THE MACRAMÉ 1024 NODE SWITCHING NETWORK: STATUS AND FUTURE PROSPECTS

R.W. DOBINSON, S. HAAS and M. ZHU
CERN, 1211 Geneva 23, Switzerland and
University of Liverpool, Liverpool, UK
Email: Bob.Dobinson@cern.ch

B. MARTIN
CERN, 1211 Geneva 23, Switzerland

D.A. THORNLEY,
University of Kent, Canterbury, UK, and
CERN, 1211 Geneva 23, Switzerland

22nd November 1996

ABSTRACT

Within the framework of the MACRAMÉ Esprit project a large switching network is being constructed using IEEE 1355 DS link technology. The final system will allow the systematic study of the performance of a 1024 node testbed using a variety of network topologies and traffic patterns, including those met in High Energy Physics trigger and data acquisition applications. The status of the project and some preliminary results are presented.

1. Introduction

It is recognised that the data acquisition and trigger systems for LHC experiments will depend critically on the use of high speed point to point links and packet switches [1,12]. A range of alternative technologies are presently under study including, Fibre Channel [2], ATM [3], SCI [4] and IEEE 1355 [5] (DS and HS links).

To date, practical experience 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 [6] have been carried out for large networks of up to one thousand nodes connected by different switching topologies. However, the traffic patterns used are not those relevant to High Energy Physics experiments, but rather those found in the telecommunications industry.

The work reported involves the on going construction of a large modular testbed using DS link and switch technology [7]. Up to one thousand nodes can be interconnected by a switching fabric based on the STC104 packet switch [8] which interconnects 32 bi-directional ports. The system has been designed and constructed in a modular way to allow a variety of different network topologies to be investigated (Clos, mesh, grid, tree, etc.).

Network throughput and latency are being studied for various traffic conditions as a function of the packet length, topology and network size. The effect of grouped adaptive and universal (random) routing features present in the STC104 will also be investigated. We report on the design of the system, its present status of construction and some preliminary results. The work is carried out within the framework of the European Union's Esprit program as part of the MACRAMÉ project [9] (Esprit project 8603).

2. System implementation

The requirement to study many different topologies, plus the need to do this for many hundreds of nodes, imposes a system design and implementation which is highly modular and flexible, for quick and reliable system re-configuration, as well as having a very low cost per node. In addition it must be possible to drive the links at their full speed in a sustained manner.

The system is built up from three elements, each one housed in VME 6U mechanics:

- traffic node modules
- packet switch units
- traffic timing and spy nodes

A traffic node can simultaneously send and receive data at the nominal link speed of 100 Mbits/sec. The traffic is pre-defined in terms of the packet destination address, the packet length and the time delay to wait before dispatching the packet. Each traffic node has memory for up to 8k such packet descriptors and the nodes are all synchronised with the same clock. The dispatch algorithm is operated by a state machine which can be reconfigured under host control. The overhead of sending a packet is only 650 ns. A control processor is used to supervise the operation of a group of 4 traffic nodes and all such processors are connected to 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 either directly or through the packet switch units which contain one switch with all 32 ports brought out to the front panel.

To measure latency, the traffic 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".

Each VME crate contains 128 traffic nodes and the entire 1024 node system can be housed within eight crates. Each crate has an Ethernet port which drives a daisy chain connection to the control processors. The STC104 packet switches have their own separate control network which is independent from the main data path.

Figure 1 shows how a two dimensional grid network topology can be constructed. Each packet switch has 16 connections to traffic nodes and four cabled connections to each of its four nearest neighbours.

A set of files is prepared off line containing: the packet descriptors, the state machine 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 not only present but also connected in the required order.

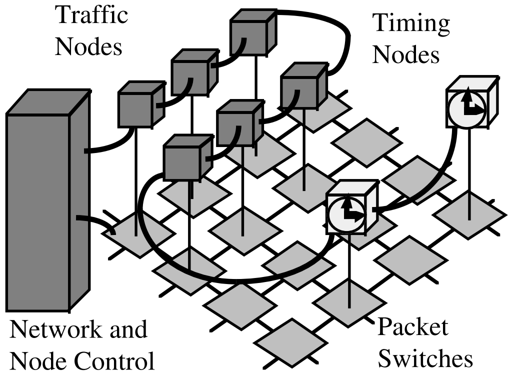

Figure 1 : Macramé Testbed Block Diagram

Each control processor has 4kbytes of on-chip memory. It is loaded at initialisation time with a "manager" 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 manages their synchronisation.

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

So far 384 nodes have been built and tested. They have been assembled as a 256 node 2D grid as shown in Figure 1, as well as a 128 node multistage network, as shown in Figure 2. The symmetry of this network means that it can be 'folded' about the central stage to give either of the two graphical representations. This network is referred to as a Clos network [13]. Results are presented for these two configurations.

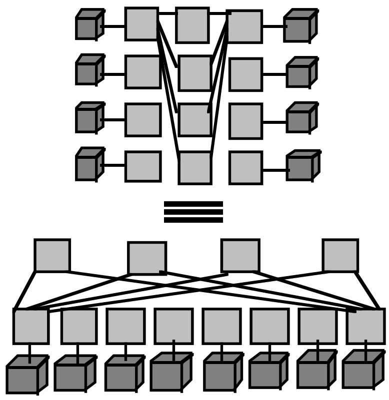
Figure 2 : 128 Node Clos network

3. Results

The results shown are preliminary and show that large networks of DS links and packet switches can be built and operated in a reliable manner. They also illustrate some of the different aspects of network behaviour which can be investigated.

3.1 Traffic patterns

Figure 3 shows the total network data throughput obtainable as a function of the packet length for a 128 node Clos network. Two types of traffic are shown, namely random traffic where transmitting nodes choose a destination at random, and systematic traffic which involves fixed pairs of nodes transmitting to each other. For random traffic, contention at the destinations reduces the network throughput to about 60% of that obtained for systematic traffic, where there is no destination contention.

Figure 4 shows random and systematic traffic through a 256 node two-dimensional grid. Two different systematic traffic patterns are shown. The best case corresponds to communication between nodes attached to nearest neighbour switches and the worst case to communication between “mirror image” nodes. Because of contention for the links internal to the network, the 2D grid is more sensitive to the traffic pattern than the Clos. This shows that good performance requires locality.

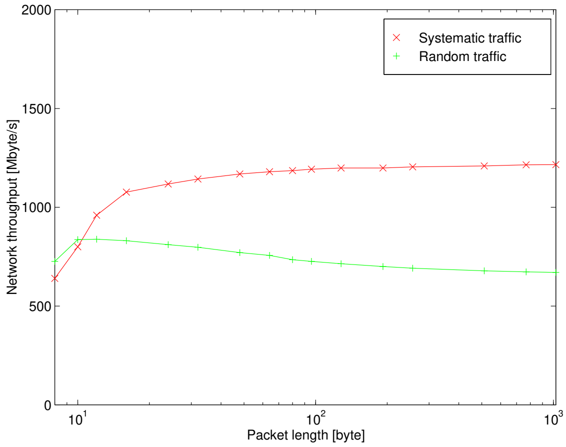

Figure 3 : 128 node Clos network under random and systematic load

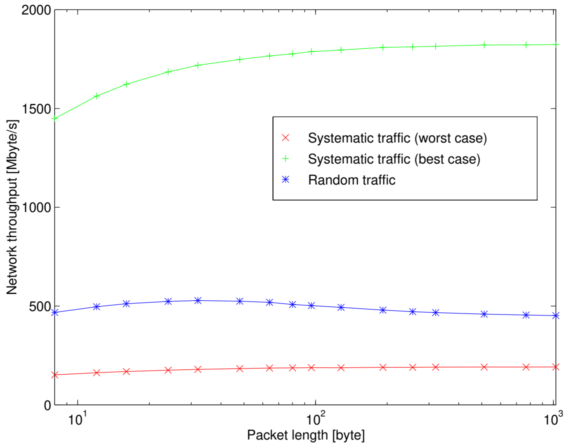

Figure 4 : 256 node 2D grid under random and systematic load

3.2 Packet overhead

The overhead in dispatching packets in the traffic nodes is determined only by hardware and is very small, less than one microsecond. 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. Figures 5 and 6 show the dependence of data throughput and packet rate on packet overhead for a 128 node Clos network 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.

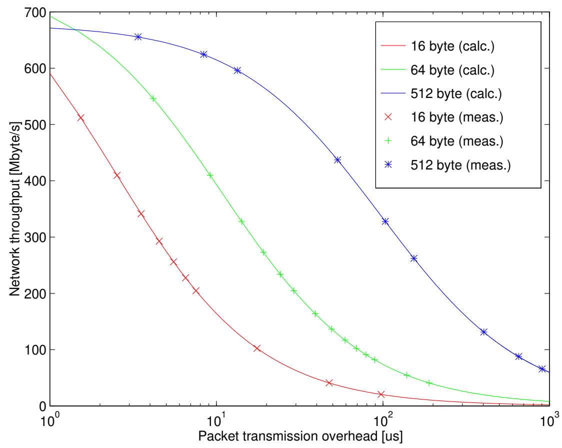

Figure 5 : Network throughput for a 128 node Clos network

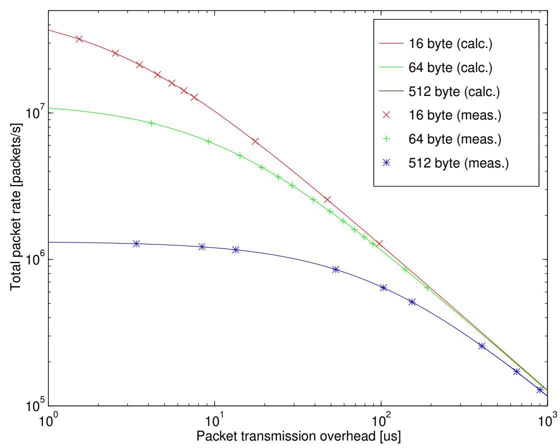

Figure 6 : Packet rate for a 128 node Clos network

This underlines the importance of an efficient processor to link interface for applications requiring high packet rates.

3.3 Topology

Random traffic has been applied to a Clos network with 128 nodes and to a 2D grid, of 3 by 3 switches, with 144 nodes. The results in figure 7 show that a better per node performance is obtained with the Clos compared to the 2D grid. This is because of the larger cross-sectional bandwidth of the Clos network.

3.4 Scalability

Figures 7 shows the per node data rate for two Clos networks and for three different size 2D grids under random traffic. Although the overall data throughput increases with the network size, the per node performance decreases. This effect is less marked for the Clos networks. These results show that the scaling behaviour of Clos networks nearer to the linear case than that of 2D grids.

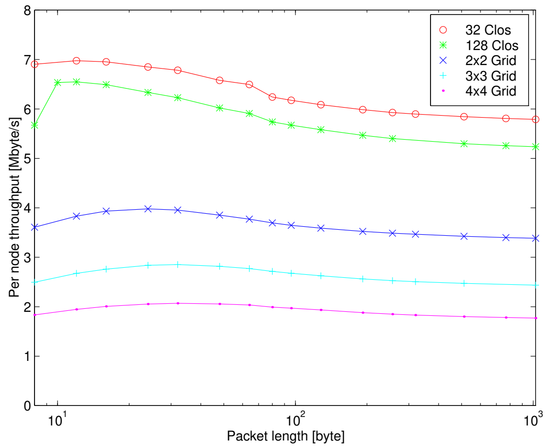

Figure 7 : Per node throughput for Clos and 2D grid networks under random traffic

3.5 Latency

The packet latency shown in figure 8 is for 15 nodes connected through a single switch together with two timing nodes. The traffic pattern is random. Trace packets are injected by the timing nodes and their delay through the switch is measured. The network load must be reduced well below the saturation throughput, about 6Mbyte/s, to significantly reduce the latency. The steps in the graph show the number of data packets in the queue in front of the trace packet.

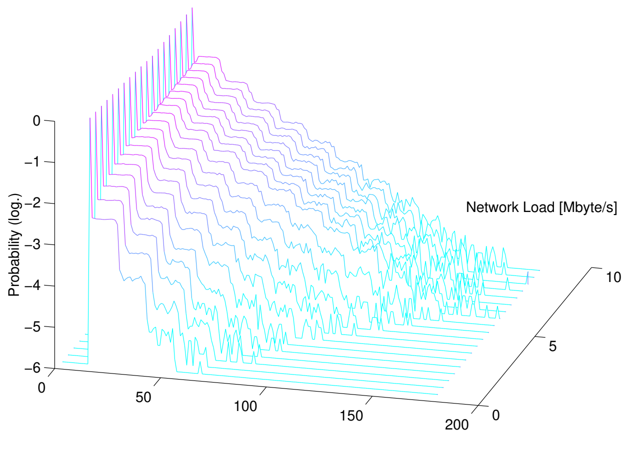

Figure 8 : Latency distribution versus network load for a single STC104

3.6 High Energy Physics trigger traffic

Network performance under traffic patterns as expected in the Atlas and HERA-B second level trigger systems have also been investigated. Figures 9 and 10 show the total network throughput versus the event rate for a 128 node Clos network.

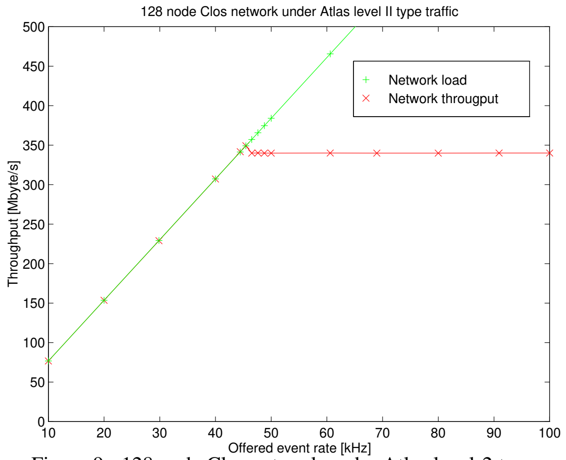

Figure 9 : 128 node Clos network under Atlas level-2 type traffic

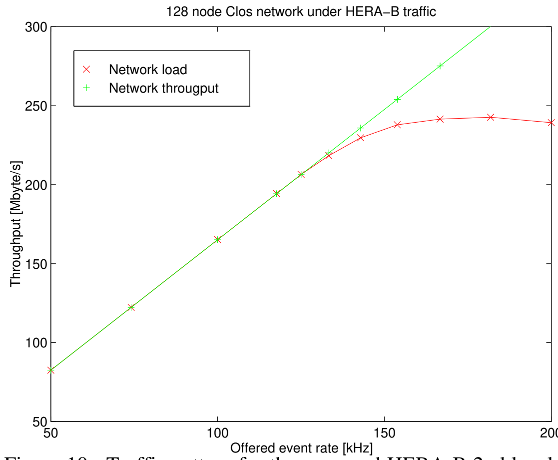

Figure 10 : Traffic pattern for the proposed HERA-B 2nd level trigger on a 128 node Clos network

For both applications the traffic shows a fan-in pattern, i.e. several sources (level-2 buffers) are sending to the same destination (trigger processor). In the case of HERA-B there is only one destination per event, for the Atlas level-2 trigger, several destinations are selected per event. The values of the traffic parameters for Atlas are based on references [1] and [10], more detail on the proposed HERA-B second level trigger is given in [11]. The traffic parameters used for these two applications are as follows :

	Atlas Level 2 Trigger	HERA-B 2nd Level Trigger
No. of buffers	64	64
No. of processors	64	64
No. of processors selected per event	3	1
No. of active buffers per event	15	16 to 18
Data size per event [Kbytes]	7.68	1.65
Packet size [byte]	512	90 to 160

For the Atlas trigger traffic the maximum sustainable event rate is about 45kHz. Because of the near linear scaling of the Clos topology it is expected that larger networks will be able to achieve the required event rate of 100kHz. For the HERA–B application the achieved event frequency is well above the required performance of 50 kHz.

4. Conclusions and future work

The construction of a very large, modular, testbed for evaluating the performance of IEEE1355 DS links and switches is well advanced. About half of the 1024 node system is functioning reliably and preliminary results have been obtained. We expect to complete construction by the end of 1996 and an extensive program of measurements and analysis will then be carried out, including the study of possible architectures for the Atlas 2nd and 3rd level triggers.

5. 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 and the Niels Bohr Institute (Copenhagen) for providing the input for the HERA-B studies. Encouragement throughout was given by Professor Erwin Gabathuler, University of Liverpool and Professor Peter Welch, University of Kent.

6. References

[1] The ATLAS Technical Proposal, CERN/LHCC/94-43, ISBN:92-9083067-0.

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

[3] Marc Boisseau, Michel Demange and Jean-Marie Munier, An introduction to ATM technology, International Thomson Publishing,
London, 1995. ISBN 1-850-32140-X.
`http://www.atmforum.com`

[4] ISO/ANSI/IEEE Std 1596-1992, The Scalable Coherent Interface (SCI)
`http://sunrise.scu.edu/HowToGetSCIdox.html`

[5] IEEE Std. 1355-1995, Standard for Heterogeneous InterConnect (HIC)
`http://stdsbbs.ieee.org/groups/1355/index.html#Index`
`http://www.cern.ch/HSI/dshs`

[6] A. Klein, Interconnection Networks for Universal Message-Passing Systems, Esprit '91 Conference Proceedings, pp. 336-351, Commission of the European Communities, 1991, ISBN 92-826-2905-8.

[7] B. Martin et al., Realisation of a 1000 node high speed packet switching network, Procedings of the International Symposium on Problems of Modular Information Computer Systems, St Petersberg, Russia, June 1995

[8] STC104. Asynchronous Packet Switch, Preliminary Data Sheet, June 1994, SGS-THOMSON Microelectronics.

[9] `http://www.pact.srf.ac.uk/macrame/index.html`

[10] R. Bock and P. LeDu, Detector and Readout specification for Level-2 Trigger Demonstrator Program, ATLAS DAQ note 53, June 1996.
`http://atlasinfo.cern.ch/Atlas/GROUPS/DAQTRIG/NOTES/notes_ps.html`

[11] Proposal for a DS-Link based SLT for HERA-B, the NBI group, November 1996

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

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