Topologies and routing in Gigabit switching fabrics

Sven-Arne Reinemo*, Olav Lysne and Tor Skeie
Department of Informatics, University of Oslo
Box 1080 Blindern, N-0316 Oslo, Norway
Phone: +47 22 00 85 56 +47 22 85 24 28 +47 22 85 24 07
Fax: +47 22 85 24 01
Robert Dobinson, Stefan Haas and Brian Martin
CERN
Phone: +41 22 76 73066 +41 22 76 78646 +41 22 76 74915

Abstract— **The ever increasing demand for bandwidth in computer networking has seen an evolution from the 3 Mbps Ethernet of 1976 to 1 Gbps Ethernet today with 10 Gbps available soon. This evolution in bandwidth has presented the network vendors with an increasing range of challenges when producing high capacity switches with a large number of ports.**

In this paper we study three different scalable topologies suitable for Gigabit switching fabrics. We consider the Hybrid, the Dual Hybrid and the Hierarchical Clos together with several routing strategies.

We demonstrate that the Hybrid is the topology that has the best price/performance ratio while the Dual Hybrid has the best performance when cost is not considered. We also show that universal routing is necessary to achieve fairness in the network and this combined with grouped adaptive routing gives us high throughput without sacrificing fairness. Finally, our results show that excess capacity in a network is critical to performance and drastically reduces the differences between the routing schemes.

Keywords— **Gigabit switching fabrics, scalability, crossbar, routing.**

I. Introduction

IN recent years the demand for high speed switching has risen in conjunction with the increase in bandwidth available in all types of computer networks. Local area networks (LANs) in many corporations have made the transition from 10 Mbps Ethernet to 100 Mbps Ethernet and on the Internet many home users have switched from 56 Kbps to high speed alternatives such as ADSL[1] and Cable modems. This has resulted in a need for high-bandwidth network switches with a large number of ports.

The basic function of a switch is to forward packets from any input port to any given output port. A generic switch architecture therefore consists of a number of port interfaces linked through some form of interconnect. This is illustrated in Figure 1. The design and performance of the internal switch interconnect is the topic of this paper. There are several ways of implementing the internal interconnect, such as a bus, a shared memory or a crossbar switch. We will focus our study on the crossbar based interconnect, since the other two architectures have limited scalability in terms of bandwidth and number of ports [1].

Crossbar based switching fabrics have been successfully used in multiprocessor systems for a long time, but only recently have they been applied to LAN switches. Important features of the switching fabrics studied in our analysis are the aggregate throughput and fairness as well as scalability. The requirement for the switching fabrics under study was to scale from 8 to 64 external switch ports while sustaining wire speed traffic at 1 Gbps.

We use traffic data collected from an extensive program of measurements performed at CERN

*Primary contact author

[1] Asymmetric Digital Subscriber Line

to compare three different network topologies suitable for large, scalable Gigabit switching fabrics. These measurements where performed on the MACRAME test-bed , which is a hardware test-bed based on the IEEE 1355 standard [2] and developed as a part of the European Unions ESPRIT project MACRAME[2] [3] [4]. It consists of crossbar switches and 100 MBaud DS-links and allows the implementation of various network topologies with up to 512 terminal nodes (In the rest of the paper we will simply write *node* whenever we mean terminal node) [5]. Even though the experiments reported here have been performed using 10 Mbit links, our results can easily be converted to be valid for similar networks with Gbit links or even 10 Gbit links. Since we use crossbars we neither have to take bus speed nor memory speed into account. The performance of a crossbar based interconnect therefore only depend on the collision rate in the crossbars, and that rate will be proportional to the link speed for any given link utilisation factor.

The main objective of the work presented here is to evaluate the performance of the three topologies under study and to determine which network is best suited for application in Gigabit switching fabrics.

The paper is organised as follows. In Section II, we present the network topologies studied and the associated routing schemes. In Section III, we describe the traffic patterns and the performance parameters used in our measurements. Section IV presents and discusses the results for the different topologies and routing schemes and gives a simple cost analysis. Finally, in Section V we present our conclusions.

II. Topologies and Routing

We have studied three different network topologies which we designated the Hybrid, the Dual Hybrid and the Hierarchical Clos. All three networks are so-called multi-stage interconnection networks (MINs) [6] [7].

The network topologies studied are scalable and support Ethernet switches with 8 to 64 Gigabit ports. In our context the term *port* refers to the external connections on the outside of a switch while the term *node* refers to the end-points in the switching fabric inside the switch. Each port is connected to two nodes so the switching fabric will always have twice as many nodes as it has ports. The reason for this is explained in the next section.

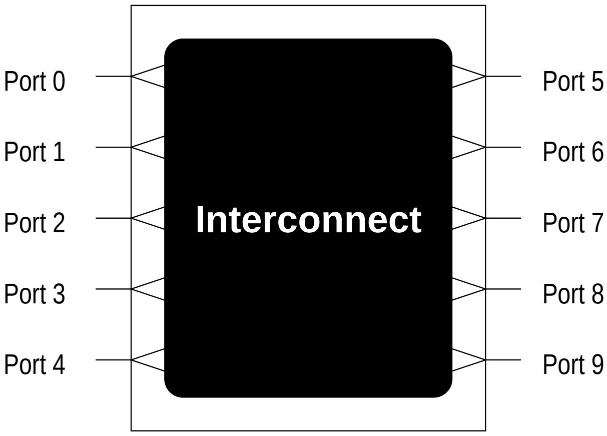

Fig. 1. Conceptual sketch of a switch.

To achieve scalability the switching fabrics are all constructed in a modular fashion. There are two types of modules: an *I/O module* which has 8 Gigabit Ethernet ports and a *matrix module* which is used to connect up to 8 I/O modules for a total of 64 ports. The I/O modules can also be used as stand-alone switches. The basic building blocks for the topologies studied are self-routing 8-way [3] crossbar switches, using worm-hole routing and input queueing which a small amount of buffer space on-chip [8].

The results presented in the following sections focus on the full scale networks, for an in-depth study of the performance of the individual modules and the modular architecture refer to [9]. For performance studies of similar and other topologies refer to [10].

A. The Hybrid Network

The Hybrid is a MIN offering multiple paths between any source-destination pair and in that perspective belongs to the same class of networks as the Beneš [11] and the Clos [12]. Those type of

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

[3] By 8-way we mean a crossbar with 8 I/O links also known as a 4-by-4 crossbar.

networks are often denoted *rearrangeable* due to the multi-path property [13]. The complete 128 node Hybrid is built from 80 8-way crossbars organised as 4 × 4 ($k \times k$ in general terms) switching elements (figure 2). The figured Hybrid consists of 5 stages and is the maximal sized Hybrid that can be constructed from 8-way crossbars (set up as 4 × 4 switching elements). Scaling of a larger sized Hybrid by deploying recursive principles as discussed in [12] and [14] is possible, but is out of the scope of this paper.

The adopted connectivity of the Hybrid is motivated by its flexibility for grouping links. This is utilised by the routing function as discussed later on. The connection pattern between the first and second stages (the interconnecting of the IO-modules) actually applies the style of the Butterfly topology which belongs to another class of MINs that possesses only a unique path between source-destination pairs [15]. Moreover, the pattern may be described as follows:

$$(i\,mod\,k) * k + \lfloor \frac{i}{k^2} \rfloor * k^2 + (\lfloor \frac{i}{k} \rfloor\, mod\,k)$$

Where i is the node number and $k = 4$. The connections between the second stage and the mid stage (i.e. the pattern interconnecting the IO-modules and the Matrixes) are described as follows[4]:

$$(i\,mod\,k^2) * k + \lfloor \frac{i}{k^2} \rfloor$$

B. The Dual Hybrid Network

The Dual Hybrid topology is implemented as two parallel network layers, where each layer consists of a Hybrid network with only half its terminal links connected. Therefore the total number of terminal links for the Dual Hybrid is the same as in the Hybrid. This topology features a significant amount of excess capacity, since the load on each layer is only half that of a Hybrid. However the Dual Hybrid network requires a total of 160 8-way crossbars.

C. Routing Algorithms

The routing schemes applied on all three network topologies make use of Universal Routing (UR) and Grouped Adaptive Routing (GAR).

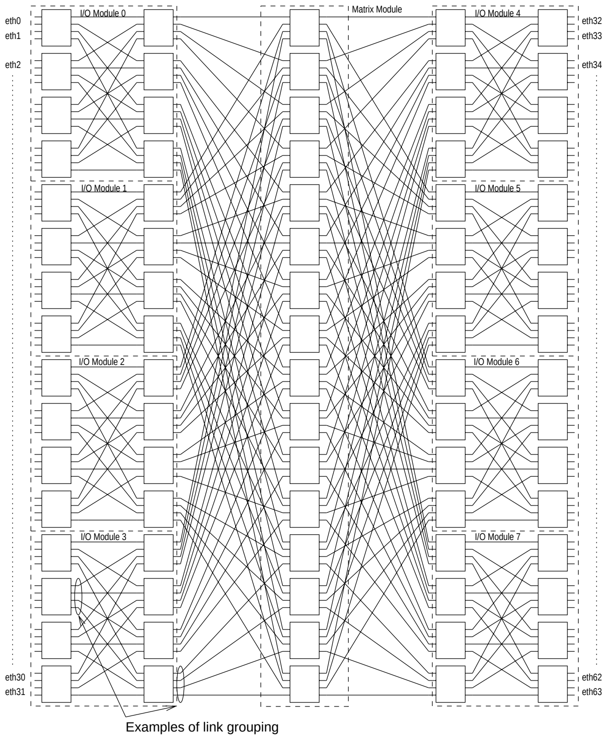

Fig. 2. Link grouping on The Hybrid network.

C.1 Grouped Adaptive Routing

Grouped Adaptive Routing is a form of adaptive routing which can be used to improve the usable bandwidth and fault-tolerance in a network. It is a feature supported by the crossbar switch chips used, which allows a set of links to be logically grouped (see Figure 2 and 3). This then enables a packet to take one of several possible path through the network, depending on the local state of contention, by allowing a packet to be sent down any link in the group which is not currently in use. This scheme improves performance by ensuring that there are no packets waiting to use one link when an equivalent link is idle.

C.2 Universal Routing

Universal Routing works by routing packets in two phases [16]. First the packet is routed to a random intermediate destination from where it is then routed deterministically to its final destination. UR can provide fair sharing of bandwidth between terminal nodes and good load balanc-

[4] Note that the defined inter stage connectivity is symmetrical around the mid stage

ing on multistage networks, however this comes at the cost of reducing the overall throughput, since links that are already busy can be selected again by the random routing scheme.

C.3 Deadlock

An important property of a network and its routing function is that it should be deadlock free. Deadlock occurs when a packet gets blocked forever because of a resource dependency cycle in the network. The resources can be buffers or links and a conflict can occur when a packet holding one resource is allowed to request another. The routing algorithms presented here have been designed to avoid deadlock.

D. Routing on the Hybrid Networks

Results for two different routing schemes, labelled *RS#1* and *RS#2*, are presented. The same routing schemes are used on the two Hybrid network topologies.

RS#1 uses UR by sending each packet to any of the 16 centre stage switches in the matrix module at random. From there the packet is then routed deterministically to its destination I/O module. This routing scheme also ensures that the number of switches a packet has to traverse on its path from source to destination is always 5 hops.

RS#2 uses GAR on the links from the terminal nodes to the matrix, since a packet can be sent through any of the 16 centre stage switches in the matrix module. The fact that the two terminal links associated with one external switch port are equivalent allows another level of grouping to be exploited.

E. The Hierarchical Clos Network

The Hierarchical Clos topology is built from several rearrangeable non-blocking 32-way Clos networks [12]. Each I/O module consists of one 32-way Clos and the matrix module used to interconnect the I/O modules consists of four 32-way Clos subnetworks. The complete 128 node network shown in Figure 3 requires a total of 144 8-way crossbars. Note that the Clos subnetworks in the I/O modules and the matrix modules are identical, but they seem different in figure 3 since the Clos networks for the I/O modules have been folded so all the nodes are on the same side. This is done just to make the drawing of the figure more convenient.

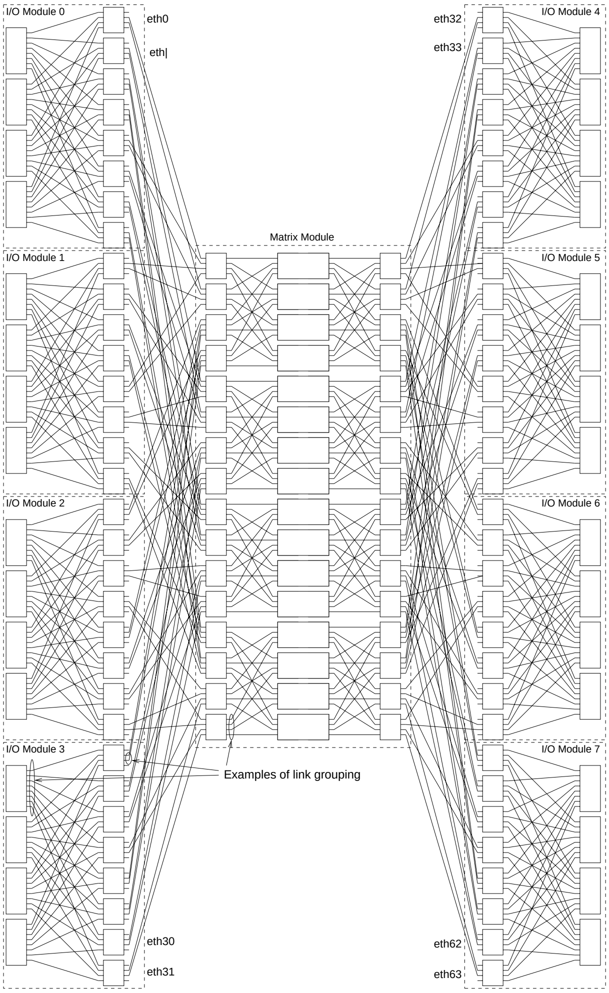

Fig. 3. Link grouping on The Hierarchical Clos.

F. Routing on the Hierarchical Clos

On the Hierarchical Clos we present results from three different routing schemes, labelled *RS#1*, RS#2 and *RS#3*.

RS#1 uses a combination of GAR and UR, where all packets pass through the centre stage of the matrix module and the centre stage of the destination I/O module before it reaching their destination terminal. The only exception to this rule is for packets destined for a terminal node connected to the same crossbar as they originated from. UR is used for routing packets from the first stage of the matrix module to its centre stage and also for routing packets from the first stage of the destination I/O module to its centre stage. With this routing scheme the path length of a packet is normally seven hops, except in the case where source and destination are on the same switch.

RS#2 uses GAR exclusively. In addition it differentiates between local [5] and non-local traffic. In the centre stage of the I/O modules two of the crossbars are dedicated to local traffic and the other two are dedicated to non-local traffic. This is necessary to avoid deadlock [9]. With this scheme the path length of a packet can be one, three, seven or nine hops, depending on the location of the source and destination terminals.

RS#3 extends *RS#2* with UR thereby spreading the load evenly across the matrix module. This also ensures that the path length for all packets equal to nine hops. For a more detailed description of these and other routing schemes tested refer to [9].

III. Traffic measurements

A traffic pattern describes, among other things, how the source and destination pairs are chosen when running measurements. In order to obtain meaningful measurement results, the traffic patterns used should reflect the expected traffic load. Packet length, packet source/destination mapping and applied load are key parameters which define the traffic pattern. For the measurements presented here, the packet length was kept constant[6], since variable length Ethernet frames would normally be fragmented into smaller constant length packets before transmitting them across the internal switching fabric. Results for two different traffic patterns are presented, namely *random destinations* and *permuted pairs*.

A. Random destinations

With random destinations each source randomly sends packets to any destination in the network except itself. This traffic pattern causes destination contention, since it allows several nodes to send to a given destination simultaneously. The destination selection is based on a uniform distribution.

B. Permuted pairs

With permuted pairs traffic, fixed pairs of nodes are communicating, i.e. each nodes only sends packets to one destination and only receives packets from one source. Therefore the permuted pairs pattern does not cause destination contention, however internal network contention might still occur, e.g. when several packets have to share the bandwidth of a link along the path to their destination.

We have used sets of 64x2 pairs to better emulate how the 64 external Gigabit Ethernet ports are split across two internal fabric links by using the same set of 64 independent permutations twice. Thus the two terminals associated to one external switch port are transmitting and receiving correlated traffic.

The networks under study here are not strictly non-blocking such as a full crossbar. We have therefore performed measurements with 100 permutations selected at random, in order to evaluate the effect of blocking due to varying source-destination pairings.

C. Performance parameters

Our results will focus on the two performance parameters throughput and fairness, which is characterised by the variation in throughput across all the nodes. The following results are presented for each measurement:

- Average throughput: the achieved transmit rate averaged over all the active terminals.
- Maximum throughput: the maximum transmit rate achieved by any of the source terminals.
- Minimum throughput: the minimum transmit rate achieved by any of the source terminals.
- Standard deviation of the throughput: the standard deviation of the transmit rate of all the active terminals.

The results are summarised in Table I–III and in Figure 4 and 5. All values shown are shown as a percentage of the maximum internal fabric link bandwidth, i.e. they represent the link utilisation of the connected terminal links.

[5] Traffic with source and destination on the same I/O module

[6] For results from variable length packet measurements see [9]

RANDOM DESTINATIONS				
Routing	Avg	Max	Min	Std
RS#1	38.2%	39.2%	37.1%	0.46
100 RANDOM PERMUTATIONS				
Routing	Avg	Max	Min	Std
RS#1	40.3%	41.2%	39.8%	0.21
RS#2	45.7%	55.2%	41.9%	1.87

TABLE I
The Hybrid per node link utilisation with 64 byte packets at 100% load.

The link utilisation histograms in Figure 4 and 5 were obtained by measuring the network performance for 100 permutations selected at random and creating histograms for the utilisation of the terminal links.

IV. Results

A. The Hybrid

From the results in Table I and Figure 4 it is clear that the routing scheme *RS#1* gives the best performance on the Hybrid. *RS#1* gives an average link utilisation of 40.3% combined with a low spread in throughput for random permutations, compared to 45.7% for *RS#2*. This 5.4% increase comes at the cost of an increase in the spread in throughput, i.e. less fair use of network resources. The average throughput is higher for *RS#2* because there is less contention towards the centre stage since the links are grouped. A packet can therefore use any of the four links towards the centre stage instead of just one which is the case with *RS#1*, which uses UR. However, the spread in throughput is also higher, since the contention is no longer uniformly distributed across all the links to the centre stage, which is the case when using the UR scheme.

For the random destinations it is worth observing that the performance is not much affected when we compare it to the 100 permuted pairs. This is because of the UR which introduces fairness for all packets independent of the applied traffic pattern.

RANDOM DESTINATIONS				
Routing	Avg	Max	Min	Std
RS#1	59.3%	59.4%	59.1%	NA
100 RANDOM PERMUTATIONS				
Routing	Avg	Max	Min	Std
RS#1	70.0%	73.0%	69.1%	0.46
RS#2	74.8%	79.2%	72.6%	0.90

TABLE II
The Dual Hybrid per node link utilisation with 64 byte packets at 100% load.

B. The Dual Hybrid

Table II and Figure 4 shows the results for the Dual Hybrid. As with the Hybrid we see an increase in throughput and a decrease in fairness between RS#1 and RS#2. For the Dual Hybrid the increase in throughput is 4.8%. More interesting though is the comparison of the Hybrid and the Dual Hybrid results. From Figure 4 it is clear that the Dual Hybrid gives us a major leap in performance. Comparing the results for RS#1 in Table I and II we see a 29.7% increase in link utilisation for the Dual Hybrid. In addition to the high throughput values the Dual Hybrid also has a low spread. This shows that excess capacity in the network improves link utilisation of the terminal links, since contention occurs less frequently.

C. The Hierarchical Clos

Looking at the Hierarchical Clos and its three routing schemes we notice major differences in the spread of the terminal link utilisation. The results in Table III and Figure 5 clearly show that RS#1 gives the highest average through-

100 RANDOM PERMUTATIONS				
Routing	Avg	Max	Min	Std
RS#1	64.9%	94.7%	29.0%	10.8
RS#2	57.4%	95.0%	32.3%	7.4
RS#3	55.1%	57.0%	53.4%	0.62

TABLE III
The Hierarchical Clos per node link utilisation with 64 byte packets at 100% load.

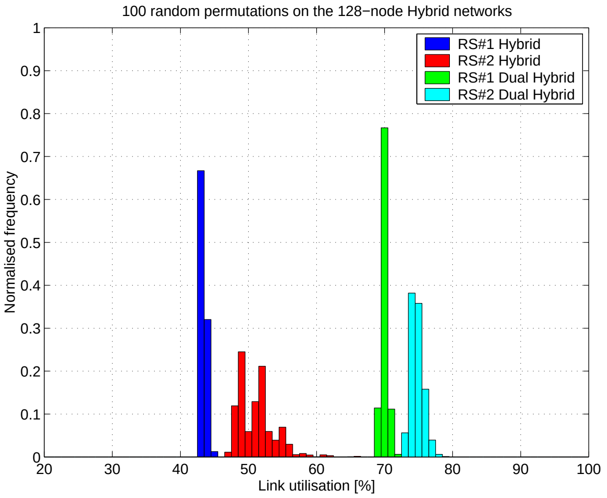

Fig. 4. Link utilisation on the Hybrid networks with 64 byte packets, 100% load and a 100 random permutations.

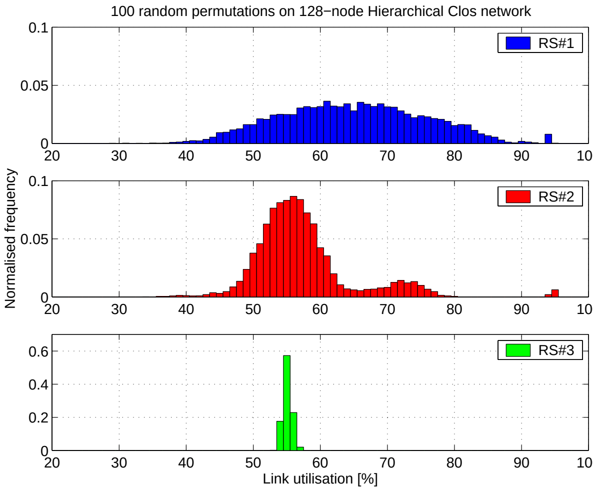

Fig. 5. Link utilisation on the Hierarchical Clos with 64 byte packets, 100% load and a 100 random permutations.

put with 64.9% link utilisation, however this is at the price of a very wide throughput variation. There is a difference of 65.7% between the maximum and minimum throughput. The main reason for this large variation is that the load is not spread evenly over the matrix, since *RS#1* routes packets from a given source terminal always through the same matrix switches.

Moving on to RS#2 we observe a reduction in the average link utilisation from 64.9% to 57.4%. The results in Table III also show a slight reduction in the spread of the link utilisation values. The histogram in Figure 5 shows three distinct peaks in the graph for RS#2. which can be explained as follows. The first peak is for non-local traffic, the second peak is for local traffic within an I/O module and the third one is for local traffic within a single crossbar. The variation in throughput is high because the path length of the packets travelling through the network can be either one, three, seven or nine hops.

RS#3 results in a small reduction in the average link utilisation from 57.4% for RS#2 to 55.1%. However the fairness has improved dramatically with RS#3. This can be clearly seen in Figure 5, which only shows a small variation in the link utilisation for *RS#3*. The improvement in fairness is due to the use of UR, which helps to spread the load over the matrix, and due to the fact that the path length for all packets is now equal (9 hops).

D. *Performance and implementation cost*

We use the number of crossbars necessary to build a 128 node network as a parameter to quantify the cost of implementing a given network topology. Table IV shows the best result obtained for the 100 permuted pairs, the number of 8-way crossbar switches required and the performance to cost ratio for the three topologies studied.

Topology	Utilisation	Switches	Ratio
H. Clos	55.1%	144	0.38
Hybrid	45.7%	80	0.57
Dual Hybrid	74.8%	160	0.47

TABLE IV
Performance to cost ratio of the Hierarchical Clos, the Hybrid and the Dual Hybrid.

The Hybrid network provides the best cost to performance tradeoff, however it also has the worst performance of the three networks. The best performance by far is obtained with the Dual Hybrid network, followed by the Hierarchical Clos network. The cost to performance ratio of these two networks is similar. However it is worth noting that our simple cost analysis does not take the cost of the connections between the I/O modules and the matrix module into account. The cost associated with this is likely to be significantly higher in the case of the Dual Hybrid network, which means that the performance/cost ratio would not be as good as shown in Table IV.

V. Conclusions

Based on our results we draw the following general conclusions concerning the routing strategies applied to the chosen network topologies:

- Universal routing provides good load balancing on multistage networks by distributing the load evenly across the centre stage switches. It also helps the fair sharing of the network bandwidth between the terminal nodes. Therefore UR can provide predictable per-node throughput at the expense of a reduction in the average performance.
- Grouped adaptive routing gives the highest average throughput, however it also results in large variations of the per-node throughput for permutation traffic.
- Ensuring an equal path length for all packets is essential to obtain a small spread of the terminal node throughput.
- A combination of these routing strategies results in a good balance between achievable throughput and fairness.

Concerning the network topologies we observe the following:

- The Dual Hybrid gives the highest performance for a 128 node network. The main drawback of this topology is its implementation complexity.
- The Hierarchical Clos has the worst cost/performance ratio. However it might be preferable to the Dual Hybrid, since it is simpler to implement.
- We saw that excess switching capacity in the network contributes strongly to achieve good link utilisation.

Acknowledgements

The authors would like to acknowledge the support from the European Union through the ESPRIT project Swift (EP28742).

References

[1] S. Keshav and R. Sharma. Issues and trends in router design. *IEEE Communications Magazine*, 36(5), 1998.

[2] *IEEE Standard 1355-1995, Standard for Heterogeneous InterConnect (HIC).* IEEE, 1995.

[3] David Arnould Thornley. *A Test Bed for Evaluating the Performance of Very Large IEEE1355 Networks.* PhD thesis, University of Kent at Canterbury, 1998.

[4] S. Haas, D.A. Thornley, M. Zhu, R.W. Dobinson, and B. Martin. The macramé 1024 node switching network. *Microprocessors and Microsystems*, (21):511–518, 1998.

[5] Stefan Haas. *The IEEE1355 Standard: Developments, Performance and Application in High Energy Physics.* PhD thesis, University of Liverpool, 1998.

[6] Jose Duato, Sudhakar Yalamanchili, and Lionel Ni. *Interconnection Networks an Engineering Approach.* IEEE Computer Society, 1997.

[7] Daniel A. Reed and Richard M. Fujimoto. *Multicomputer Networks, Message-Based Parallel Processing.* The MIT Press, 1987.

[8] SGS Thomsom Microelectronics. *STC104 Asynchrouns packet switch, Engineering data*, 1995.

[9] Sven-Arne Reinemo. Topologies and routing in gigabit switching fabrics. Master's thesis, Department of Informatics, University of Oslo, November 2000.

[10] A.M. Jones, N. J. Davis, M. A. Firth, and C. J. Wright. *The Network Designer's Handbook.* SRF-PACT, 2 edition, 1997.

[11] V. Benes. *Mathematical Theory of Connecting Networks and Telephone Traffic.* Academic Press, New York, 1965.

[12] Charles Clos. A study of non-blocking switching networks. *The Bell System Technical Journal*, March 1953.

[13] Y.M. Yeh and T.Y. Feng. Fault-tolerant routing on a class of rearrangeable networks. In *In Proc. 1991 Internat. Conf. on Parallel Processing*, pages 305–312, 1991.

[14] Tor Skeie, Olav Lysne, and Geir Horn. Scalable non-blocking networks with fixed size routers. In *Proc. of the Intern. Conf. on Parallel and Distributed Processing Techniques and Applications*, pages 1308–1314, 1997.

[15] H.J. Siegel, W.G. Nation, C.P. Kruskal, and L.M. Napolitano Jr. Using the multistage cube network topology in parallel supercomputers. In *Proc. IEEE*, pages 1932–1953, 1989.

[16] L. G. Valiant. A scheme for fast parallel communication. *SIAM J. on Computing*, 11:350–361, 1982.