Cost-Efficient SCI-based Banyan Networks

Harald Richter*[+], Richard Kleber[+] and Hermann Hellwagner[+],
*Max-Planck-Institute for Plasma Physics, Garching, Germany,
EURATOM Association,
[+]Lehrstuhl für Rechnertechnik und -organisation,
Technical University of Munich,
D-80290 Munich, Germany,
Tel. +49 89 289 22382, Fax. +49 89 289 28232,
{richterh, hellwagn}@informatik.tu-muenchen.de

Keywords: Cluster Computing, Scalable Coherent Interface, Banyan Networks.

I. Introduction

Today, various high speed transmission technologies and protocols such as Fiber Channel, FDDI, ATM and Fast Ethernet are wide spread, complemented by system- or local area communication networks such as HIPPI, OMI-HIC [IEEEP], Myrinet [Boden95], RACEway [Mercury95], Memory Channel [Gillet96] and SCI (Scalable Coherent Interface) [IEEE92]. This paper discusses the efficient use of 4-port SCI switches, resulting in 5-fold performance compared to SCI switches in the standard operating mode. Furthermore, Banyan-like topologies are proposed based on such efficient switches. The novel networks are suited for applications in parallel computers, clusters of workstations [Ander95] and local area multiprocessors [Gustav94]. Their performance improvements are the higher the larger the switch size is, while exhibiting lower costs at the same time. The prerequisite is that data locality in the traffic patterns is present.

In chapter 2, a short review on the SCI technology and the internal set-up of SCI switches is given. In chapter 3, the efficient use of SCI switches in uniform and non-uniform memory-access architectures (UMA/NUMA) is presented. In the fourth chapter, the achieved improvements are applied for multistage Banyans, that are constructed by an intra-switch wiring on higher number basis than 2, resulting in enhanced topologies. In the sixth chapter, performance figures for the 4-port switch and the multistage topologies are given for both, bi- and unidirectionally-connected operating modes. The results show better behaviour for bandwidth, latency and packet losses compared to standard SCI-based Banyans. Figures were obtained by means of a self-written simulation program called SCINET.

II. The SCI technology

SCI is an IEEE and ANSI standard [IEEE92] for high-speed, low-latency data exchange between processor and memory nodes and peripheral devices. The basic topological element of SCI is an unidirectional ring. Data rates can achieve up to 2 Gbytes/s by using optical flat-ribbon cables [Enge96]. Latency times are reported to be in the range of a few µs between user applications on different nodes [Omang96]. SCI incorporates all features of a modern bus such as split transactions, pipelining and broadcast, but operates in a functional and spatial distributed environment. In each of the up to 64 K nodes that are possible in a SCI system, a local memory of 2^{48} bytes can be addressed to allow for a distributed shared-memory with an address range of 64 bits. Split transactions are established by separate request and response subactions which are individually acknowledged by echo packets. The ring is not blocked between a request and its response to allow for maximum utilization of the ring's bandwidth by pipelining multiple requests. Up to 64 requests can be pending per SCI ring.

III. SCI switches

In addition to commercial SCI products for PCs and workstations, SCI switches [Dolphin95] have also become available that allow to connect two or more SCI rings, thereby forming a static or dynamic SCI network. Each SCI net can be composed of nodes such as computers, processors, memories, peripherals, routers, bridges and switches. By proper address management, a commercial 4-port SCI switch can act as a router, if it is connected with one port to a SCI node and with the remaining ports to a static network such as a torus. It can act as a bridge, if it is located between adjacent rings to allow data to pass, and it can be employed to establish multistage networks (Figure 1a, b and c).

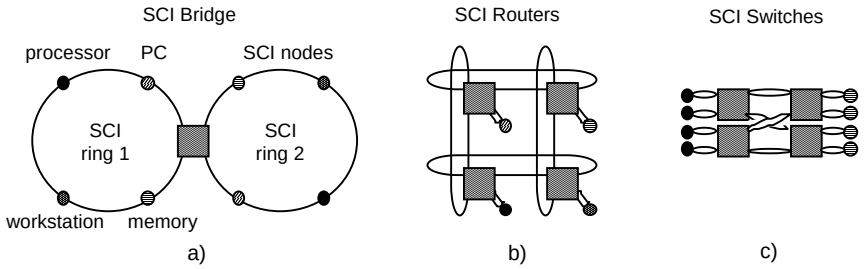

Figure 1: SCI switches in bridge- (a), router- (b) and multistage-net (c).

An SCI switch differs in various respects from conventional switches. Firstly, each SCI switch is part of 2-4 rings on which data are unidirectionally transferred. Secondly, there exists a port-internal bypass-FIFO connecting the in- and out-terminal of each port to allow a very fast bypass (<50ns). Thirdly, in each port two separate buffers for SCI requests and responses are available to prevent from deadlocks induced by mutual waiting of resources (Figure 2).

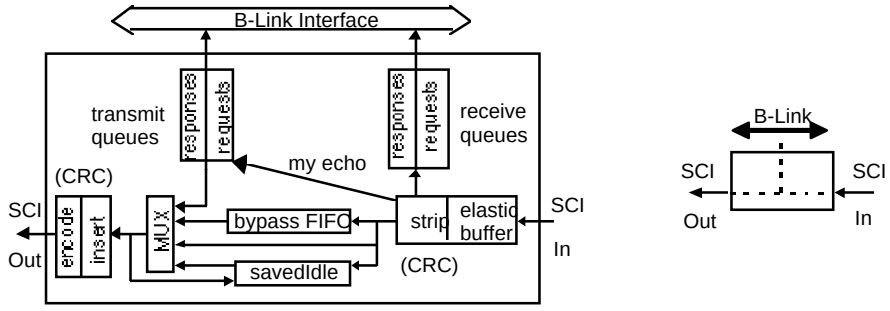

Figure 2: SCI switch port (a) and its symbolic representation (b).

In the following, we consider switches according to Dolphin's implementation [Dolphin95]. The transmit- and receive buffers of the ports of such a switch are connected to a high-speed packet-bus called B-Link [Dolphin94], which has a transmission rate of 600 MB/s in the latest version. Inside the switch, the B-Link connects 4 ports, from which each is capable of 500 MB/s data rate per direction. By this, a high-speed SCI switch is established. Between any pair of ports, the maximum port rate of 500 MB/s can be achieved for unidirectional transfers (either read or write) as long as the remaining other pair of ports produces not more than 100 MB/s of traffic. If both pairs simultaneously operate in full duplex mode, the individual port rate per direction is reduced to 150 Mbytes/s due to the B-Link's bandwidth limitations.

For better clarity, in Figure 3a, a block diagram of the 4-port switch is given, and a functional equivalent representation is shown in Figure 3b.

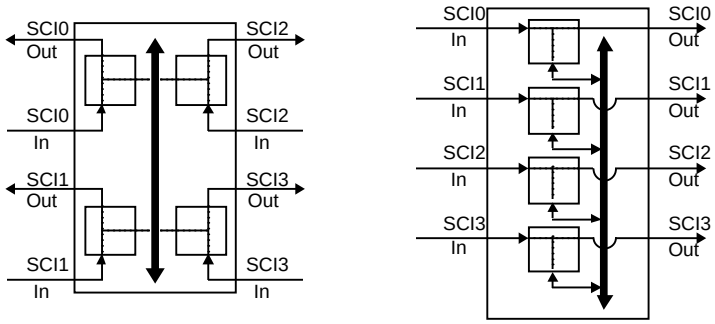

Figure 3: Two equivalent representations of a 4-port SCI switch.

IV. Efficient use of SCI switches

Let us define the throughput T of a switch as the sum of the ports' throughputs. With SCI, it is possible to push T above the bandwidth limit B of the switch-internal B-Link by using the ports' bypass-fifos for additional data transfers. This efficient switch usage is illustrated in the following figures. For comparison with SCI, in Figure 4 two UMA/NUMA-multiprocessors are depicted that employ conventional switches.

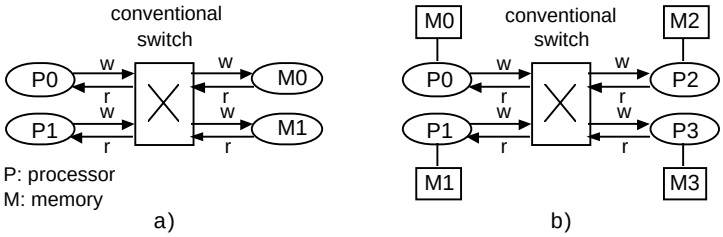

Figure 4: UMA (a) and NUMA (b) multiprocessor with conventional switches.

In Figure 4a, the throughput of the conventional switch is assumed to be T_{con}, with $T_{con} \leq 2t$, where t is the throughput of a single switch port to which a processor is connected. In the NUMA example of Figure 4b, every pair of computing nodes can simultaneously communicate with each other (up to two pairs at the same time), thus pushing the throughput to T_{con}' with $T_{con}' \leq 4t$. In both cases, the switch-internal transfer capacity B_{tr} is assumed to be sufficiently large to carry the produced traffic ($B_{tr} \geq T_{con}$).

In Figure 5a, the UMA architecture of Figure 4a is upgraded to an SCI-switch, and the bidirectional transmission lines are replaced by bidirectional SCI rings. The total throughput is T_{SCI}, with $T_{SCI}=\min\{2t, B_{tr}\}$ which is the same as with conventional switches. This solution is published in the literature [Krist94, WU94]. In Figure 5b however, the nodes are coupled more efficiently. Here, we have T_{SCI}' as throughput which can become larger than B_{tr}, provided that some fraction of the data can stay on the ring where it originated to reach the target. The reason for higher throughput is that data may enter and leave the switch through the ports' bypass FIFOs, so that the B-Link bottleneck is circumvented.

Because of the fact that in SCI the intra-ring communication is faster than the inter-ring communication, the latency is also reduced. Additionally, from Figure 5a to Figure 5b the amount of required hardware has decreased from 4 to 2 rings and from a 4-port to a 2-port switch while the performance has increased. For larger switch sizes, the same factor of two can be achieved.

Finally as shown in Figure 6a and b, more flexibility with respect to the number of installable processors and memories can be obtained by the proposed usage of SCI switches. In the shown example of Figure 6a, twice as much processors and memories can be connected to the same 4-port switch without

degradation. In the example of Figure 6b an other factor of two can be obtained. Of course, in the latter case the maximum date rate per processor is halved.

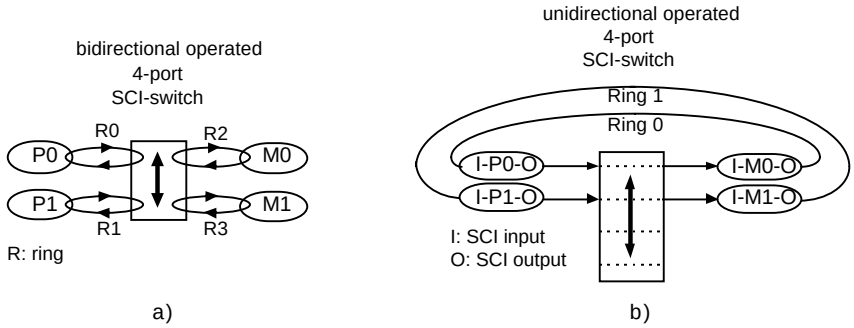

Figure 5: SCI net using B-Link (a) or bypass-Fifo (b) as main data path.

The prerequisite that T_{SCI}' surpasses B_{tr} is that in the communication pattern of the data packets exists some data locality. This means for the example of Figure 6a that processor P_i (i=0,1,2,3) mainly wants to communicate with memory M_i, so that data can stay on their rings and travel through the bypass fifos. Normally, this is the case since data locality is common to most parallel applications. If not, it can be manually forced by proper allocation of tasks and data structures to computing nodes.

To summarize, data locality together with SCI technology allows to increase the performance of a 4-port SCI switch by conducting traffic not via the switch-internal B-Link but via the ports' bypass-Fifos. In chapter 6, the performance improvement will be quantitatively specified.

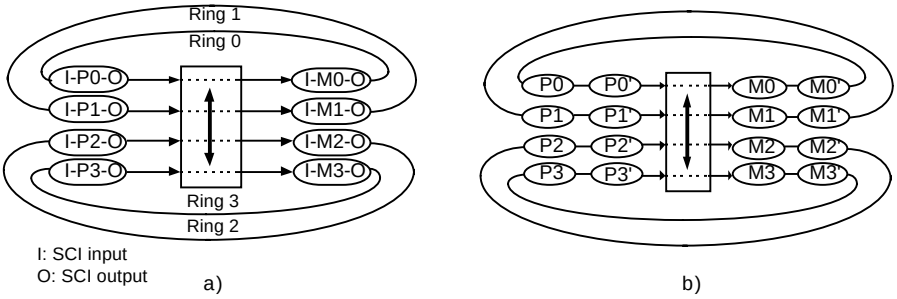

Figure 6: Enhanced flexibility in the number of connected processors.

V. Multistage SCI networks

In this chapter, the described optimization technique is transferred to multistage networks that allow for larger switch sizes than four. Generally, the most cost-effective multistage networks are of the Banyan type [Goke73], since they can be built with the minimum number of stages. However, Banyans are blocking networks which do not have redundancy and therefore also no fault tolerance. Typical Banyans are Baseline-, Omega-, Flip-, Butterfly-, Indirect Binary n-Cube, and Generalized Cube networks.

In Figure 7a, the standard implementation according to [Wu94] of an SCI-based Baseline-network is shown: Nodes and switch ports of adjacent stages are connected by SCI-ringlets. The optimized bypass-FIFO-solution, is shown in Figure 7b. In the optimized network, large toroidal rings which are going through all stages of the network are required.

Furthermore, in the example of Figure 7b, the switch size was reduced from 4 to 2-port switches, and the minimum latency to travel from input to output of the net of size N has decreased from $L=\alpha\log_2 N$ to

L'=$\beta \log_2 N$, where α is the time to travel inside a switch from one port to an other (≈some μs) and β is the time to travel through a bypass FIFO(≈some tens of ns).

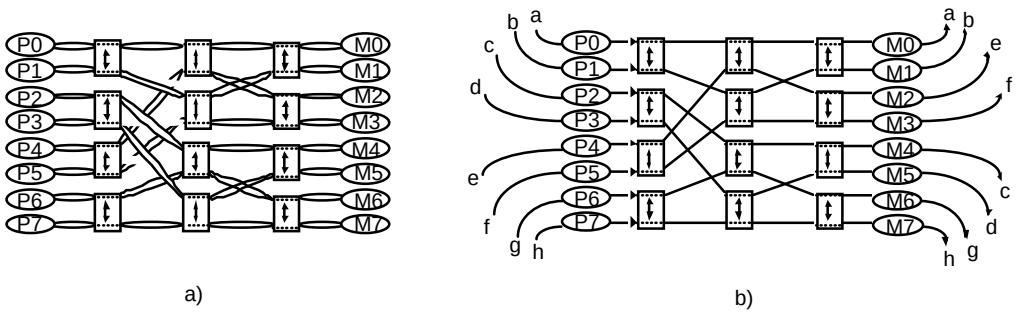

Figure 7: Baseline nets with B-Links or with bypass-Fifos as main paths.

A disadvantage of the Baseline-Network of Figure 7b is that an additional permutation wiring is required to connect memory out-terminals with processor in-terminals to obtain closed rings. Fortunately, two of the known Banyans allow by virtue of their topological structure a 1:1 connection from output to input. These topologies are the Omega- and the Generalized Cube network and their mirror images, the Flip net and the Indirect Binary n-Cube. Therefore, it is here proposed that SCI-nets that have bypass-Fifos as their main data paths should be one of these 4 topologies. In the following, the 1:1 nets are termed *first grade optimized*, and an example based on the Omega-topology is given in Figure 8.

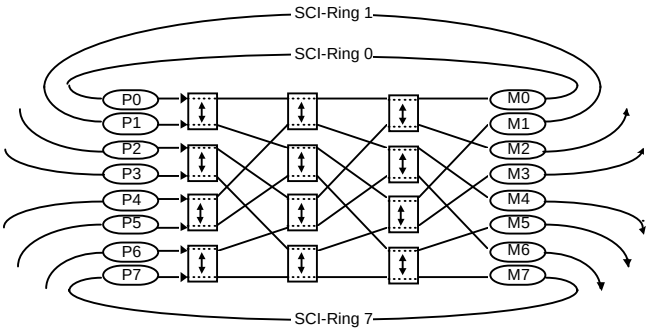

Figure 8: First-grade-optimized SCI-based Omega-net.

First-grade optimized SCI-Banyans can be further improved by using s-port switches with s>2 which results in a factor of improvement in terms of costs and latency of $\frac{\log_2 N}{\log_s N}$. For example, if s = 4 is assumed such a network of size 16x16 (depicted in Figure 9) requires 32 ports, while a first-grade optimized net of the same size needs 64 ports, and a SCI-net with bidirectional ringlets similar to Figure 7a requires already 128 ports. If s>2 holds we call such structures *second grade optimized*. In chapter 6, the performances of first and second grad optimized nets will be compared with each other.

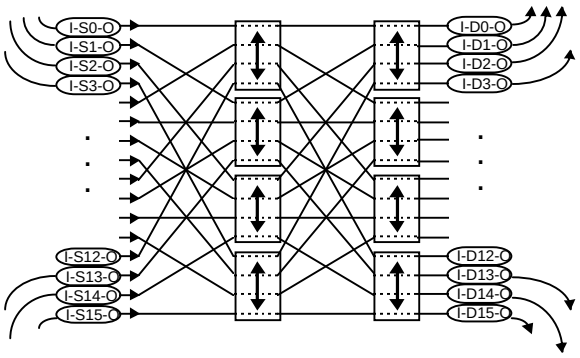

Figure 9: Second-grade optimized SCI-based Omega-net.

VI. Simulation Results

The first suite of simulations was conducted to evaluate the performance of a single 4-port SCI switch which uses conventional bidirectional ringlets to connect processors and memories, and to compare that solution with a switch with unidirectional rings. The performance metrics are throughput, latency and packet losses. To pinpoint the performance discrepancy between both concepts, a data locality of 100 % was chosen, i.e. processor P_i (i=0,1) communicates exclusively with memory M_i. In practice, the factor will be lower, but as long as there is some data locality a performance improvement is visible.

For all simulations, the responseless DMOVE64 transaction was taken, and the processors are configured to send DMOVE64 request packets at the same time and with the same rate to the switch. The data rate, i.e. the packet period, is chosen to be deterministic.

On the x-axis of the following graphs, the total raw data rate of the traffic that is injected into the switch is shown as simulation parameter. Its value is varied from 0 to 500 MB/s per processor which is also the maximum rate of an SCI-port. The data packets have 64 Bytes payload, an overhead of 16 Bytes for header and trailer and an additional 4 Bytes overhead on the ring for idle packets, so that the ratio of payload length to raw length is 64/84.

The memories are assumed to have a cycle time of 40 ns to store a 64 Byte value. The link delays between processors, switch and memories are set to 1 ns each. The timing parameters for all SCI ports are 20 ns address decoder delay, 48 ns bypass fifo delay, 106 ns fifo to B-Link delay and 82 ns B-Link to fifo delay. All ports are modelled to have input and output buffer space for 4 request and response packets each. By this, the simulations are compliant with the “LC-II” SCI-port specification of Dolphin [Dolphin97].

On the y-axis of Figure 10, the achieved network throughputs of a bidirectionally connected 4-port switch (bO2nS2) as depicted in Figure 5a, and a unidirectionally connected switch (uO2nS2) as depicted in Figure 5b are given. With bidirectional ringlets, the switch saturates with 272 MB/s output payload at 400 MB/s raw input rate. At the same input rate, the packet losses become significant and reach up to 1823 MB/s at 1 GB/s (Figure 11). A packet loss occurs each time when a new packet is generated while the switch is still occupied to transfer previous packets, provided that all input buffers are full. Because of the constant period with which data are issued by the processors, the switch has to accept packets in real time which is possible only up to a certain speed. Above that limit, packets are lost.

The latency from initiating a DMOVE64 packet to its storing jumps in the vicinity of the saturation point from 1242 ns up to 5230 ns (Figure 12). The unidirectionally connected switch instead behaves much better: it saturates with 682 MB/s output payload at 900 MB/s input rate, has 83 MB/s packet losses at 1 GB/s and a maximum latency of 1895 ns. Please observe, that for its simulations only 2 of the four ports of the switch were connected.

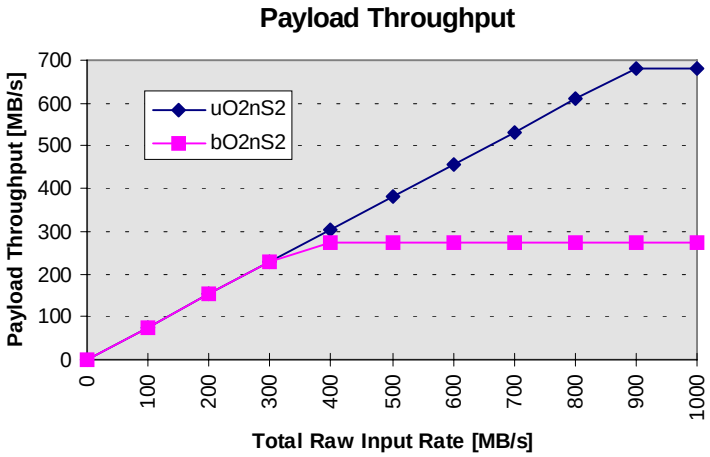

Figure 10: Throughput of a bi-/unidirectionally connected 4-port switch.

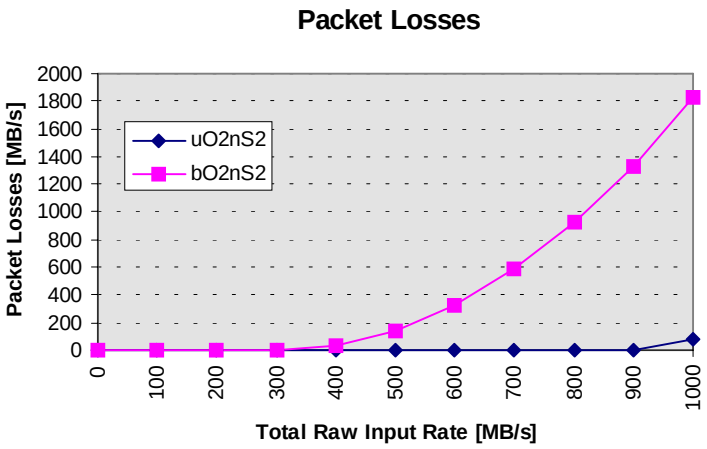

Figure 11: Packet losses of a bi-/unidirectionally connected 4-port switch.

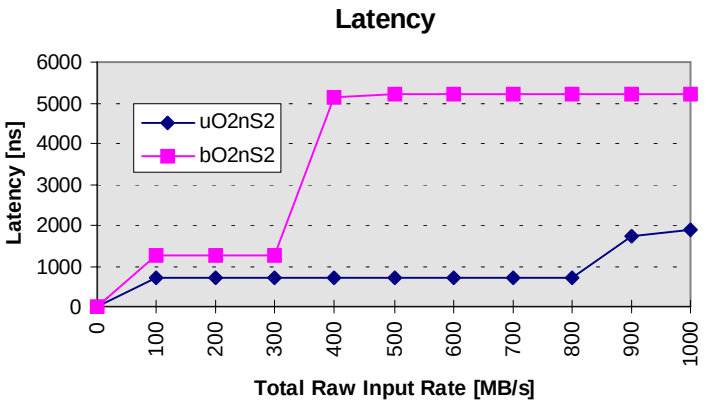

Figure 12: Latency of a bi-/unidirectionally connected 4-port switch.

If all 4 ports of the switch are coupled with unidirectional rings to processors and memories (as depicted in Figure 6a), a linear performance improvement compared to 2 processors/memories is achieved. The switch saturates with 1364 MB/s output payload at 900 MB/s input rate, has 167 MB/s packet losses at 1 GB/s and a maximum latency of 1895 ns. This means that with the same switch as in the bidirectionally connected case, a 5-fold throughput improvement can be achieved.

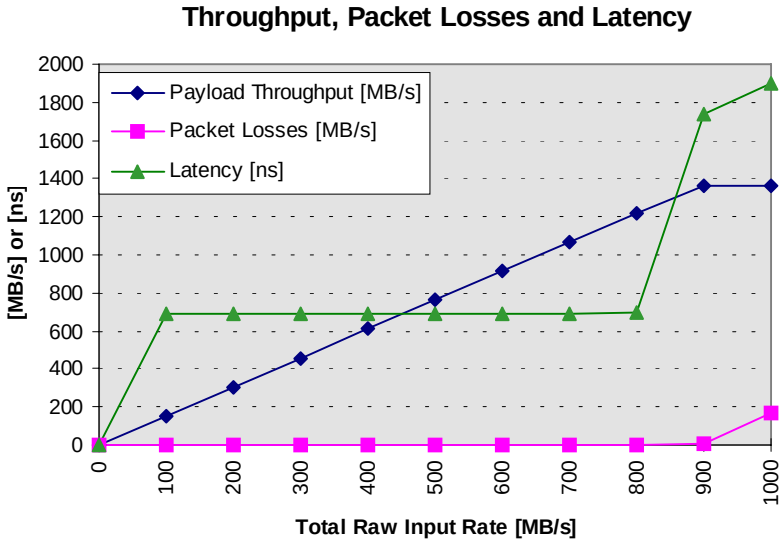

Figure 13: Performance of 4-port switch with 4 processors/memories.

The second simulation suite compares a 16x16 first-grade optimized Omega-net, abbreviated uO2nS16, with a second-grade optimized one (uO4nS16). The latter multistage network shows better performance in all respects. Both are superior to a bidirectionally connected net of same type and size while having significantly lower costs.

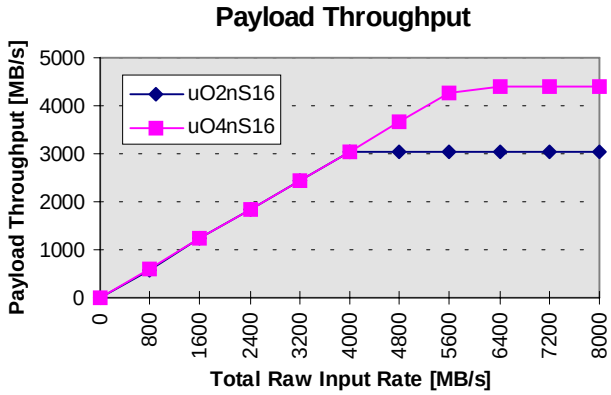

Figure 14: Throughput of first- and second-grade optimized Omega-net.

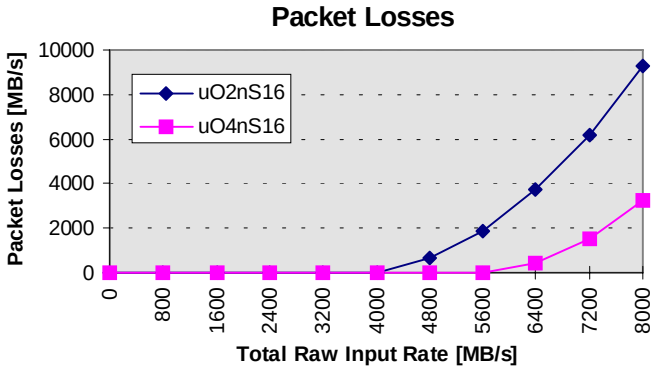

Figure 15: Packet losses of first- and second-grade optimized Omega-net.

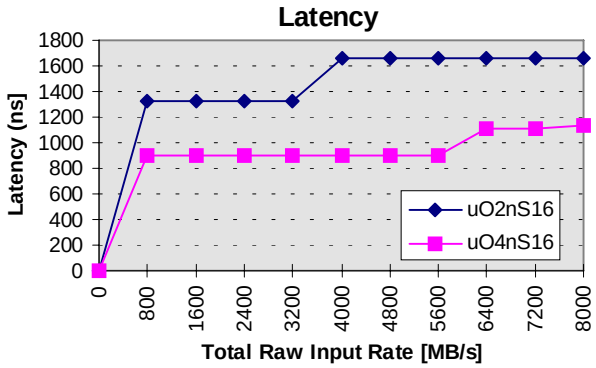

Figure 16: Latency of first- and second-grade optimized Omega-net.

VII. Summary and conclusion

We have shown how the bypass-FIFOs that are inherent to SCI ports can be used to push the performance in terms of throughput and latency for a 4-port SCI switch beyond its internal transfer-capacity limit, provided that some data locality is present in the traffic pattern. The principle is to redirect packets from switch-internal transfer links to the ports' bypass-Fifos. By this method, more flexibility in the number of installable nodes can also be achieved. Compared to a bidirectionally connected switch, a 5-fold throughput improvement can be achieved for 100% data locality. Latency is reduced by a factor of 2.8. The number of lost packets decreases by a factor of 11.

The redirection technique was also applied to multistage Banyan networks of larger size, and new first and second grade-optimized SCI-based Omega-nets were proposed that are using unidirectional SCI-rings with permutation bases of s with s≥2. The given figures of throughput, latency and packet losses show better performance compared to SCI-Banyans based on bidirectional ringlets while having lower costs at the same time. The Results were achieved by the self-developed SCINET simulation program.

VIII. Literature

[Ander95] T. E. Anderson, D. E. Culler, D. A Patterson, “A Case for NOW (Networks of Workstations)”, *IEEE Micro*, Feb. 1995, pp.54-64.

[Boden95] N. Boden, D. Cohen, R. Felderman, A. Kulawik, C. Seitz, J. Seizovic, and W.-K. Su, “Myrinet: A gigabit-per-second local area network,” *IEEE Micro*, vol. 15, pp. 29-35, Feb. 1995.

[Bogaerts94] A. Bogaerts, R. Keyser, G. Mugnai, H. Müller, P. Werner, B. Wu, B. Skaali, J. Ferrer-Prietro, “SCI Data Acquisiton Systems: Doing more with less,“ *CHEP’94*, San Franciso, April 1994.

[CACI95] CACI Products Company, “Modsim II, The Language for Object Oriented programming”, *Reference Manual*, La Jolla, California, 1995.

[Dolphin94] Dolphin, “A Backside Link (B-Link) for Scalable Coherent Interface (SCI) Nodes”, *Dolphin Interconnect Solutions Inc*., Oslo, Norway, 1994.

[Dolphin95] Dolphin, “4-way SCI Cluster Switch”, *Dolphin Interconnect Solutions Inc*., Oslo, Norway, 1995.

[Dolphin97] Dolphin, “Link Controller LC-2 Specification”, *Dolphin Interconnect Solutions Inc*., Oslo, Norway, 1997.

[Enge96] D.R. Engebretsen, D.M. Kuchta, R.C. Booth, J.D. Crow, W.G. Nation, “Parallel Fiber-Optic SCI Links”, *IEEE Micro*, Vol. 16, No 1, Feb. 1996, pp. 20-26.

[Gillet96] R. B. Gillet, “Memory Channel Network for PCI”, *IEEE Micro*, Vol. 16, No 1, Feb. 1996, pp. 12-19.

[Goke73] L. R. Goke, G. J. Lipovski, "Banyan Networks for Partitioning Multiprocessor Systems", *Proc. of the 1st Annual Symposium on Comp. Architecture*, 1973, pp. 21-28.

[Gustav94] D. Gustavson, Q. Li, “Local Area Multiprocessor: the Scalable Coherent Interface”, in *Defining the Global Information Infrastructure*, S.F. Lundstrom ed., SPIE Press, vol. 56, pp. 141-160, 1994.

[IEEE92] “Standard for Scalablew Coherent Interface (SCI)”, IEEE std. 1596-1992.

[IEEEP] “Standard for Heterogenous Interconnect (HIC), IEEE P1355 proposed standard.

[Krist94] E. H. Kristiansen, G. Horn, S. Linge, “Switches for point-to-point links using OMI/HIC technology,”, in *Int. Data Acquisition Conference on Event Buildung and Data Readout*, Fermi National Accelator Laboratory, Batavia, Illinois, USA, Okt. 1994.

[Mercury95] Mercury, “RACEway”, *Mercury Computer Systems Inc*., USA, 1995, (http://www.mc.com/mtb2pci/mtb2_main.html).

[Omang96] K. Omang, B. Parady, “Performance of Low-Cost UltraSparc Multiprocessors connected by SCI”, Research Report No 219, University of OSLO, Norway, June 1996, http://www.ifi.uio.no/~sci.

[Scott96] S. Scott, “The GigaRing Channel”, *IEEE Micro*, Vol. 16, No 1, Feb. 1996, pp. 27-34.

[Siegel78] H. J. Siegel und S. D. Smith, "A Study of Multistage SIMD Interconnection Networks", *Fifth Annual Symposium on Computer Architecture*, April 1978, pp. 9-17.

[Wu80] C. I. Wu und T. Y. Feng, "On a Class of Multistage Interconnection Networks", *IEEE Transactions on Computers*, Vol. C-29, No. 8, August 1980, pp. 694-702.

[Wu942] “Applications of the Scalable Coherent Interface in Multistage Networks”, IEEE TENCON, Aug. 1994.

[Wu94] B. Wu, “SCI Switches”, *Int. Data Acquisition Conference on Event Buildung and Data Readout*, Fermi National Accelator Laboratory, Illinois, USA, Okt. 1994.