the expected devices are present and connected in the required order.

Each control processor has 4 Kbytes of on-chip memory. It is loaded at initialisation time with a kernel which handles the control link traffic and the dynamic loading of the application programs. Application programs for self-test, hardware configuration, storing of traffic descriptors and run time supervision are loaded in turn by the host which also controls their synchronisation.

Once the system is running each control processor maintains local histograms of results. These are returned to the host on request for on-line display, data logging and subsequent analysis.

3.3 Network topologies

Figure 2 shows how a two dimensional grid network topology can be constructed. Each packet switch has 16 on board connections to traffic nodes and four external connections to each of its four nearest neighbours. A 3-stage folded Clos [7] network is shown in Figure 3.

On the grid topology a deterministic routing algorithm with dimension order routing is used, i.e. packets are sent first in the north/south direction and then in the east/west direction to their respective destinations. This algorithm ensures deadlock-free routing on the grid. Furthermore, grouped adaptive routing is used on parallel links between adjacent routers, i.e. a packet will be sent on the first free link in a group.

---

[3] Field Programmable Gate Array

[4] 20 Mbits/s Over Sampled Transputer links

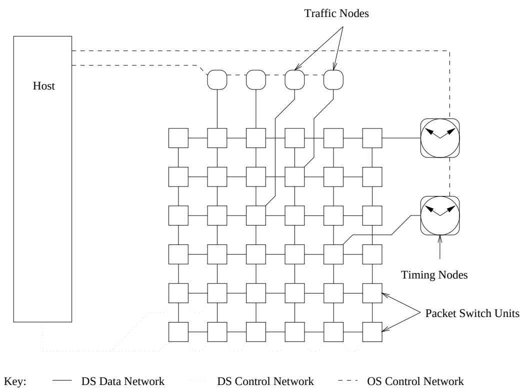

Figure 2: Architecture of the Macramé testbed

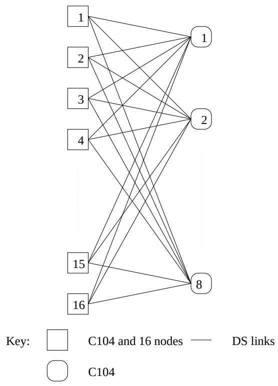

Figure 3: 256 node Clos network

4 Results

So far 656 nodes have been built and tested. They have been assembled as a range of 2-dimensional grid, and multistage Clos networks. Results are presented for these configurations. All the results presented below make use of the grouped adaptive routing feature of the STC104 packet switch, unless otherwise noted.

4.1 Comparison of network topologies

Figure 4 shows the per node saturation throughput for different size 2-dimensional grids and Clos networks as a function of the packet length. The traffic pattern is random, i.e. transmitting nodes choose a destination from a uniform distribution. The packet length is 64 bytes. The Clos shows better per node performance, this is because of the higher cross-sectional bandwidth. The effect of packet length on throughput can also be observed, 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.

4.2 Scalability of Clos and grid networks

Figure 4 also shows that the throughput of Clos and 2-dimensional grid networks does not scale linearly with network size under random traffic, the per node throughput is reduced as the network size increases. Figure 5 shows saturation network throughput for different sizes of Clos and 2-dimensional grid networks under random and systematic traffic for 64 byte packets. Systematic traffic involves fixed pairs of nodes sending to each other. For the grid this traffic pattern involves 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.

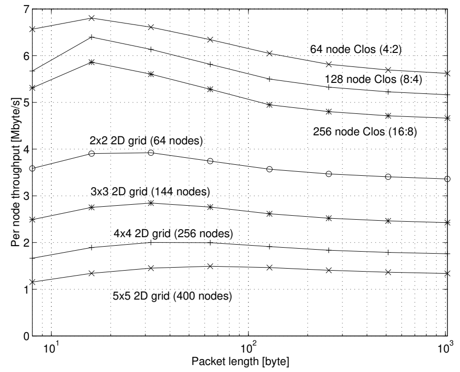

Figure 4: Per node throughput for 2-dimensional grid and Clos networks under random traffic

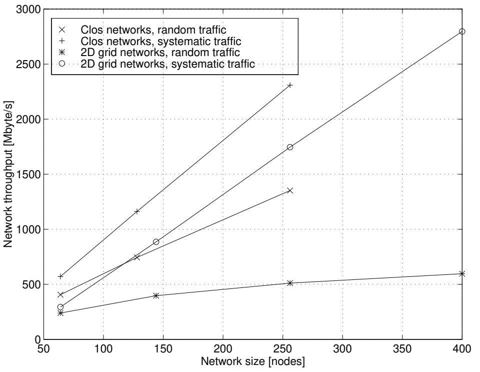

Figure 5: Throughput versus network size for Clos and grid networks

4.3 Network latency for Clos networks

Figure 6 shows the latency of three different size Clos networks under random traffic as a function of the aggregate network throughput. The packet length is 64 bytes. The results are produced by varying the network load and measuring the corresponding throughput and latency. It can be seen that the average latency increases exponentially as the network throughput approaches saturation.

Figure 7 shows the probability that a packet will have a latency greater than a given value for various network loads. Again the traffic pattern is random, with a packet length of 64 bytes. For 10% load the latency distribution is very narrow compared to higher loads. Near the saturation throughput (about 60% load) a significant percentage of the packets experience a latency many times the average value, which is 18 μs. To reduce the probability of very large latency values the network load must be far below the saturation throughput.

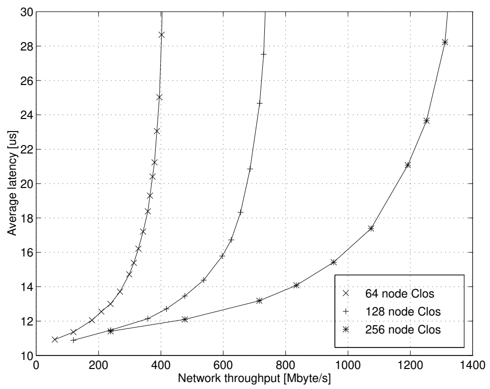

Figure 6: Latency versus throughput for 64, 128 and 256 node Clos networks

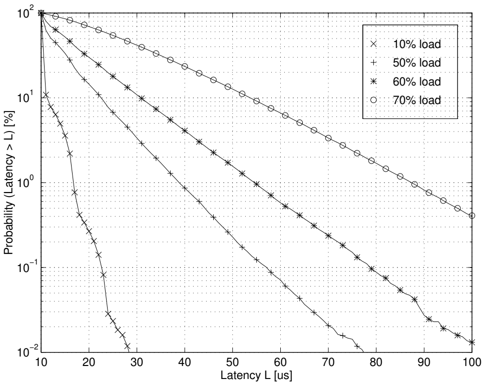

Figure 7: Probability that a packet will have greater than a given latency value for a 64 node Clos network

4.4 Packet transmission overhead

The overhead in dispatching packets in the traffic nodes is determined by hardware and is small, 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 8 shows the dependence of network throughput on packet overhead for a 128 node Clos under random traffic. The fall off in performance is particularly marked for short packets; the throughput 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.

4.5 Comparison of simulation and measurement

A 64 node Clos network has been simulated using the commercial OPNET simulation package [9]. A model of the DS link and the STC104 switch has been developed within the Macramé project for this simulator. Results from simulation and measurement have been compared and are shown in Figure 9, which shows the latency distribution for 64 byte packets and random traffic at 50% load. The majority of packets pass through the network without being queued, corresponding to the peak at 12 μs. It can be seen that the agreement between simulation and measurement is very good.

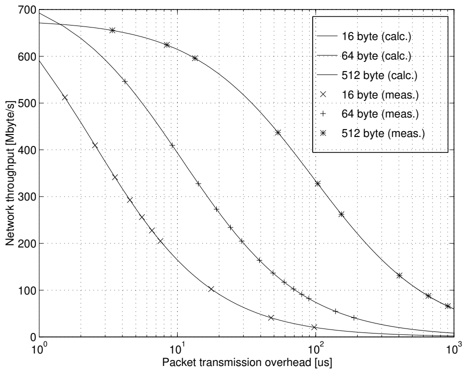

Figure 8: Network throughput versus packet transmission overhead for a 128 node Clos network

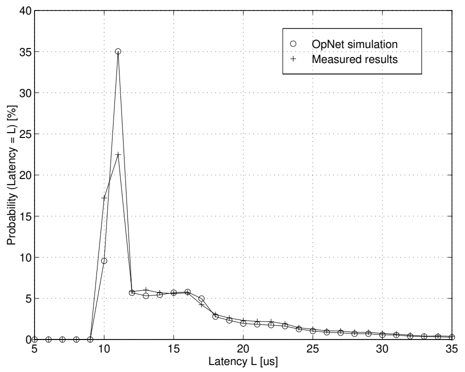

Figure 9: Simulated and measured latency distributions for a 64 node Clos network

4.6 Routing algorithms

The effect of grouped adaptive routing has been studied on the Clos network. Figure 10 shows the average network latency versus the network throughput for a 256 node 3-stage Clos network under random traffic with 32 byte packets. Deterministic routing and grouped adaptive routing are compared. Using grouped adaptive routing gives about 15% better maximum network throughput as well as slightly lower average latencies. This is because of the better utilisation of the links to the centre stage of the Clos network when using grouped adaptive routing.

The STC104 packet switch also supports universal routing, the routing algorithm on the 2-dimensional grid is as follows: as a packet enters the network it is first routed along the x-dimension to an intermediate destination switch which is chosen at random. The random header is deleted there and routing proceeds as in the deterministic case. Universal routing has the effect of distributing the traffic evenly over the entire network, thereby reducing worst case latencies.

Figure 11 shows the average latency of a 400 node 2-dimensional grid under random traffic as a function of the network throughput for universal and deterministic routing. It can be seen that the network saturates much earlier with universal routing, and that the latency increases much faster. This is because links some links have to be reserved for the random phase, so that in the x-dimension, the width of the link groups is only two and not four links. This causes greater contention and hence an increase in latency.

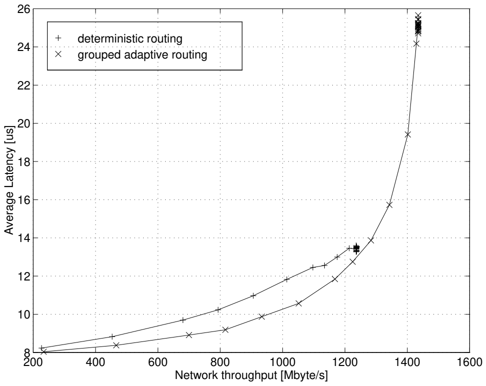

Figure 10: Deterministic and grouped adaptive routing on a 256 node Clos network

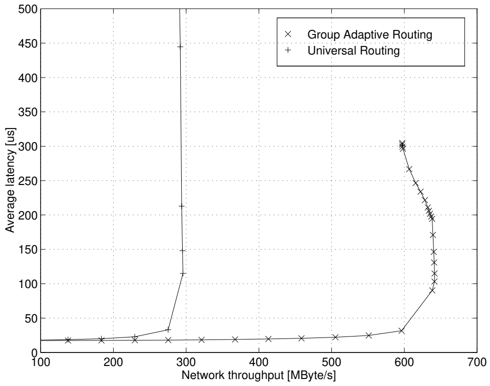

Figure 11: Universal routing on a 400 node 2-dimensional

5 Conclusions

We have demonstrated a large packet switching system, based on the DS link technology, that is performing reliably, and has provided quantitative measurements of the performance of 2-dimensional grid and Clos topologies. This work will be extended to cover other topologies and a systematic study of performance on networks of up to 1024 nodes will be carried out.

References

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

[2] 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.

[3] 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`

[4] The STC104 Asynchronous Packet Switch, Data sheet, April 1995. SGS-THOMSON Microelectronics. `http://www.st.com/stonline/books/`

[5] The Esprit Project Macramé. `http://www.pact.srf.ac.uk/macrame/welcome.html`

[6] 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/ieeehic/copper.ps.gz`

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

[8] R.W. Dobinson, B. Martin, S. Haas, R. Heeley, M. Zhu, J. Renner Hansen, Realization of a 1000-node high-speed packet switching network, ICS-NET '95 St Petersburg, Russia. `http://www.cern.ch/HSI/dshs/`

[9] The OPNET Modeler, `http://www.mil3.com/`.