6. SCI Rings, Switches, and Networks for Data Acquisition Systems

Harald Richter[1], Richard Kleber[1], Matthias Ohlenroth[2]

[1] Institut für Informatik, Technische Universität München
D–80290 München, Germany
email: richterh@informatik.tu-muenchen.de
[2] Fakultät für Informatik, Technische Universität Chemnitz-Zwickau
D-09111 Chemnitz, Germany

6.1 Introduction

In plasma-physical fusion devices, the ionized hydrogen isotopes Deuterium and Tritium are fusing to helium ions, provided that they can be kept long enough and dense enough at very high temperatures (typically $\gg 1$ million degrees), thereby delivering energy according to the equation $E = (m_{Deuterium/Tritium} - m_{Helium}) \times c^2$. To control the plasma confinement and to get information about its physical behavior, a high-speed and high-volume data acquisition system is needed for the on-line and off-line monitoring and evaluation of measured plasma and fusion device data. State-of-the-art experiments produce approx. 100 MByte of measurement values during a 10 second experimental period. Future plasma devices will deliver one to two orders of magnitude more data in the same time interval, while additionally operating continuously in a steady-state mode. This imposes high requirements on the real-time behavior, i.e., the transmission latency, as well as on the bandwidth of the underlying communication network that is part of the data acquisition system. A low and guaranteed latency is very important for the closed-loop feedback systems of the fusion device that keep the burning plasma hot, stable, and away from physical material.

Potential candidates for the communication network of such a future data acquisition system are Fiber Distributed Data Interface (FDDI), Gigabit Ethernet, Asynchronous Transfer Mode (ATM), and Scalable Coherent Interface (SCI). Among them, SCI seems to be especially attractive because of its forward progress guarantee, prioritized bandwidth allocation and extremely low latency. For the same reasons, also other high-end physical experiments such as the "Large Hadron Collider" at CERN are considering SCI as the primary communication medium [3, 4, 5, 20, 30]. In industry, SCI gains more and more importance since large computer manufacturers are offering off-the-shelf SCI-based products for cluster computing [22, 7]. However, for data

[†] Partially reprinted from: H. Richter and M. Ohlenroth: Data acquisition with the SCINET, a scalable-coherent-interface network, *Fusion Engineering and Design*, vol. 43, pp. 393–400, Copyright 1999, with permission from Elsevier Science

H. Hellwagner, A. Reinefeld (Eds.): SCI'99, LNCS 1734, pp. 125-149, 1999

acquisition purposes commercially available SCI products are very rare, and for high-end applications no off-the-shelf solutions are provided since this market segment is too small. A lot of research remains to be done in this field [23, 25, 31, 21, 19].

One research project is SCINET, the goal of which is to investigate the applicability of SCI networks for data acquisition systems in large-scale fusion reactor experiments. SCINET investigates how data acquisition computers can be efficiently connected to each other and to sensors and actuators by means of SCI, which topological structures large-scale networks should have, and the bandwidth and latency that can be expected. Therefore, this chapter first presents the results of SCI test beds that were established as a sample SCI-based data acquisition system. The test beds demonstrate up to 45 MByte/s of throughput and $< 5\mu$s latency for end-to-end data transfers. Second, it is shown how commercial SCI switches can be used more efficiently resulting in more than sevenfold higher throughput and half the latency at the same costs. Third, SCI-based Banyan topologies are proposed which are highly efficient for multi-port switches in parallel computers, clusters of workstations, and local area multiprocessors. The networks have up to four times the performance in terms of throughput and latency, as compared to a conventional SCI-based multistage network, while requiring only one fourth of its costs. The prerequisite for the improvements is that some data locality is present in the traffic patterns between senders and receivers. All results were achieved by means of our SCINET simulator.

The chapter is organized as follows. In Section 6.2, the basic requirements of a data acquisition system for a fusion reactor experiment are explained. In Section 6.3, the two SCI test beds used for benchmarking are described. Section 6.4 shows the performance results of the test beds which indicate the principal suitability of SCI for data acquisition systems in plasma physics. Section 6.5 is devoted to SCI switches as the basic components of large-scale data acquisition systems. Section 6.6 explains and validates (by means of simulation) that SCI switches can be used more efficiently. Section 6.7 deals with multistage networks; two new cost-efficient Banyan topologies are proposed that can replace conventional solutions. In Section 6.8, simulation results for the new topologies are given. The chapter concludes with a summary in Section 6.9.

6.2 SCI-based Data Acquisition Systems

Data acquisition is a real-time task with reaction times determined by the sampling rates of the probing devices that collect the measurement values. Each data acquisition system can be evaluated by six key parameters:

– The probes' data rates in terms of samples per second. This value determines mainly the technology that has to be employed. For a fusion reactor,

data normally need not to be sampled with frequencies higher than 100 MSamples/s per sensor.
- The total number of sensors that have to be read out, maintained, and operated over a longer period of time (normally 1-2 decades).
- The amount of data collected by the system. For today's data acquisition systems, this value lies in the range of 10-100 MByte/s. Future systems will deliver 1-2 orders of magnitude more data which imposes high requirements on the database and data mining systems.
- The system's delay time between sampling and remote storing of the measurement values. This is a crucial factor if any additional feed-forward or feedback control system has to make use of the measured values because it determines the time constant, i.e., the reaction delay of the control system.
- The allowable data loss rate of the system. This rate determines the degree of redundancy that has to be built into the system. Since fusion reactor experiments are mission-critical, high reliability is required. Thus, the data loss rate has to be kept significantly lower than in telecommunication systems, for instance.
- The scalability of the system. Large-scale experiments take about 10 years to be built and are operated another 10-15 years. This means that their data acquisition system has to be flexible enough to be expanded over the years.

In this chapter, throughput and latency of a sample SCI-based data acquisition system are investigated. Data losses are partly evaluated by registering the error rates that occur on the transmission lines. Other sources of data loss, for instance those due to buffer overflows, are not considered. The management of the data retrieval and the overall system's scalability are outside the scope of this contribution.

SCI is believed to be a proper technology for high-end data acquisition systems since it has, by construction, guaranteed data delivery, high throughput, low latency and scalability. In the framework of the SCINET project, a sample acquisition system was implemented to validate this. Throughput and latency of data transfers from memory to memory were measured by means of test beds.

6.3 SCINET Test Beds

Two different SCI test beds were set up to enable comparisons of, and to achieve a higher confidence in, the obtained results. Each system represents an SCI data transmission ring, comprising two PCs that are connected by SCI interfaces. For both test systems, commercially available SCI cards from Dolphin Interconnect Solutions [11] are used, together with appropriate copper cabling for the transmission lines.

In the first system, two 200 MHz Pentium Pro PCs with the 440FX PCI chip set are employed (GA- 686DX mother board), running the Linux 2.0.30 operating system. The Dolphin boards are controlled by our own device driver, and also the benchmarks that measure the throughput and latency are self-developed. The second system comprises two 100 MHz Pentium PCs, Windows NT 4.0, and the standard FX chip set. Here, device driver and benchmark software were delivered by Dolphin. In both cases, the responseless DMOVE64 transaction of SCI was used for all data transfers.

Each interface card is based on Dolphin's 200 MByte/s Link Controller (LC) chip [10, 12] that implements the physical layer of SCI, and a PCI bridge that converts the PC's PCI bus protocol into SCI packets. On the card, the PCI bridge and the LC chip are connected internally via a proprietary high-performance bus, the so called B-Link [8]. The LC conforms to the IEEE interface specification, with the exception that the attached user device is a PC that is connected via the PCI/B-Link bridge to the SCI interface. The bridge consists mainly of a PCI master/slave building block, a DMA engine and 8 buffers for read and write transactions, respectively. Further components are an address translation cache for the mapping of PCI to SCI addresses, configuration, control, and status registers, and an interrupt mechanism. The block diagram of the PCI/B-Link bridge is shown in Figure 6.1. The adapter card is further described in Chapter 3.

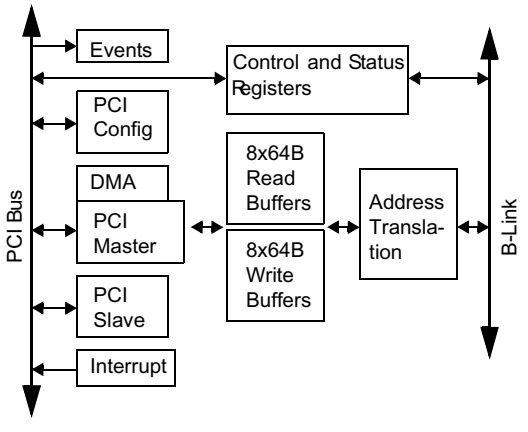

Fig. 6.1. Dolphin's PCI/B-Link bridge

The bridge is capable of either splitting its buffers in order to support up to 8 external PCI masters, for read and write, respectively, or of combining the buffers to allow for eightfold throughput for a single master. The prerequisite for buffer combining is that PCI data that has to be transferred to the SCI link

is located on contiguous addresses, and that the control and status registers of the bridge are properly set.

6.4 Measurement Results

The throughput results shown in Figure 6.2 were obtained for both, remote read and write transactions between the memories of the Pentium Pro test bed (system 1). Data transfer took place exclusively between user address spaces, thus the values represent the real end-to-end data rate. As one can see, the SCI ring already reaches its maximum throughput at short block lengths of 256 bytes; at 64 bytes, about half of that can be achieved. Remote read is significantly slower than remote write, by a factor of 3.5. This stems from the fact that, for the requester, a remote write transaction is already finished as soon as its data are stored in the transmit buffer of the PCI/B-Link bridge. From there on, reliable delivery is guaranteed. In contrast, a remote read always needs to wait until the requested data is obtained because of the semantics of the read transaction. Obviously, to achieve maximum throughput only remote writes should be employed. This means for the data acquisition system that a "push" strategy should be preferred to "pull"-style operation. That is, the sensors should on their own write their measured data to the final destinations after they have received a data-sampling trigger from those locations, and the remote computers should not read out the probes. For that reason, only remote write transactions are further investigated in the following.

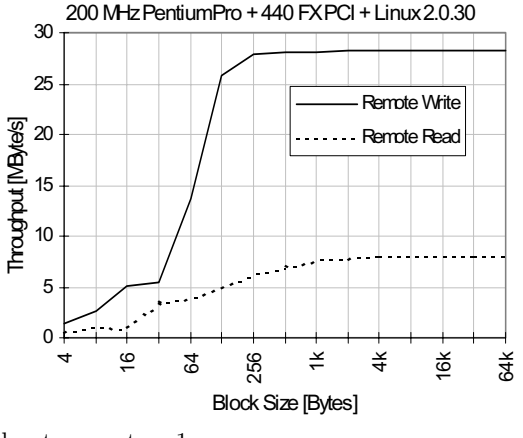

Fig. 6.2. Throughput on system 1

When one considers the elapsed times of the data transfers vs. the block size (Figure 6.3), it becomes clear that the setup times to initiate a transfer

are quite small: $< 10\mu$s for both, read and write. That would allow a very small reaction time of an SCI-based control system. The elapsed times for remote read (RR) are higher, compared to remote write (RW) due to its slower transfer rate.

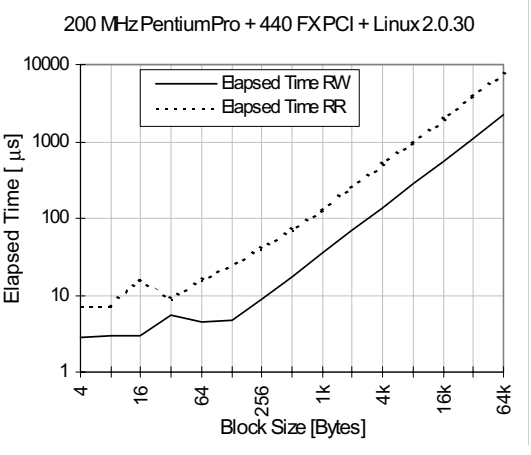

Fig. 6.3. Latencies of remote read and write on system 1

The Pentium-based test bed with the Dolphin software (system 2) shows a throughput behavior as depicted in Figure 6.4 for comparison; here, only 12 MByte/s can be achieved. Additionally, Figure 6.4 shows how the transfer rate behaves in case that, on top of SCI's hardware error detection and correction, a software error check with optional retry of the last sent block is applied. For both cases (HW correction only and HW+SW correction), the maximum throughput turns out to be the same, but for the latter it can be achieved only for large block sizes of > 64 kByte. Without software overhead, Dolphin's solution reaches its maximum transfer rate at 64 bytes already. The latencies for larger block sizes (depicted in Figure 6.5) are in compliance with the measured throughput. For small blocks, the latency becomes as low as 2-3 μs. Of course, more time is needed (11-12 μs) if additional software error correction is used.

Dolphin's device driver allows the combining of the PCI/B-Link bridge write buffers. As one can see from Figure 6.6 and Figure 6.7, the throughput scales nearly linearly with 2 and 4 buffers combined. Only the minimum required buffer size for full transmission speed is doubled each time.

The maximum achievable data transfer rate is 45 MByte/s since with 8 combined buffers saturation effects are showing up (Figure 6.8). Since each write buffer has a capacity of 64 bytes, the combining of 2 or 4 buffers requires that at least 128 bytes or 256 bytes, respectively, that are aligned to 64-byte

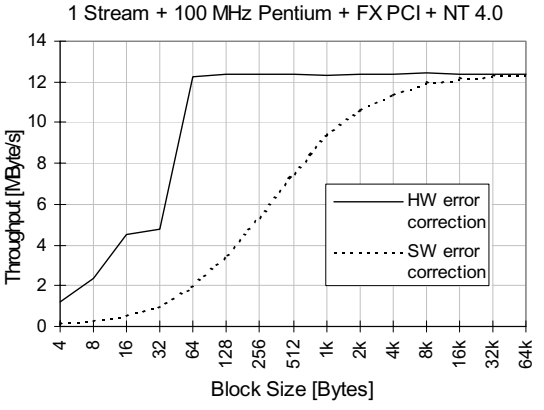

Fig. 6.4. Throughput of 1-stream remote write on system 2

.

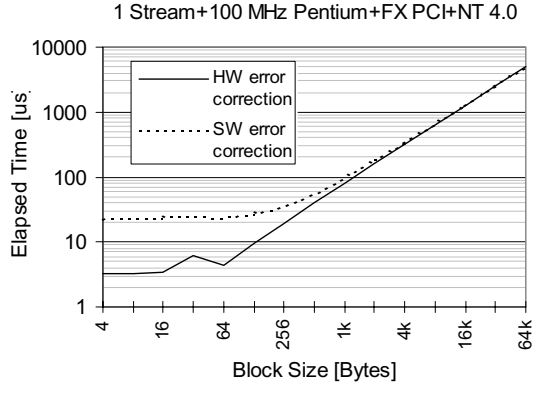

Fig. 6.5. Latencies of 1-stream remote write on system 2

.

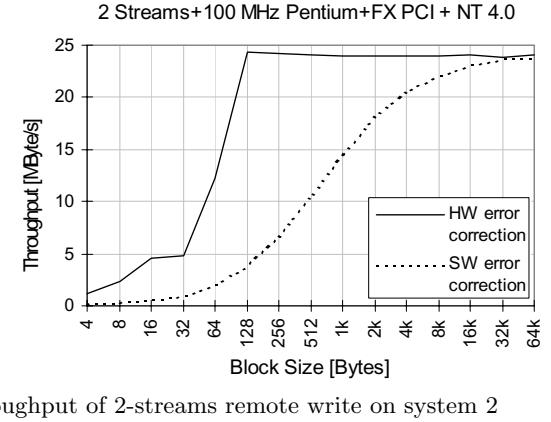

Fig. 6.6. Throughput of 2-streams remote write on system 2

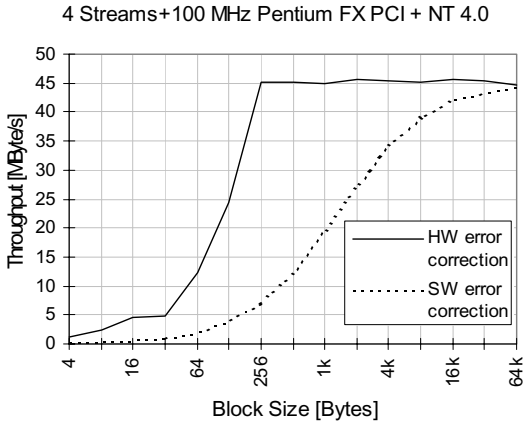

Fig. 6.7. Throughput of 4-streams remote write on system 2

boundaries in the PCI address space, are available for transfer. At the latter block size, maximum throughput is achieved. In the case of combining 8 buffers, bursts of 512 bytes and larger suffer to be transferred at a reduced speed of only 34 MByte/s. Either the PCI bus, the slow Pentium CPU or some other limiting factor induced this saturation effect that causes the significant performance drop.

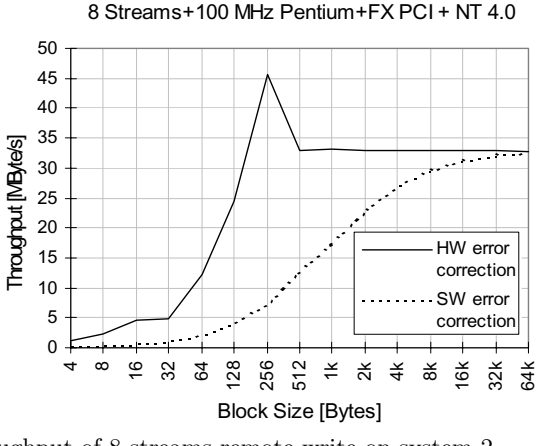

Fig. 6.8. Throughput of 8-streams remote write on system 2

The latency times for 2, 4, and 8 combined streams are given in Figures 6.9, 6.10, and 6.11, respectively.

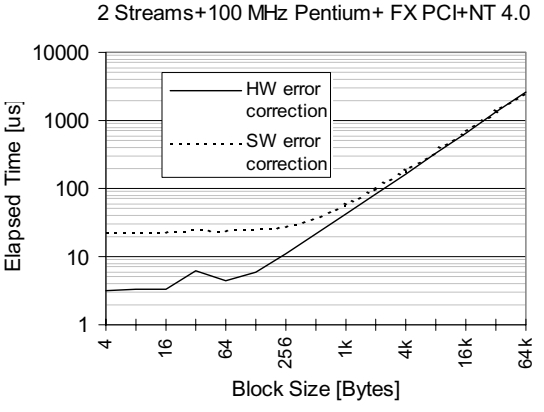

Fig. 6.9. Latency of 2-streams remote write on system 2

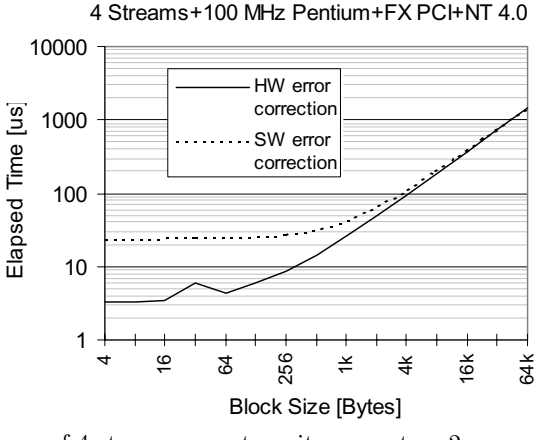

Fig. 6.10. Latency of 4-streams remote write on system 2

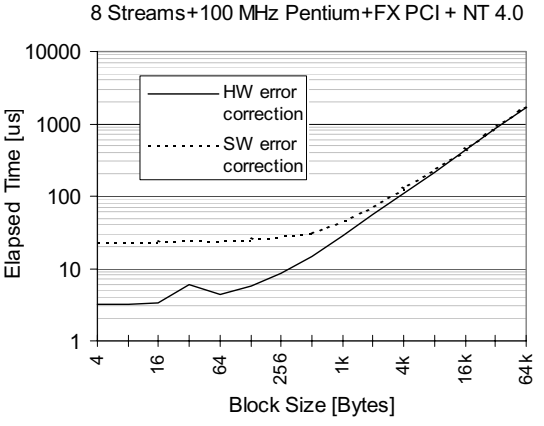

Fig. 6.11. Latency of 8-streams remote write on system 2

6.5 SCI Switches

In addition to commercial SCI products for data transfer between PCs and workstations, SCI switches [9] were investigated, since they are the prerequisite for large-scale data acquisition systems. Switches allow to connect two or more SCI rings, thereby forming a static or dynamic SCI network. Each SCI network 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 an 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 6.12). An SCI switch differs in various respects from conventional switches. First, each SCI switch is part of 2 to 4 rings on which data are unidirectionally transferred. Second, there exists a port-internal bypass FIFO connecting the in and out terminals of each port to allow a very fast bypass (< 50 ns). Third, in each port two separate buffers for SCI requests and responses are available preventing deadlocks caused by cyclic waiting on resources (Figure 6.13).

In the following, we consider switches according to Dolphin's implementation [9]. The transmit and receive buffers of the ports of such a switch are connected to a high-speed packet bus called B-Link [8], which has a transmission rate of 600 MByte/s. Inside the switch, the B-Link connects 4 ports, each of which has a data rate of 500 MByte/s per direction. By this, a high-speed SCI switch is established. Between any pair of ports, the maximum port rate of 500 MByte/s can be achieved for unidirectional transfers (either read or write) as long as the remaining other pair of ports produces no more than 100 MByte/s of traffic. If both pairs simultaneously operate in full duplex

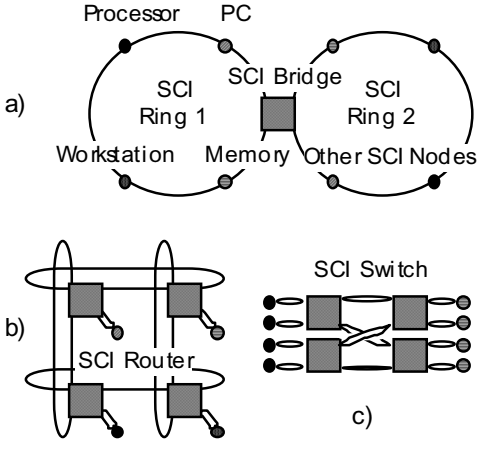

Fig. 6.12. SCI switches can serve as bridges (a), routers (b), and building blocks for multistage networks (c)

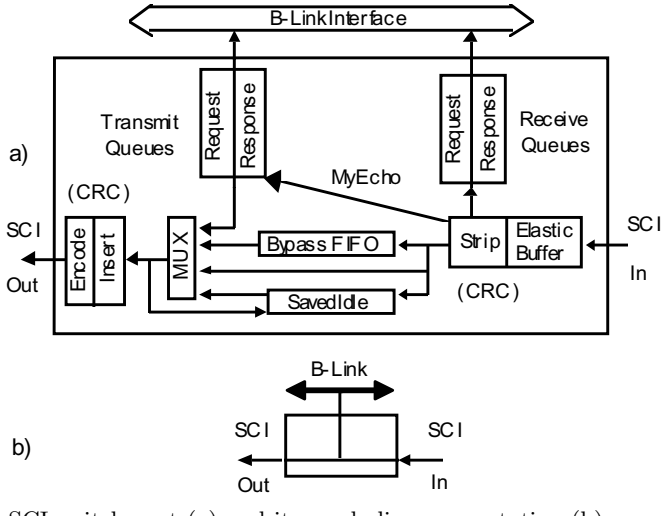

Fig. 6.13. SCI switch port (a) and its symbolic representation (b)

mode, the individual port rate per direction is reduced to 150 MByte/s due to the B-Link's bandwidth limitations.

For illustration, Figure 6.14 depicts a block diagram of a 4-port SCI switch as well as its equivalent representation that will be used later in this chapter.

.

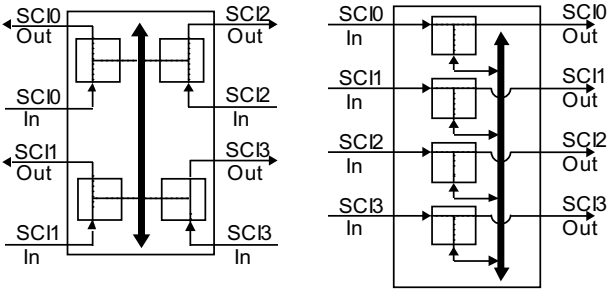

Fig. 6.14. Two equivalent representations of a 4-port SCI switch

6.6 Efficient Use of SCI Switches

Let the throughput T of a switch be the sum of the ports' throughputs. For a subsequent comparison with SCI systems, Figure 6.15 shows two simple multiprocessors (UMA and NUMA variants) that employ conventional switches.

In Figure 6.15(a), 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 6.15(b), 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 ($B_{tr} \geq T'_{con} \geq T_{con}$).

With SCI, it is possible to push T'_{con} above the bandwidth limit B_{tr} by using the ports' bypass FIFOs for additional data transfers. This special switch usage will be explained in the following. In Figure 6.16(a), the UMA architecture of Figure 6.15(a) is upgraded to an SCI switch, and the bidirectional transmission lines are replaced by SCI ringlets. Now, 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 [20, 30]. In Figure 6.16(b) however, the processor and memory nodes are coupled differently: the SCI ringlets are replaced by long rings connecting a sender, a switch, and a receiver in one instead of two rings so that no B-Link is in between.

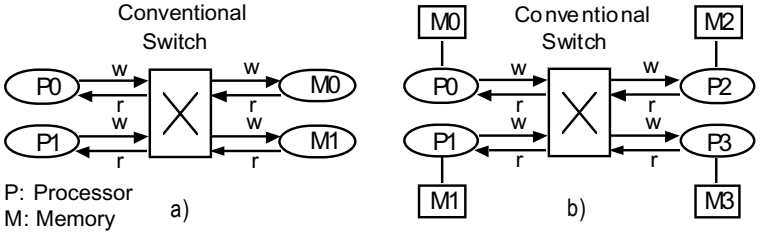

Fig. 6.15. Simple UMA (a) and NUMA (b) multiprocessors based on conventional switches

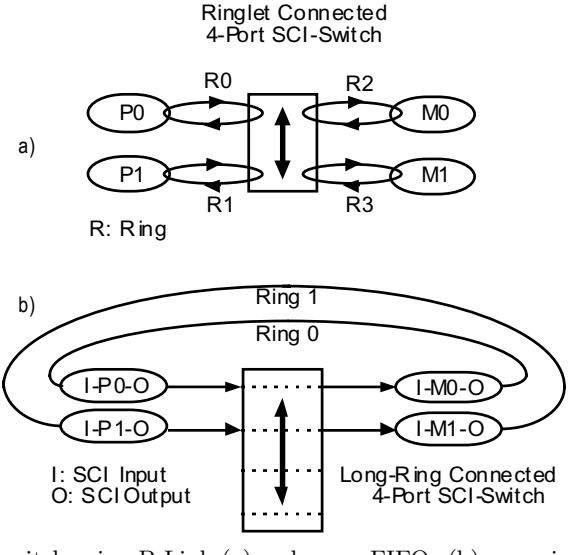

Fig. 6.16. SCI switch using B-Link (a) or bypass FIFOs (b) as main data paths

Here, we obtain throughput T'_{SCI} which can become larger than B_{tr}, provided that some fraction of the data can stay on the ring where it originated. 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.

The prerequisite that T'_{SCI} exceeds B_{tr} is that the communication patterns between senders and receivers exhibit some data locality. Data locality is common to most parallel applications; if not, it can be explicitly forced by proper allocation of tasks and data structures to computing nodes. In the example depicted in Figure 6.17(a), this means that processor $P_i (i = 0, 1, 2, 3)$ mainly communicates with memory M_i, so that data packets can stay on the rings where they originated and can travel to the destinations through the bypass FIFOs.

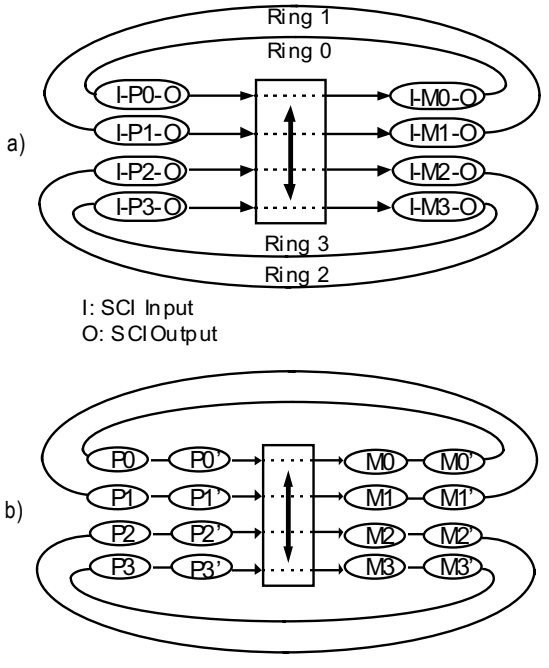

Fig. 6.17. Flexibility in the number of nodes by means of long rings

In addition, the latency is reduced since in SCI the intra-ring communication is faster than the inter-ring communication. Furthermore, from Figure 6.16(a) to Figure 6.16(b) 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

is expected to increase. For larger switch sizes than 4 the same improvement would be achieved.

Finally, as shown in Figure 6.17(a) and (b), more flexibility with respect to the number of attachable processors and memories can be obtained by using long rings that pass through SCI nodes and switches. In the example of Figure 6.17(a), twice as many processors and memories are coupled to the same 4-port switch without degradation in performance, as compared to the ringlet configuration. In the example of Figure 6.17(b), the number of connectable devices is again doubled. However, in the latter case, the maximum data rate per processor may be halved. In Section 6.8, the predicted performance improvements are quantified.

6.7 Multistage SCI Networks

In this section, the long-ring connection technique is applied to multistage networks comprising 2-port or 4-port SCI switches, in order to construct more efficient networks. Generally, the most cost-efficient multistage networks are of the Banyan type [15], 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 [28].

In Figure 6.18(a), the standard implementation of an SCI-based Baseline network according to [30] is shown: nodes and switch ports of adjacent stages are connected by SCI ringlets. A functional equivalent but bypass FIFO-based solution that consists of large toroidal rings is shown in Figure 6.18(b).

In this example, the switch complexity is reduced from 4-port to 2-port switches. The minimum latency of data to travel from an input to an output of a network of size N has decreased from $L = \alpha \log_2 N$ to $L' = \beta \log_2 N$. The factor α denotes the time for a packet to cross a switch (usually some μs), while β is the time to travel through a bypass FIFO (some tens of ns). Obviously, two orders of magnitude in latency decrease can be expected by employing long rings instead of ringlets, while halving the switch costs at the same time.

A disadvantage of the Baseline network of Figure 6.18(b) is that an additional permutation wiring is required to connect the memory-out links a–h with the corresponding processor-in links to obtain closed rings. Fortunately, by virtue of their topological structure, two of the known Banyans allow a one-to-one connection from outputs to inputs. These topologies are the Omega and the Generalized Cube networks. (Their "mirror images", the Flip network and the Indirect Binary n-Cube, have the same property.) Therefore, we propose that SCI networks which have bypass FIFOs as their main data paths should be built according to one of these 4 topologies. In the following,

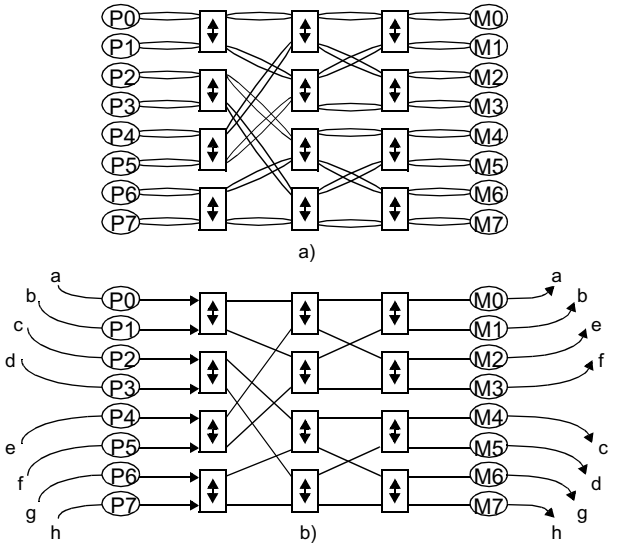

Fig. 6.18. Baseline networks with B-Links (a) or bypass FIFOs (b) as main paths

the networks without permuted wiring from output to input are termed *first-grade optimized*. An example of such a network is given in Figure 6.19.

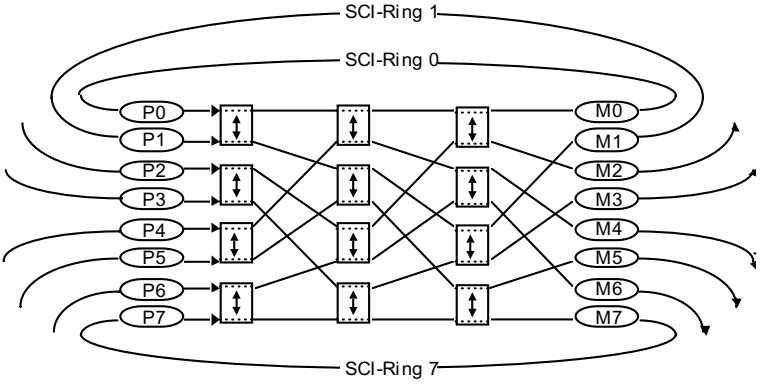

Fig. 6.19. First-grade optimized SCI-based Omega network

First-grade optimized SCI Banyans can be further improved by using s-port switches with $s > 2$, which results in additional improvements by a factor of $\log_2 N/log_s N$ in terms of costs and latency. For example, when $s = 4$ and a network size of 16×16 is assumed, then 32 ports are necessary to build up all network switches, while the equivalent first-grade optimized network

with $s = 2$ needs twice as many, i.e. 64 ports. Since the port count is the dominant cost factor of a network, the prize is approx. halved for $s = 4$. A ringlet network of the same size and $s = 2$ would require 128 ports.

If $s > 2$, we call such a structure a *second-grade optimized* network. An example of a second-grade optimized network is depicted in Figure 6.20. In Section 6.8, the performance of first-grade and second-grade optimized networks will be compared.

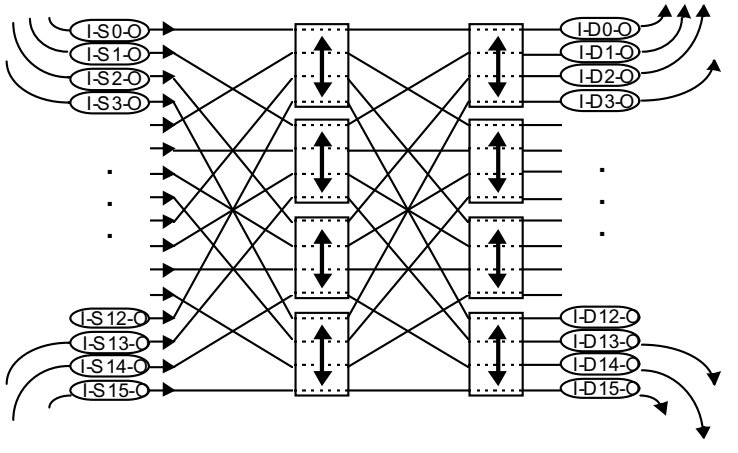

Fig. 6.20. Second-grade optimized SCI-based Omega network

6.8 Simulation Results

A first suite of simulations was conducted to evaluate the performance of a single 4-port SCI switch, which uses ringlets to connect processors and memories, and to compare that solution with a switch based on long rings (Figure 6.16). The performance metrics are throughput, latency, and packet losses. To pinpoint the performance discrepancy between both concepts, 100% data locality was chosen, i.e. processor P_i communicates exclusively with memory $M_i, i = 0, 1$. In practice, the locality will be lower, but as long as there is some data locality, a performance improvement will be visible.

For all simulations, the DMOVE64 SCI command was chosen, and all processors are configured to simultaneously send DMOVE64 packets at the same rate. The input data rate to the switch was decided to be deterministic, no random traffic is applied.

In the following graphs, the achieved data throughput of the switch (net output rate) versus the generated input traffic (gross input rate) are shown.

The input is varied from 0 to 500 MByte/s per processor, which is also the maximum ring speed, to study the input/output behavior of the switch. The data packets carry 64 bytes of payload with an overhead of 16 bytes for header and trailer. Together with additional 4 bytes for idle symbols, the ratio of payload length to raw length is 64/84. The memories are assumed to have an access time of 40 ns for each block of 64 bytes. The link delays between processors, switch, and memories are set to be 1 ns each. The remaining 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 modeled to have input and output buffer space for 4 request and response packets each. By this parameter set, the simulations are compliant with the latest SCI Link Controller (LC-2) of Dolphin [12].

The achieved throughput rates of a ringlet and a long-ring connected 4-port switch are shown in Figure 6.21. With ringlets, the switch already saturates at 250 MByte/s raw input rate, delivering 176 MByte/s output payload. At the same input rate, the packet losses become significant and eventually reach a value of 585 MByte/s at 1 GByte/s gross input rate. A packet loss occurs each time a new packet is generated that cannot be injected into the ring by a sender's SCI interface; this happens when the ring is still occupied by transferring previous packets and the transmit buffer of the interface is full. Because of the constant rate with which data are issued by the processors, the ring has to accept packets in real time which is only possible up to a certain speed. Above that limit, packets are lost.

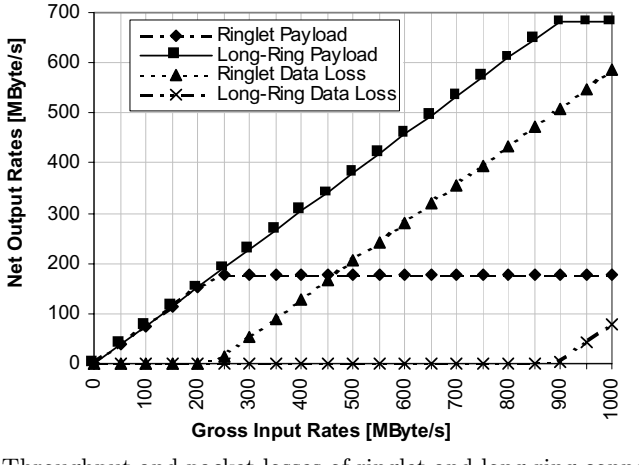

Fig. 6.21. Throughput and packet losses of ringlet and long-ring connected 4-port SCI switches with two senders and receivers

The latency behavior of a ringlet-connected 4-port switch is depicted in Figure 6.22. The time from initiating a DMOVE64 packet to storing it in its destination shows a value of 2344 ns as long as the latency saturation point of 200 MByte/s is not reached. The latency increases and becomes non-deterministic above that point, assuming values between 7362 ns (minimum) and 12797 ns (maximum).

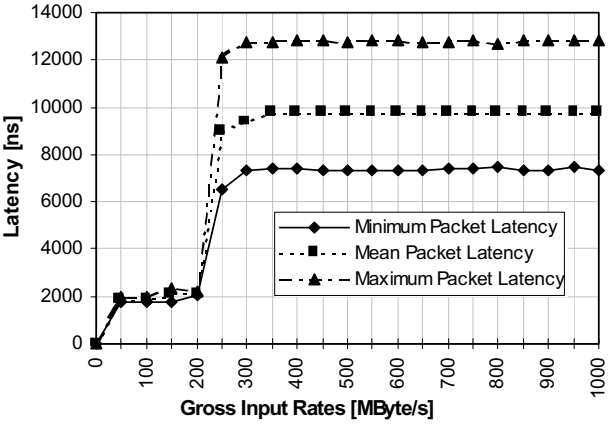

Fig. 6.22. Latency of ringlet-connected 4-port SCI switch with two senders and receivers

The long-ring connected switch coupling two senders and receivers behaves much better: it saturates at 900 MByte/s gross input rate with 682 MByte/s output payload, and at 1 GByte/s input rate it has 80 MByte/s packet losses. Below the saturation point, a latency of 1127 ns can be expected. Compared to the ringlet-case, the throughput has increased by a factor of 3.9 while the latency has decreased by 52%. However, above the saturation point, latency not only becomes non-deterministic but it also jumps by two orders of magnitude, varying between 75 μs and 301 μs (Figure 6.23). Obviously, for 100% data locality the long-ring connected switch behaves much better than a conventionally ringlet-coupled one, but it should not be overloaded.

If all 4 ports of the switch are coupled with long rings to 4 processors and memories, as depicted in Figure 6.17(a), a linear performance improvement compared to 2 processors and memories is achieved. As shown in Figure 6.24, the switch saturates at 1800 MByte/s gross input rate with 1365 MByte/s output payload, and it has 160 MByte/s packet losses at 2 GByte/s. With 1127 ns, the latency below the saturation point is identical to the case of 2 processors and memories. Identical latencies are also obtained above the saturation point.

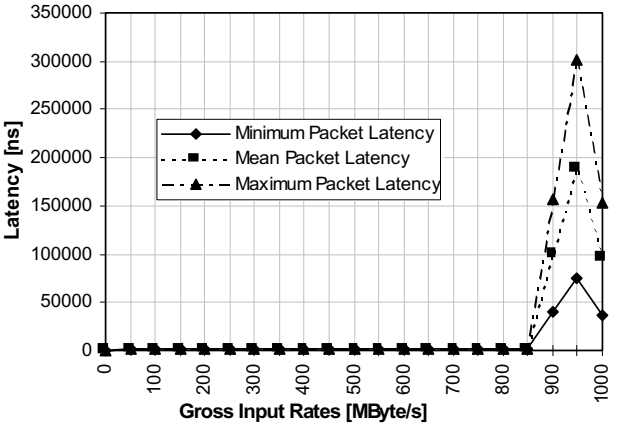

Fig. 6.23. Latency of long-ring connected 4-port SCI switch with two senders and receivers

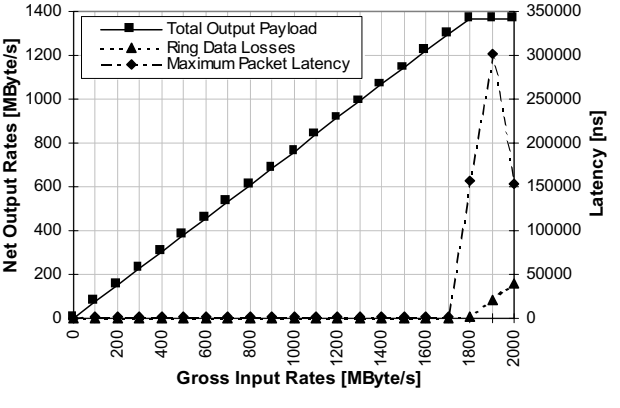

Fig. 6.24. Performance of 4-port switch with 4 processors and memories

This means that with the same switch as in the ringlet-connected case, a 7.8-fold throughput improvement can be achieved, provided that 100% data locality is present. Latency drops to one half if the switch is not overloaded. Other simulations show that ringlet and long-ring switches will be identical in performance if no data locality is present. This means that for operations below the saturation limit the long-ring coupled switch can always be preferred.

The second series of simulation experiments that was performed with the SCINET tool compares a 16×16 first-grade optimized Omega network with a conventional ringlet-based network of the same size and topology. Furthermore, first-grade and second-grade optimized networks are compared. Again, 100% data locality is assumed to demonstrate the upper limit of the performance improvements.

The behavior of the 16×16 ringlet-based SCI network is depicted in Figure 6.25. The network saturates at 2 GByte/s gross input rate with 1412 MByte/s output rate. At that point, the network has 112 MByte/s packet losses that increase up to 4684 MByte/s at 8 GByte/s input rate. The maximum latency is 6670 ns below saturation and jumps up to 20056 ns above saturation. Latency saturation occurs earlier than throughput saturation, at 1600 MByte/s.

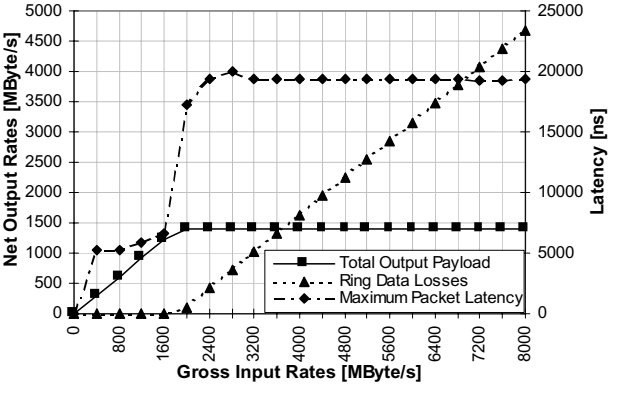

Fig. 6.25. Performance of a ringlet-connected 16×16 Omega network with 100% data locality

The performance of the first-grade optimized Omega network is shown in Figure 6.26. It has 5333 MByte/s output rate at 7200 MByte/s gross input rate, a 3.8-fold performance improvement over the ringlet network. The latency is much better as well: below the latency saturation point of 6800 MByte/s, we have 1771 ns which are 27% of the ringlet latency. Above

that point, a maximum of 5067 ns is reached which is roughly one fourth of the ringlet's latency.

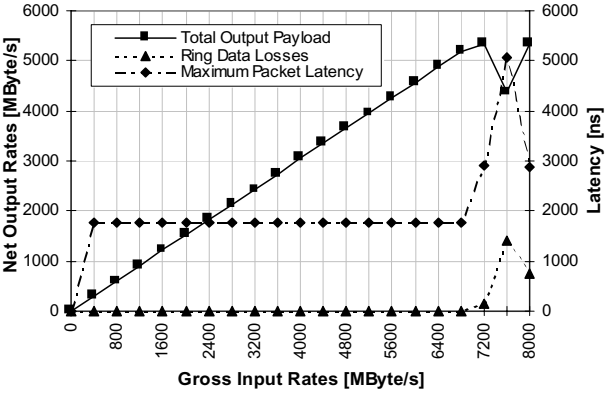

Fig. 6.26. Performance of first-grade optimized 16×16 Omega network with 100% data locality

This means that the first-grade optimized network of size 16×16 has approx. the 4-fold throughput and one fourth of the latency of a conventional SCI-based Omega network, at only half of its costs. The results can be extrapolated to sizes > 16.

The simulation results for the second-grade optimized network are shown in Figure 6.27. It can be seen that throughput saturation also occurs at 7200 MByte/s gross input rate, but with 5456 MByte/s output rate it delivers a slightly higher throughput than the corresponding first-grade optimized network. Also the latency before saturation is better, 1525 ns. However, after saturation latency jumps by two orders of magnitude and reaches 134 μs. Note that the fully connected 4-port switch shows the same behavior.

The second-grade optimized network delivers slightly better performance in all respects compared to the first-grade optimized network as long as it is not overloaded, but at only half the costs. Both have roughly four times the performance in terms of throughput and latency of the ringlet-coupled network of the same type and size, while requiring only one half or one fourth of the costs, respectively.

6.9 Summary and Conclusions

In this chapter, the feasibility of an SCI-based data acquisition system was demonstrated by the achieved throughputs of 45 MByte/s and latencies of

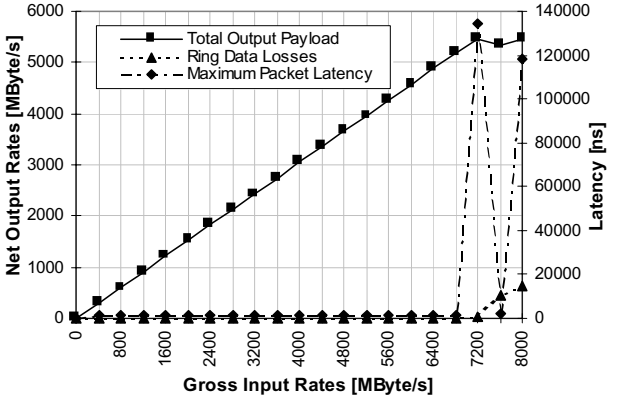

Fig. 6.27. Performance of second-grade optimized 16 × 16 Omega network with 100% data locality

$< 10\mu$s. The sample DAQ system was based on PC test beds where a decent performance was also measured for the case of additional software error checking and correcting. In the future, the test beds will be upgraded to fiber links and an SCI switch will be included to study its influence in practice. Additionally, the interactions of the transmission system and the higher software levels have to be analyzed, and a programming model must be devised that is suitable for the physicists' needs.

Furthermore, we have shown by means of simulation how commercial SCI switches, which are the basic blocks of *large-scale* data acquisition systems can be used more efficiently. The principle is to use the bypass FIFO that is part of every SCI port, for establishing long SCI rings comprising a pair of nodes where one node is a transmitter and the other node is a receiver. Then, packets can be redirected from the switch-internal bus (i.e. the B-Link) which is a bottleneck, to the port's bypass FIFO. The prerequisite for redirection is that there exists data locality in the traffic pattern between the sender and the receiver residing on the same ring. For 100% data locality, up to a 7.8-fold throughput improvement is achievable compared to a ringlet-connected switch. Latency is reduced by a factor of 2 and the number of lost packets by a factor of 7. Additionally, by grouping pairs of sender and receiver nodes into each ring, more nodes can be connected. However, latency increases by two orders of magnitude above the saturation point of the ring.

A new network type, called first-grade optimized network, was proposed. It is based on long rings passing through all switch stages of a Banyan network. Each long ring replaces a number of ringlets connecting neighboring nodes in the conventional network. The new network type improves multistage Banyan networks that are based on SCI ringlets. With long rings and data locality, packets can stay on the ring where they originated, thus remo-

ving traffic from each switch-internal bus (B-Link). For 100% data locality, a first-grade optimized network has four times the performance in terms of throughput and latency of a conventional SCI network of same size and type while exhibiting one half of its costs.

Finally, so-called second-grade optimized networks were suggested, further improving first-grade ones. They use long rings as well as intra-stage wirings with permutation functions on a number base higher than two. With a permutation base of four, for instance, one half of the costs of a first-grade optimized network can be achieved, even with slightly better performance in throughput and latency. The performance improvement is proportional to the number of network ports. All results were obtained by the newly developed SCINET simulation program.

References

1. T. E. Anderson, D. E. Culler, D. A. Patterson, A Case for NOW (Networks of Workstations). *IEEE Micro*, pages 54–64, Feb. 1995.
2. N. Boden, D. Cohen, R. Felderman, A. Kulawik, C. Seitz, J. Seizovic, W.-K. Su. Myrinet: A Gigabit-per-Second Local Area Network. *IEEE Micro*, pages 29–35, Feb. 1995.
3. A. Bogaerts, R. Divia, H. Müller, J. Renardy. SCI-based Data Acquisition Architectures. *IEEE Transactions on Nuclear Science*, Vol. 39, No. 2, Apr. 1992.
4. A. Bogaerts, R. Keyser, G. Mugnai, H. Müller, P. Werner, B. Wu, B. Skaali, J. Ferrer-Prietro. SCI Data Acquisition Systems: Doing More with Less. *CHEP'94*, San Francisco, April 1994
5. A. Bogaerts et al. *RD 24 Status Report: Application of the Scalable Coherent Interface to Data Acquisition at LHC*. Oct. 1996. http://nicewww.cern.ch/~hmuller/~HMULLER/docs/report96.pdf.
6. CACI Products Company. *Modsim II, The Language for Object Oriented Programming*. Reference Manual, La Jolla, California, 1995.
7. R. Clark, K. Alnes. An SCI Interconnect Chipset and Adapter. *Proc. Hot Interconnects Symposium IV*, Stanford University, Aug. 15-17, 1996.
8. Dolphin Interconnect Solutions. *A Backside Link (B-Link) for Scalable Coherent Interface (SCI) Nodes*. Dolphin Interconnect Solutions Inc., Oslo, Norway, 1994.
9. Dolphin Interconnect Solutions. *4-way SCI Cluster Switch*. Dolphin Interconnect Solutions Inc., Oslo, Norway, 1995.
10. Dolphin Interconnect Solutions. *Link Controller LC-1 Specification*. Dolphin Interconnect Solutions Inc., Oslo, Norway, 1995.
11. Dolphin Interconnect Solutions. *PCI/SCI Cluster Adapter Specification*. Dolphin Interconnect Solutions Inc., Oslo, Norway, 1996.
12. Dolphin Interconnect Solutions. *Link Controller LC-2 Specification*. Dolphin Interconnect Solutions Inc., Oslo, Norway, 1997.
13. D. R. Engebretsen, D. M. Kuchta, R. C. Booth, J. D. Crow, W. G. Nation. Parallel Fiber-Optic SCI Links. *IEEE Micro*, pages 20–26, Feb. 1996.
14. R. B. Gillett. Memory Channel Network for PCI. *IEEE Micro*, pages 12–19, Feb. 1996.
15. L. R. Goke, G. J. Lipovski. Banyan Networks for Partitioning Multiprocessor Systems. *Proc. 1st Int'l. Symposium on Computer Architecture*, pages 21–28, 1973.

16. D. B. Gustavson, Q. Li. Local Area Multiprocessor: the Scalable Coherent Interface. *Defining the Global Information Infrastructure*, S. F. Lundstrom (ed.), SPIE Press, Vol. 56, pp. 141–160, 1994.
17. Standard for Scalable Coherent Interface (SCI). *IEEE Std. 1596-1992*
18. Standard for Heterogeneous Interconnect (HIC). *IEEE P1355 Proposed Standard*
19. D. V. James. The Scalable Coherent Interface: Scaling to High-Performance Systems. *Proc. COMPCON Spring'94*, 1994.
20. E. H. Kristiansen, G. Horn, S. Linge. Switches for Point-to-Point Links Using OMI/HIC Technology. *Int. Data Acquisition Conference on Event Building and Data Readout*, Fermi National Accelerator Laboratory, Batavia, Illinois, USA, Oct. 1994.
21. M. Liebhart, A. Bogaerts, E. Brenner. A Study of an SCI Switch Fabric. *Proceedings IEEE MASCOTS'97*, Haifa, Israel, 1997.
22. K. Omang, B. Parady. Performance of Low-Cost UltraSparc Multiprocessors Connected by SCI. *Proceedings Communication Networks and Distributed Systems Modeling and Simulation (CNDS'97)*, Phoenix, Arizona, USA, Jan. 1997.
23. H. Richter, M. Liebhart. Performance Optimizations of Switched SCI-Rings. *Proceedings 11th Annual International Symposium on High Performance Computing Systems (HPCS'97)*, Winnipeg, Canada, July 1997.
24. H. Richter. *Interconnection Networks for Parallel and Distributed Systems* (in German). Spektrum Akademischer Verlag, Heidelberg, Germany, 1997.
25. S. Scott, J. Goodman, M. Vernon. Performance of the SCI Ring. *Proc. 19th Int'l. Symp. on Computer Architecture*. ACM Press 1992.
26. S. Scott. The GigaRing Channel. *IEEE Micro*, pages 27–34, Feb. 1996.
27. H. J. Siegel, S. D. Smith. A Study of Multistage SIMD Interconnection Networks. *Proc. 5th Int'l. Symposium on Computer Architecture*, pages 9–17, April 1978.
28. C. I. Wu, T. Y. Feng. On a Class of Multistage Interconnection Networks. *IEEE Transactions on Computers*, Vol. C-29, No. 8, pages 694–702, August 1980.
29. B. Wu. Applications of the Scalable Coherent Interface in Multistage Networks. *IEEE TENCON*, Aug. 1994.
30. B. Wu. SCI Switches. *Int'l. Data Acquisition Conference on Event Building and Data Readout*, Fermi National Accelerator Laboratory, Illinois, USA, Oct. 1994.
31. B. Wu, A. Bogaerts, B. Skaali. A Study of Switch Models for the Scalable Coherent Interface. *Proceedings of the Sixth IFIP WG6.3 Conference on Performance of Computer Networks*, Istanbul, 1995.