The Macramé 1024 Node Switching Network

S. Haas[1,2], D.A. Thornley[1,3], M. Zhu[1,2], R.W. Dobinson[1,2,4], R. Heeley[1], N.A.H. Madsen[1,4], B. Martin[1]

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

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. 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 and switching technology, as defined in the IEEE 1355 standard. One thousand 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. Results obtained with the current 656 node setup are presented.

This 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 link protocols which form the basis of the IEEE 1355 standard, these are:

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

The work reported here is based on the the DS link protocol as shown in Figure 1. The data line carries the binary data values and the strobe line only changes state when the next data bit has the same value as the previous one. 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}.

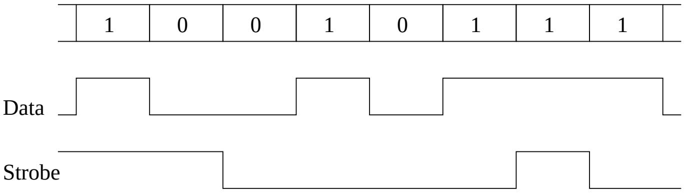

Fig. 1. DS link protocol

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. 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 are lost internally due to buffer overflow.

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

3 The Macramé network testbed

3.1 Hardware

The requirement to study different topologies, plus the need to do this for hundreds of nodes, imposes a system design and implementation which is highly

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

modular and flexible, for quick and reliable system re-configuration, as well as having a very low cost per node. The system is built up from three elements, each one housed in VME 6U mechanics:

- Traffic node modules
- STC104 packet switch units
- Timing and spy nodes

A traffic node can simultaneously send and receive data at the full link speed of 100 Mbits/s. A series of packet descriptors define the traffic pattern. The packet destination address, the packet length and the time delay to wait before dispatching the next packet is programmable. Each traffic node has memory for up to 8k such packet descriptors. The nodes are all synchronised with the same clock.

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 and all these processors are connected via a control network.

To reduce cabling, sixteen traffic nodes are hard-wired to an on board STC104 packet switch. The remaining 16 ports of the switch are brought out to the front panel for inter module connection. Boards can be interconnected either directly, or via packet switch units which contain one switch 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, in spy mode, can be inserted into any cabled connection to provide a snapshot of the traffic passing through that point. This provides 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[4] link daisy chain connection to the control processors. The STC104 packet switches have their own separate DS control network which is independent from the main data path.

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 cabled connections to each of its four nearest neighbours.

So far 656 nodes have been built and tested. They have been assembled as a range of 2D grid, and multistage Clos networks [7]. An example of a 256 node Clos is shown in Figure 3. Results are presented for these configurations. Further details on the design of the testbed are presented in [8].

[3] Field Programmable Gate Array

[4] 20 Mbits/s Over Sampled Transputer links

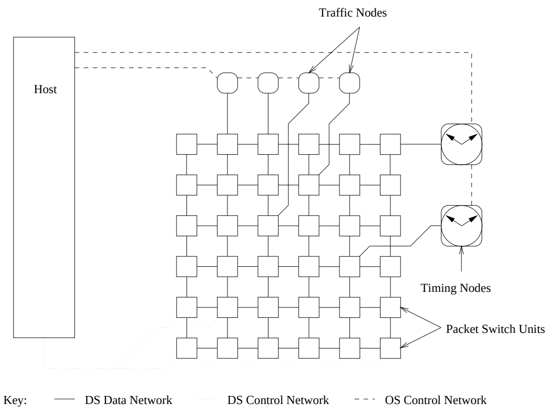

Fig. 2. Architecture of the Macramé testbed

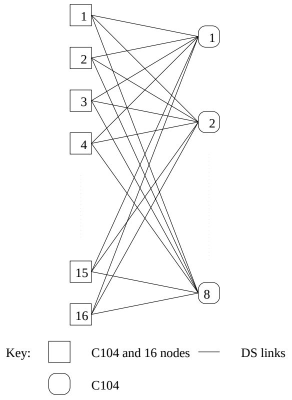

Fig. 3. 256 node Clos network

3.2 Software

A set of files is 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.

4 Results

4.1 Network latency for Clos networks

Figure 4 shows the latency of three different size Clos networks as a function of the aggregate network throughput. The traffic pattern is random, i.e. transmitting nodes choose a destination from a uniform distribution. 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. Therefore, to achieve low average latencies the network load must be below the saturation throughput.

Figure 5 shows the probability that a packet will have a latency 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 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.

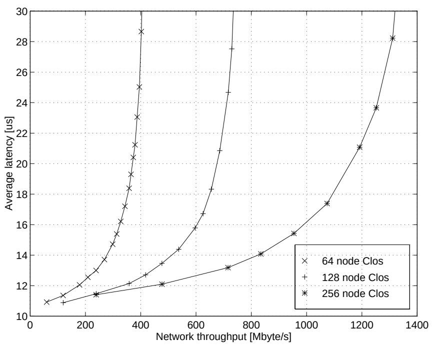

Fig. 4. Latency versus throughput for 64, 128 and 256 node Clos networks

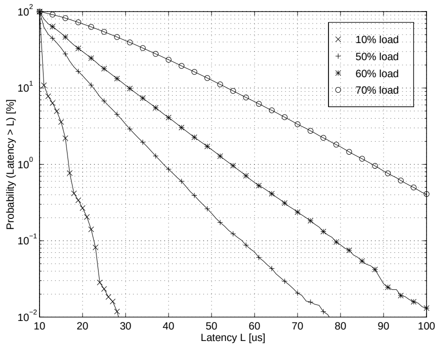

Fig. 5. Probability that a packet will have greater than a given latency value for a 64 node Clos network

4.2 Comparison of network topologies

Figure 6 shows the per node saturation throughput for different size 2D grids and Clos networks under random traffic as a function of the packet length. 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 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.

4.3 Scalability of Clos and grid networks

Figure 6 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 7 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 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. The throughput of the grid network increases logarithmically with the network size for random traffic.

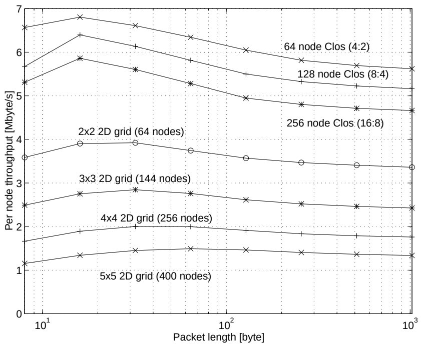

Fig. 6. Per node throughput for 2D grid and Clos networks under random traffic

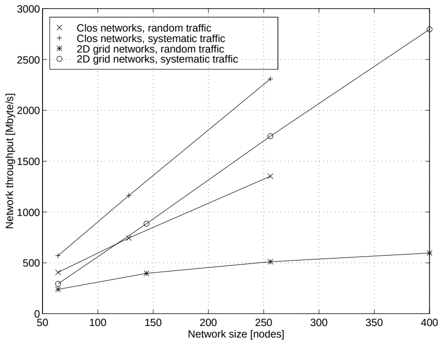

Fig. 7. Throughput versus network size for Clos and grid networks

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.

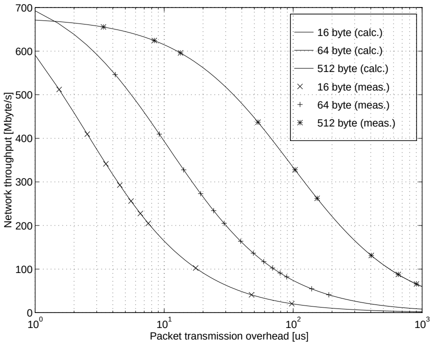

Fig. 8. Network throughput versus packet transmission overhead for a 128 node Clos network

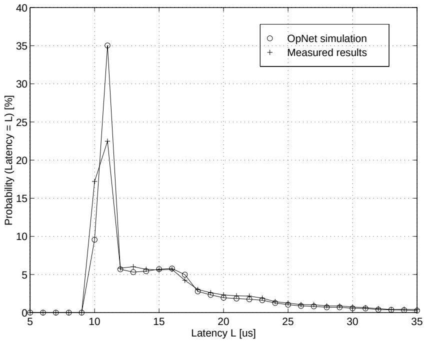

Fig. 9. A comparison of the simulated and measured latency distributions for a 64 node Clos network

5 Conclusions

We have demonstrated a large packet switching system, based on DS Link technology, that is performing reliably, and has provided quantitative measurements of the performance of 2D grid and Clos topologies. Data from this system has been used to calibrate the simulation models which now closely agree with our measurements. This work will be extended to cover other topologies and a systematic study of performance, working up to the design target of 1024 nodes.

Acknowledgements

We are very grateful for the support of the European Union through the Macramé project (Esprit project 8603). We would also like to thank PACT (Partnership in Advanced Computing Technologies, UK) for providing the simulation results presented within this paper.

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