Realisation and Performance of IEEE 1355 DS and HS Link based, High Speed, Low Latency Packet Switching Networks

M. Zhu[1,2], D.A. Thornley[1,3], J. Pech[1,4], B. Martin[1], N.H. Madsen[1,5], R. Heeley[1], S. Haas[1,2], R.W. Dobinson[1,2,5], C.R. Anderson[2]

[1] CERN, 1211 Geneva 23, Switzerland

[2] University of Liverpool, Liverpool, UK

[3] University of Kent, Canterbury, UK

[4] University of Prague, Prague, Czech Republic

[5] RHBNC, University of London, London, UK

Abstract

We report on the construction of a 1024 node switching network using IEEE 1355 DS link technology. The nodes are 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 topologies under various traffic conditions including those expected within the second level trigger of the ATLAS experiment. Initial experience using 1 Gbaud IEEE 1355 HS links and switches is also presented.

I. Introduction

We present results obtained on a large modular testbed using 100 MBaud point-to-point DS links. Up to 1024 nodes have been interconnected by a switching fabric based on the 32 way STC104 packet switch [1]. The system has been designed and constructed in a modular way to allow a variety of different network topologies to be investigated. Network throughput and latency are being studied for various traffic conditions as a function of the topology and network size. The traffic conditions include those expected within the second level trigger of the ATLAS experiment [2]. Initial work on a testbed for 1 GBaud IEEE 1355 HS links and switches is also presented.

The work presented here has been carried out within the framework of the European Union's ESPRIT[1] program as part of the OMI[2] Macramé[3] and ARCHES[4] projects.

II. The IEEE 1355 Standard

Two complementary high-speed serial link technologies have been developed within the framework of the OMI/HIC[5] Esprit project. They have been subsequently standardised and form the basis of the IEEE 1355 [3] standard :

- 100 MBaud Data-Strobe (DS) link
- 1 GBaud High Speed (HS) link

The standard allows modular scalable interconnects to be constructed based on high-speed point-to-point links and switch chips. Using the lightweight protocols of IEEE 1355 these networks can provide a transparent transport layer for a range of higher level protocols.

The IEEE 1355 protocol stack defines four protocol layers: bit, character, exchange and packet layers. Characters are 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. A credit based flow control scheme is used which operates on a per link basis. This scheme ensures that no characters will be lost due to buffer overflow.

Information in IEEE 1355 networks 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. The destination address in the header can be zero (for a directly connected link) or more bytes.

III. The Macramé Network Testbed

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 low cost per node. This has been achieved by building the testbed from three basic modules, which are packaged in VME mechanics :

Traffic Modules A traffic node can simultaneously send and receive data at the full link speed of 100 MBaud. 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. To reduce the number of external connections, sixteen traffic nodes are connected directly to an on-board STC104 packet switch

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

[2] Open Microprocessor Initiative

[3] Multiprocessor Architectures: Connectivity, Routers And Modelling Environment (ESPRIT project 8603)

[4] Application, Refinement and Consolidation of HIC, Exploiting Standards (ESPRIT project 20693)

[5] Heterogeneous InterConnect (ESPRIT project 7252)

to form a traffic module. The remaining 16 ports of the switch are brought out to the front panel for inter module connections.

Switch Units In order to build indirect networks, i.e. topologies where not all the switches have terminal nodes attached directly to them, a switch unit is required. It consists of one STC104 packet switch with all 32 ports brought out to the front panel through differential buffers.

Timing Modules To measure latency, the timing modules transmit and analyse time stamped packets which cross the network between chosen points.

Further details on the design of the testbed have been presented in [4].

IV. Network topologies

Grid, torus and Clos [5] network topologies have been studied. Figure 1 shows how a 400 node 2-dimensional grid network 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.

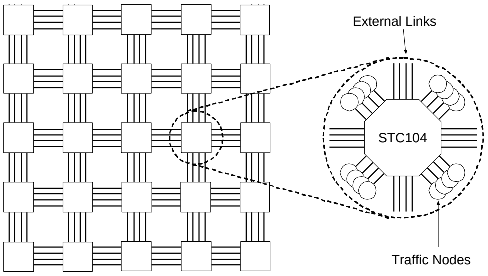

Fig. 1 : A 400 node 2-dimensional grid network

A 256 node 3-stage folded Clos network is shown in figure 2. The centre stage of the Clos consists of the switch modules described above. Each terminal stage switch connects with groups of two links to every switch in the centre stage.

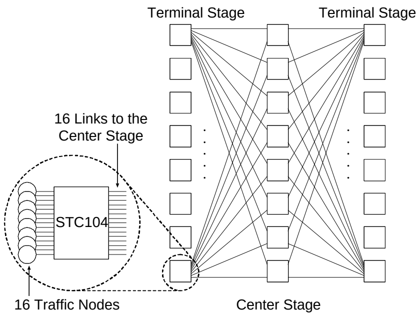

Fig. 2 : A 256 node Clos network

Grid style networks are easier to physically implement, since connections are all between adjacent switches, while Clos topologies have the constraint that each switch in the centre stage must be connected to every switch in the terminal stage.

The third topology which has been studied is the torus. A torus is similar to a grid which has its edge links wrapped around to connect to the opposite side of the network.

V. Network Routing

The STC104 contains a 32×32 non-blocking crossbar switch, which uses wormhole routing [6]. The routing decision is made as soon as the packet header arrives, the header is then sent to the chosen output link and the rest of the packet follows without being internally buffered. This implies that packets can be passing through several switches at the same time. The packet header creates a temporary circuit ('worm hole') through which the data flows. As the end of the packet passes through each device the circuit closes. Worm-hole routing minimises latency and buffering requirements compared to switches using store and forward techniques. It also has the advantage that it allows arbitrary length packets. The packet latency across the STC104 has been measured to be $1\mu s$.

Under random traffic the performance of a network is limited by head-of-line blocking. When several packets are contending for the same output link and a packet is stalled because the required output link is busy, all packets in the input queue behind it are also blocked, even if their selected output link is free. Each STC104 switch can buffer 43 bytes in the input port, and 23 bytes in the output port. This effect limits the theoretical performance of a cross-bar switch under random traffic to about 60% of the maximum cross-sectional bandwith [7].

The STC104 supports a locally adaptive routing scheme which allows packets to be sent down any free output link in a programmed set of consecutive links. This improves performance by ensuring that there are no packets waiting to use one link when an equivalent link is idle. A set of links used to access a common destination can therefore be logically grouped together, increasing the aggregate throughput to the destination. This grouped adaptive routing allows efficient load-balancing in multi-stage networks [8] and also enables a degree of automatic fault-tolerance [9]. On the grid topology grouped adaptive routing is used on parallel links between adjacent routers. For Clos networks, all the links from the terminal stage switches to the centre stage can be grouped, because all the centre stage switches are equivalent. Parallel links from the centre stage to the terminal stage are also grouped.

VI. Results

The full scale system with 1024 nodes has been built and tested. A range of 2-dimensional grid, torus and multistage Clos networks have been assembled, results are presented for these configurations. Measurements to study other topologies, such as hypercube networks, are on-going.

A. Comparison of network topologies

Figure 3 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 random traffic, nodes choose a destination from a uniform distribution.

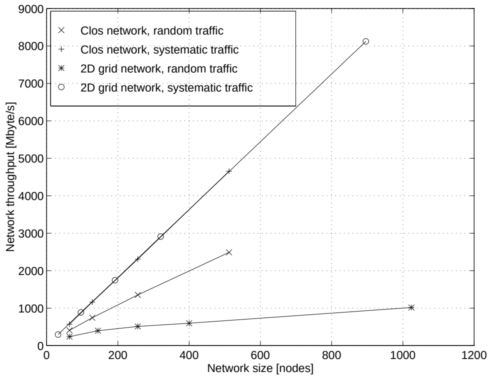

Fig. 3 : Throughput versus network size for Clos and grid networks under random and systematic traffic with 64 byte packets

The throughput of the Clos networks for random traffic is higher than for the 2-dimensional grids. This is because of the larger cross-sectional bandwidth. The maximum cross-sectional bandwidth is defined as the bidirectional data rate that can pass between two parts of the network if it is divided into two equal halves. The 256 node Clos has a maximum theoretical cross-sectional bandwidth of 2.44 Gbytes/s, whereas for the grid of the same size it is only 305 Mbytes/s. For the grid networks, the per-node throughput decreases rapidly as the network size increases, e.g. for a 64 node grid, which consists of an array of 2×2 switches, the per node throughput under random traffic is only 40% (4Mbytes/s) of the maximum link bandwidth. For a 1024 node grid (8×8 switches), the per node throughput under random traffic is only 10% (1 Mbyte/s) of the maximum link bandwidth.

The results show that the network throughput under random traffic is always significantly lower than the maximum theoretical cross-sectional bandwidth. This is because the throughput of the network under random traffic is limited by head-of-line blocking.

The fall off in performance from systematic to random traffic is more pronounced for the grid than the Clos. The degradation of performance as the network size increases agrees with analytical models presented in [8]. This study predicts the throughput of Clos networks under sustained random load to degrade by approximately 25% from linear when the network size is increased from 64 to 512 nodes. The measurement results shown in figure 3 show a reduction of about 20% under the same conditions.

The performance of the grid is strongly dependent upon the choice of pairs for systematic traffic. The results shown for the grid in figure 3 use a 'best case' scenario, this traffic pattern involves communication between nodes attached to nearest neighbour switches. A 'worst case' scenario would be the pairing of nodes with their mirror image node in the network. The throughput of a 256 node grid under this 'worst case' pattern is only 200 Mbytes/s, as opposed to 1.8 Gbytes/s under the 'best case' pattern. This shows that on the grid good performance requires locality. The throughput of the Clos under systematic traffic is independent of the choice of pairs due to its high cross-sectional bandwidth.

The throughput of the torus is about 20% higher than the grid due to the extra wrap around links which are available.

B. Network latency for Clos networks

Network latency is defined as the delay from the transmission of the packet header at the source to the reception of the end-of-packet at the destination. Figure 4 shows the latency of four 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 values. It can be seen that the average latency increases rapidly as the network throughput approaches saturation.

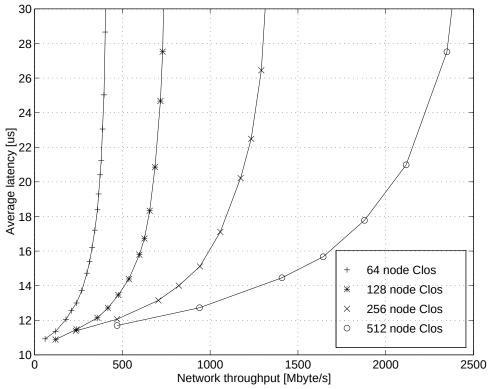

Fig. 4 : Latency versus throughput for Clos networks under random traffic with 64 byte packets

Some applications, e.g. multimedia traffic, may require statistical bounds on the maximum latency values occurring. This information can be obtained from figure 5 which shows the probability that a packet will be delayed by more than a given latency value for various network loads. The results have been obtained on a 256 node Clos network. Again the traffic pattern is random, with a packet length of 64 bytes. From figure 4 it can be seen that this network saturates at 1.35 Gbytes/s which corresponds to a throughput of 57%.

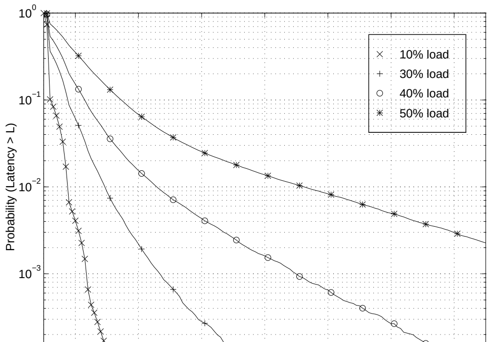

Fig. 5 : Cumulative latency distribution for a 256 node Clos network under random traffic with 64 byte packets

For 10% load the cumulative latency distribution is narrow and only a small percentage of the packets (0.01%) are delayed by more than 3 times the average latency value of about 11μs. As the network load increases, the tail of the latency distribution get wider and near the saturation throughput a significant fraction of the packets experience a latency many times the average value, e.g. at 50% load 0.7% of the packets are delayed by more than five times the average latency of 21μs. To reduce the probability of large latency values the network load must therefore be kept well below the saturation throughput.

C. Effect of grouped adaptive routing

All the measurements presented so far have been made using grouped adaptive routing. In order to quantify the impact of this feature of the STC104 packet switch, deterministic routing and grouped adaptive routing have been compared on the Clos topology. With deterministic routing, routing channels are evenly spread across the centre stage links. Figure 6 shows the average network latency versus network throughput for a 256 node 3-stage Clos network under random traffic with 64 byte packets. The network load was increased until saturation occurred. Using grouped adaptive routing results in a nearly 20% higher saturation network throughput as well as lower average latencies. This is because the adaptive routing technique minimises the effects of load imbalance, thereby allowing a better utilisation of the links to the centre stage of the Clos network.

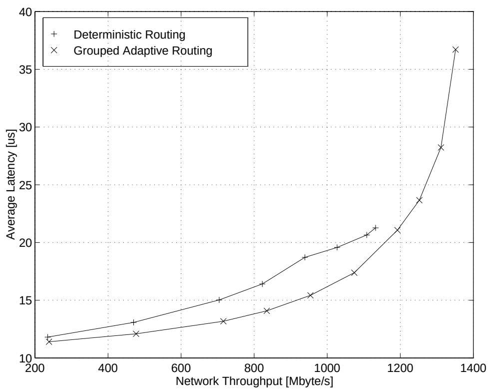

Fig. 6 : Deterministic and grouped adaptive routing on a 256 node Clos network under random traffic

D. Reliability

We have tested differential DS link connections over twisted-pair cable for immunity against electromagnetic interference according to the IEC 801-4 standard [10]. The setup passed test severity level 2, which corresponds to the environment likely to be found in the control room of an industrial or electrical plant. We have also attempted to quantify the reliability of DS link systems using the 1024 node 8×8 grid network by performing long-term error tests. The 1024 node grid contains a total of 1344 active DS links, about one third of these link use differential buffers and 2 meter twisted pair cables. The others are single-ended on-board connections. We have run the system continuously for over 200 hours without observing any link errors. This translates to a per-link error rate of better than 9.6×10^{-18}.

E. ATLAS second level trigger traffic

Network performance has been measured under the traffic patterns expected within the second level trigger of the ATLAS experiment [11]. The traffic shows a fan-in pattern, i.e. several sources (level two buffers) are sending to the same destination (Feature Extractor processors or FEX). Several FEX processors are active per event. The results presented use parameters based on the Silicon Tracker (SCT) subdetector.

The requirements of the SCT have been taken from ATLAS internal documents [12, 13]. The method used to generate event description files is described in an accompanying paper [14]. The total number of buffers in the SCT has been estimated to be 256. The number of FEX processors has been chosen to be 80.

Two different distributions of FEX processors across a 512 node Clos network have been investigated: grouped and distributed. In the grouped case all 80 FEX processors are connected to the last 5 terminal stage switches, i.e. 16 FEX processors per switch. In the distributed case the FEX processors are connected to the last 16 terminal stage switches, i.e. 5 FEX processors per switch. In all measurements the 256 buffers are connected to the first 16 terminal stage switches.

Figure 7 shows the total network throughput versus attempted event rate for a 512 node Clos. The maximum sustainable event rate is about 120 kHz which is greater than the 100 kHz expected rate within the second level trigger. The improvement from grouped to distributed FEX processors is due to reducing the contention at each terminal stage switch and therefore reducing the effect of head-of-line blocking. The average network latency is shown in figure 8.

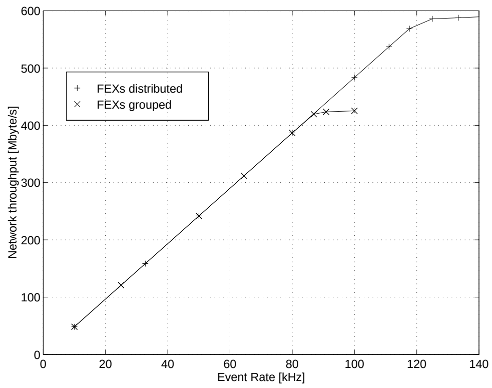

Fig. 7 : Network throughput versus attempted event rate for a 512 node Clos

The average receive rates for the individual FEX processors were 5.3 and 7.2 Mbytes/s for the grouped and distributed cases respectively. These amounts correspond to 53 and 72% of the theoretical maximum for an individual link (9.97 Mbytes/s [7]). Under random traffic the individual link data rate for a 512 node Clos is 4.2 Mbytes/s.

VII. HS LINKS AND SWITCHES

As part of the ARCHES project a 128 node 1 GBaud HS link testbed is being constructed. Initial work has been carried out to investigate the behaviour of the silicon components to be

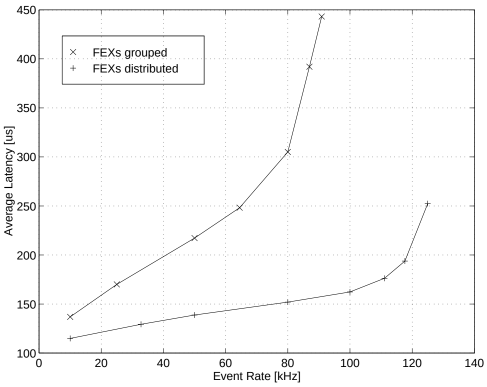

Fig. 8 : Average network latency versus attempted event rate

used. An evaluation board containing a single 8-port Rcube [15] switch and two Bullit [16] link adapters has been constructed and is shown in figure 9. The Rcube switch currently runs error free with link speeds of up to 768 MBaud in full-duplex mode, corresponding to a data rate of 62 Mbytes/s per direction for 1024 byte messages. Higher speed tests will soon be started. The packet latency across the Rcube has been measured at 180 ns.

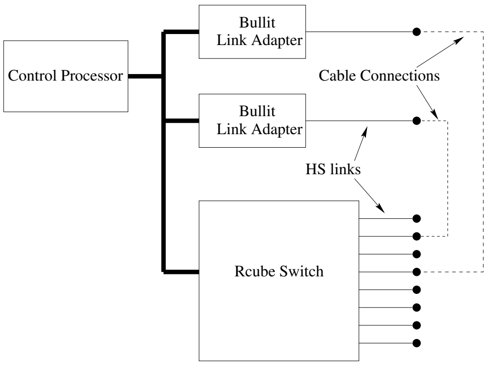

Fig. 9 : The Rcube and Bullit evaluation board

VIII. CONCLUSIONS

We have demonstrated a large packet switching system, based on the DS link technology. The network is performing reliably, and has provided quantitative measurements of the performance of 2-dimensional grid, torus and Clos topologies. The results show that, although grid and torus networks are easier to physically implement, the Clos networks clearly give better performance. Given the type of traffic, the required throughput and the maximum latency, it is possible to use the results presented to evaluate the suitability of a given topology to meet those requirements. The network designer needs to consider not only the average latency, but also the effect of the long latency tail occurring in packet switching networks due to contention. The measurements presented give an upper limit of the network performance obtainable but the performance will be reduced further if the network interfaces are unable to handle the low packet overheads required.

In practise, the system is extremely stable and measuring the upper limit of the error rate was governed principally by unstable Ethernet interfaces and mains power failures. The work presented here will be extended to cover other topologies and the application of IEEE 1355 as a switching technology for other interconnects (for example Ethernet) will be investigated.

We have shown that the event rate achieved with a 512 node Clos applied to the SCT subdetector of the ATLAS level two trigger is about 120 kHz. Work on other subdetectors and trigger architectures is on-going.

Initial experience with GBaud HS links and switches is encouraging.

ACKNOWLEDGEMENTS

We are grateful for the support of the European Union through the Macramé (ESPRIT project 8603) and ARCHES (ESPRIT project 20693) projects.

IX. REFERENCES

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

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

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

[4] 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.[6]

[5] C. Clos, “A Study of Non-blocking Switching Networks”, *Bell Systems Technical Journal*, vol. 32, 1953.

[6] W.J. Dally and C.L. Seitz, “Deadlock-free message routing in multiprocessor interconnection networks”, *IEEE Transactions on Computers*, vol. 36, no. 5, pp. 547–553, 1987.

[7] *Networks, Routers and Transputers*, edited by M.D. May, P.W. Thompson, P.H. Welch, ISBN 90 5199 129 0.

[8] 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.

[9] P.W. Thompson, “Globally Connected Fault-Tolerant Systems” in J. Kerridge (ed.), *Transputer and occam Research: New Directions,* IOS Press, 1993.

[10] International Standard IEC 801-4, *Electromagnetic compatibility for industrial-process measurement and control equipment, part 4, Electrical fast transient/burst requirements*, CEI Geneva, 1988.

[11] J. R. Hansen et al., *Local-Global demonstrator program for the ATLAS second level trigger*, Real Time 97, Beaune, France, 1997.

[12] R. Bock and P. LeDu, *Detector and readout specifications, and buffer RoI relations, for the level two trigger*, internal ATLAS DAQ note 62, Jan 1997.

[13] S. George, J.R. Hubbard and J.C. Vermuelen, *Input parameters for modelling the ATLAS second level trigger*, internal ATLAS DAQ note 70, June 1997.

[14] J. C. Vermuelen, *Discrete event simulation of the ATLAS second level trigger*, Real Time 97, Beaune, France, 1997.

[15] *The Rcube Specification*, Version 1.7, Laboratoire MASI, Université de Pierre et Marie Curie, Paris, France, 1997.

[16] *The Bullit Data Sheet*, Version 2.0, Bull Serial Link Technology, 1995.

[6] `http://www.cern.ch/HSI/dshs/`