ELSEVIER

Fusion Engineering and Design 43 (1999) 393–400

Fusion Engineering and Design

Data acquisition with the SCINET, a scalable-coherent-interface network

Harald Richter [a,b,*], Matthias Ohlenroth [c]

[a] *Computer Science Division, Technical University of Munich, Lehrstuhl für Rechnertechnik und Rechnerorganisation, Arcisstr. 21, D-80290 Munich, Germany*

[b] *Max-Planck-Institute for Plasma Physics, EURATOM Association, Boltzmannstr. 2, D-85748 Garching, Germany*

[c] *Computer Science Division, Technical University of Chemnitz-Zwikau, Lehrstuhl für Rechnerarchitektur, Straße der Nationen 62, D-09111 Chemnitz, Germany*

Abstract

The goal of the SCINET project is to investigate the applicability of the scalable-coherent-interface (SCI) for data-acquisition systems in large-scale fusion-reactor experiments. SCI is a standardized, high-speed interconnect for peripheral devices, processors, memories, PCs and workstations that provides for a distributed shared memory with optional cache coherence between computing nodes. Up to 64 K SCI nodes can be closely coupled in one or more rings that are concatenated via SCI switches. In SCINET, it is investigated how data-acquisition computers can be efficiently connected with each other and with their sensors and actuators by means of SCI, what topological structure the network should have, and which bandwidth and latency can be expected. Test stands were established as a sample SCI-based data-acquisition system that showed up to 45 MB s^{-1} of throughput and < 5 μs latency for end-to-end data transfers. © 1999 Elsevier Science S.A. All rights reserved.

Keywords: Fusion-reactor experiments; Data-acquisition systems; Scalable coherent interface; SCI test-stand; Performance analysis

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 $\gg 1$ Mio. degree), thereby delivering energy due to $E = (m_{\text{deuterium/tritium}} - m_{\text{helium}}) \cdot c^2$. To control the plasma confinement and to get information about the physical behaviour, a high-speed and high-volume data-acquisition system is needed for the online and offline monitoring and evaluation of measured plasma and fusion-device data. State-of-the art experiments produce ≈ 100 MB of measurement values during a 10 s 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 behaviour, i.e. the

[*] Corresponding author. E-mail: richterh@informatik.tu-muenchen.de

PII S0920-3796(98)00409-8

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 every physical material.

Potential candidates for the communication network of such a future data-acquisition system are fiber distributed data interface (FDDI), Gb-Ethernet, Gb-Asynchronous Transfer Mode (ATM) and SCI. From those, 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 their primary communication medium [1–5]. In industry, SCI gets more and more importance since large computer manufactures are offering off-the-shelf SCI-based products for cluster computing [6,7]. However, for data-acquisition purposes, commercially available SCI products are very rare, and for high-end applications nothing is provided since this market segment is too small. A lot of research remains to be done in this field [8–12,17]. In this paper, the performances of two SCI tests stands are reported that show the principal suitability of SCI for data-acquisition systems in future fusion reactor experiments.

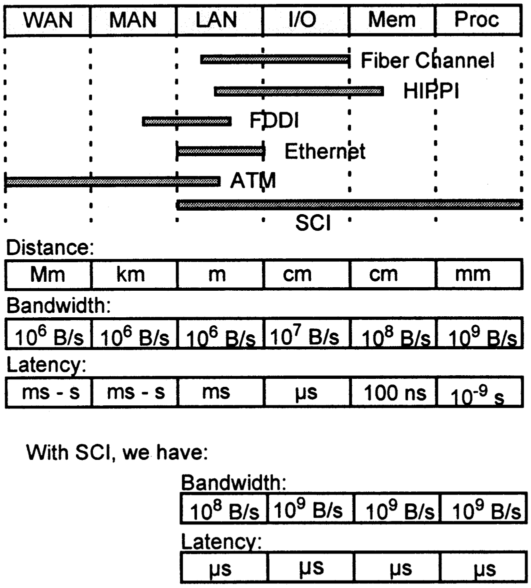
2. The SCI interconnect-technology

The scalable-coherent-interface is standardized by the IEEE and ANSI committees [13] and targets to high-speed, low-latency interconnect applications. Unlike as others, SCI establishes a common, i.e. shared, address-space between all SCI-nodes, and it optionally cares for coherency between processor caches. Because of its high data transmission rates of 1 Gb s^{-1} – GB s^{-1} and its low latencies of a few µs, it is indicated that SCI may be suitable for parallel processing with PCs and workstations in a spatially distributed environment, as well as for high-speed data-acquisition systems. In Fig. 1, the projected application domains of SCI are depicted in comparison to other interconnect technologies. It can be seen that, in contrary to the others, SCI maintains its low latency and high bandwidth in its application areas.

SCI is a state-of-the-art implementation of spatially distributed bus, and thus supplies bus services such as global clock and interrupts, decentralized access arbitration, split transactions and multiple out-standing reads or writes. It avoids the physical limitations of buses by employing unidirectional, impedance-matched transmission-lines that are forming rings out of individual point-to-point links.

For maximum throughput, data packets can be transferred simultaneously on each ring segment. A suite of protocols is defined to ensure reliable delivery of the packets. SCI is network-topology independent and supports arbitrary node types in a heterogeneous environment, such as in the sample configuration of Fig. 2.

IEEE has standardized the basis set-up of any SCI interface (Fig. 3), the flow control mechanism between SCI nodes as well as their higher proto-



WAN	MAN	LAN	I/O	Mem	Proc

Distance:

Mm	km	m	cm	cm	mm

Bandwidth:

10^6 B/s	10^6 B/s	10^6 B/s	10^7 B/s	10^8 B/s	10^9 B/s

Latency:

ms - s	ms - s	ms	µs	100 ns	10^{-9} s

With SCI, we have:

Bandwidth:

10^8 B/s	10^9 B/s	10^9 B/s	10^9 B/s

Latency:

µs	µs	µs	µs

Fig. 1. Comparison of SCI with ATM (asynchronous transfer mode), HIPPI (high performance parallel interface), FDDI (fiber distributed data interface), ethernet and fiber channel. (WAN/MAN/LAN = Wide/Metropolitan/Local Area Network, I/O bus, Mem, memory bus; Proc, processor bus).

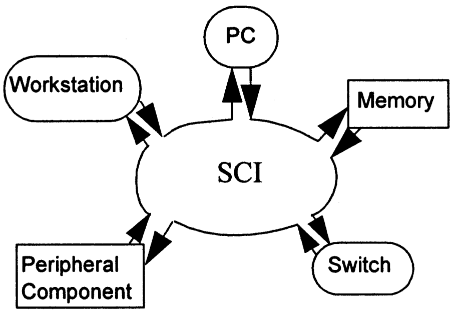


Fig. 2. Sample SCI configuration.

col layers. Each interface contains a bypass-FIFO, an input receiver and an output transmitter with separate receive and transmit queues, respectively, to avoid deadlocks while handling request, response and echo packets. The interface is capable of simultaneously transmitting, receiving and bypassing packets that are circulating in the SCI ring. Transmitted packets are inserted into the data streams according to a bandwidth allocation protocol. Each data packet is preserved in the sender until a positive echo has arrived. If the echo indicates that the packet was rejected by the receiver, for instance due to queue overflow, the sender optionally re-transmits the packet without delay. Additional 'idle' packets circulating in the ring are carrying administrative information about the packet's priority and aging. They provide for dynamic allocation of bandwidth for a subsequent transmission as well as for space allocation in the receiver FIFO. At transmitter and receiver, separate input and output FIFOs are present to avoid deadlocks and to match the attached device's data rate with the SCI link speed.

SCI supports read and write transactions that consist of request and response subactions which are both acknowledged by an echo. The request subaction transfers a data packet with address, status and read/write-command to the responder, the response subaction returns status, and in the read-case also data. Additionally, responseless move transactions are specified in the standard that consist of a request without response to allow fast block transfer. Furthermore, compound transactions such as fetch&add are available to implement locks and semaphores. In Fig. 4, the 4-phase handshake protocol is depicted. All transactions that require a response obey that protocol.

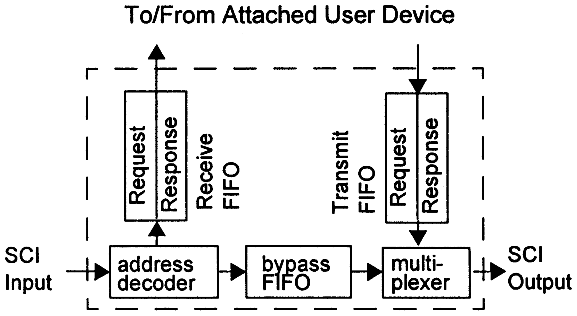


Fig. 3. Standardized SCI interface.

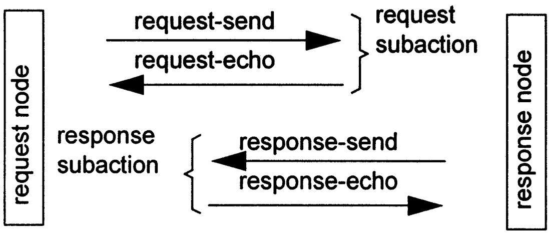


Fig. 4. SCI handshake.

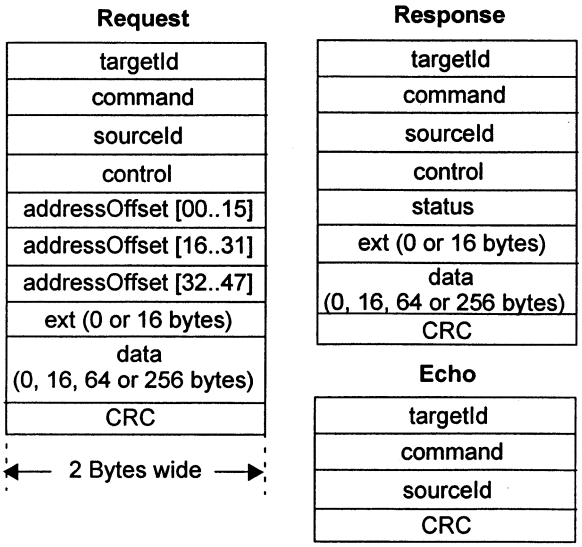


Fig. 5. SCI packet formats.

In Fig. 5, the formats of a request, response and echo packets are shown for the case of non cache-coherent transactions. Every packets is partitioned into two byte portions that are transferred simultaneously in parallel transmission lines.

3. SCI-based data-acquisition systems

Data-acquisition is a real-time task with prescribed reaction times due to 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' throughput in terms of measurements per second. This value determines mainly the technology that has to be employed. For a fusion reactor, data normally need not be sampled with frequencies higher than 100 Msamples s^{-1} per sensor.
- The total amount of sensors that have to be read-out, maintained and operated over a longer period of time (normally 1–2 decades).
- The system's collected amount of data. For contemporary data-acquisition systems, this value lies in the range of 10–100 MB s^{-1} and cycle. Future systems will deliver one to two orders of magnitude more data which imposes high requirements on the data-base- and 'mining' system.
- The system's latency time between sampling and remote storing of the measurement values. This is a crucial factor if any additional feed-forward or feed-back 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 system's data loss rate. By this rate the degree of redundancy that has to be built-in is determined. Since fusion reactor experiments are mission-critical a 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 need ~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 become expanded over the years.

In this paper, throughput and latency of sample SCI-based data-acquisition system is investigated. Data losses are partly evaluated by registering the error rates that occur in the transmission lines. Other data loss factors, for instance due to buffer over-flows, are not considered. The management of the data retrieval, and the overall systems scalability are not in the scope of this paper.

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 frame of the SCINET project, a sample acquisition system was implemented to verify this. Throughput and latency of data transfers from memory were measured by means of test stands.

4. The SCINET test stands

Two different SCI test stands were set up to allow for comparison of the obtained results and to achieve a higher reliance in the conducted tests. Each set-up establishes a SCI data transmission ring, comprising of two PCs that are connected by SCI interfaces. For both stands, commercially available SCI cards from the Dolphin Interconnect Solutions [14] are used, together with appropriate copper cabling for the transmission lines.

In the first stand, two 200 MHz PentiumPro PCs with the 440FX PCI chip set are employed (GA-686DX mother board), together with Linux 2.0.30 as operating system. The Dolphin boards are controlled by a self-written device driver, and also the bench-marks that measure the throughput and latency are self-developed. The second system comprises of two 100 MHz Pentium PCs, Windows NT 4.0 and the standard FX-chip set. Here, device driver and bench-mark software are taken from 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 MB s^{-1} 'link controller chip' [15] that implements the physical layer of SCI, and a PCI bridge that converts the PC's PCI bus protocol into SCI

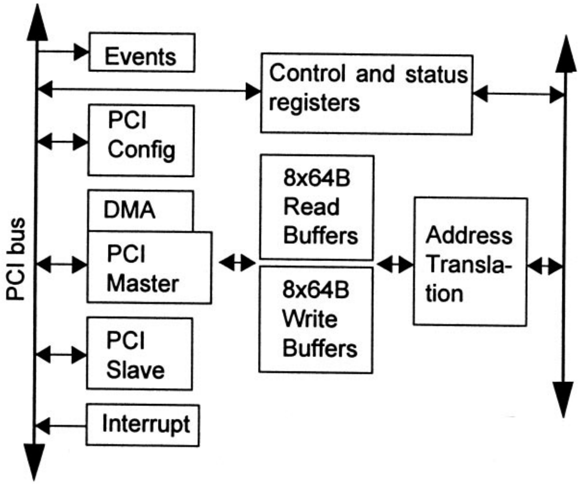


Fig. 6. Dolphin's PCI/B-link bridge.

packets. On the card, the PCI bridge and the link controller chip are connected internally via a proprietary high-performance bus, the so called B-link [16]. The set-up of the link controller chip is identical to the IEEE interface specification as it was depicted in Fig. 3, 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 eight 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 Fig. 6.

The bridge is capable of either splitting its buffers in order to support up to eight external PCI masters, for read and write respectively, or it is capable of combining the buffers to allow for 8-fold throughput for a single master. The prerequisite for buffer combination is that PCI data that has to be transferred to the SCI link exhibits contiguous addresses, and that the control and status registers of the bridge are properly set.

5. Measurement results

The throughputs shown in Fig. 7 were obtained for both, remote read and write transactions between the memories of the PentiumPro-equipped test stand. 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 reaches already for short block lengths of 256 B its maximum throughput. At 64 B, about half of that can be achieved. Remote read is with a factor of 3.5 significantly slower than remote write. This comes from the fact that a remote write transaction is for the requester 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. On the contrary, 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' -operation should be preferred to a 'pull'-strategy: the sensors should write their measured data by their own to the final destination 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.

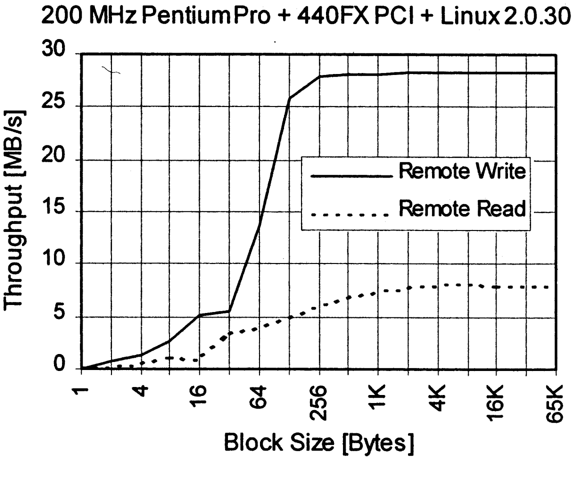


Fig. 7. Measured throughput of test stand.

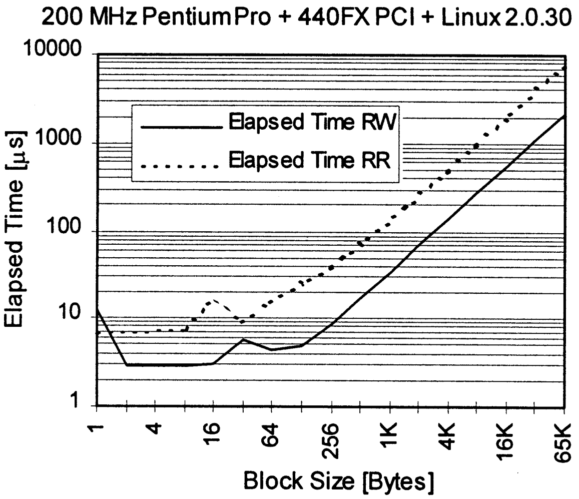


Fig. 8. Measured latencies of remote read and write transactions.

If one considers the elapsed times of the data transfers vs. the block size (Fig. 8) it becomes clear that the set-up times to initiate a transfer are considerably small: ≈10 µs for both, read and write. That would allow for a very small reaction time of an optional SCI-based control system. The elapsed times for remote read are higher, compared to remote write due to its slower transfer rate.

The Pentium-based test stand with the software from Dolphin shows a throughput that is depicted in Fig. 9 for comparison. Here, only 12 MB s^{-1} can be achieved. Additionally, it is presented in Fig. 9 how the transfer rate behaves in the case that on top of SCI's hardware error detection and correction a second error check with optional retry of the last sent block is put that is implemented in software. 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 KB. Without software over-head on the contrary, Dolphin's solution reaches for 64 B already its maximum transfer rate and remains from then on constant. The latency times for larger block sizes (depicted in Fig. 10) are in compliance with the measured throughput. For small blocks, the latency becomes as low as 2–3 µs which indicates an efficient implementation of the device driver. Of course, more time is needed (11–12 µs) if additional software error correction is used.

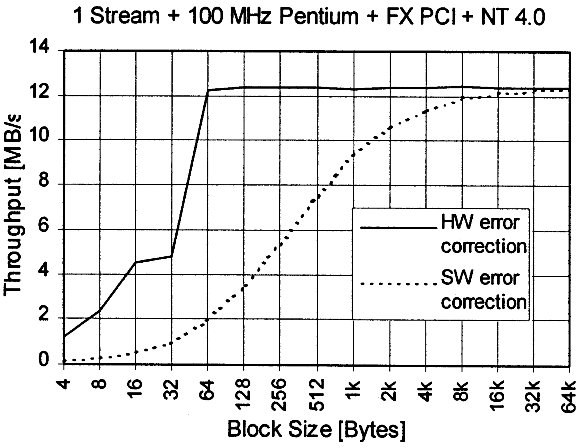


Fig. 9. Throughout of 1-stream remote write.

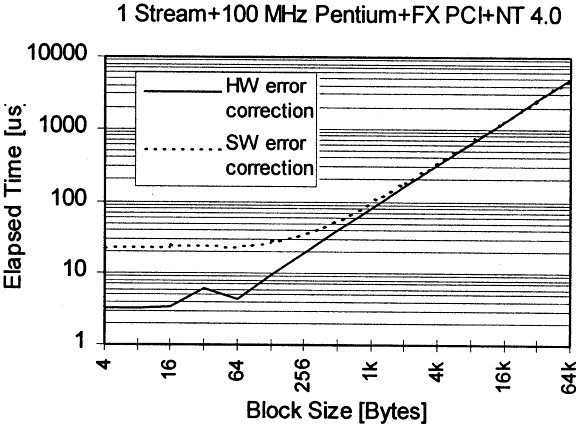


Fig. 10. Latencies of 1-stream remote write.

Dolphin's device driver allows the combining of the PCI/B-link bridge write-buffers. This valuable feature was not implementable in our self-written Linux device driver since the documentation of the bridge's control and status register is not available for public. As one can see from Figs. 11 and 12, the throughput scales nearly linearly with two and four buffers combined.

Only the minimum required buffer size for full transmission speed is doubled each time.

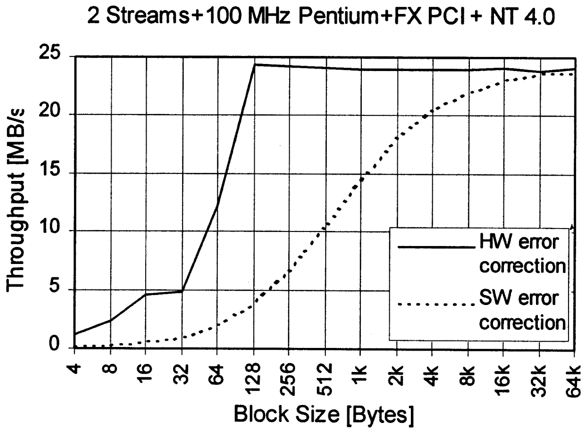


Fig. 11. Throughput of 2-streams remote write.

The maximum achievable data transfers rate proves to be 45 MB s^{-1} since with eight combined buffers, saturation effects are showing up (Fig. 13). Because of the fact that each write buffer has a capacity of 64 B, the combining of two or four buffers requires that at least 128 or 256 B, respectively, that are aligned to 64 B boundaries in the PCI address space, are available for transfer. Exactly at these block sizes, throughputs exhibit their peak performance. In the case of eight buffer-combining, bursts of ≥ 512 B suffer to be transferred with a reduced speed of only 34 MB s^{-1}. Either the PCI bus, the slow Pentium CPU or some other limiting factor causes this saturation effect that forces the significant drop in performance.

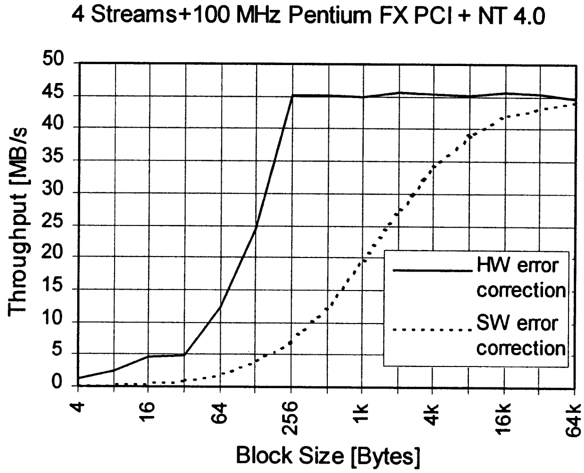


Fig. 12. Throughput of 4-streams remote write.

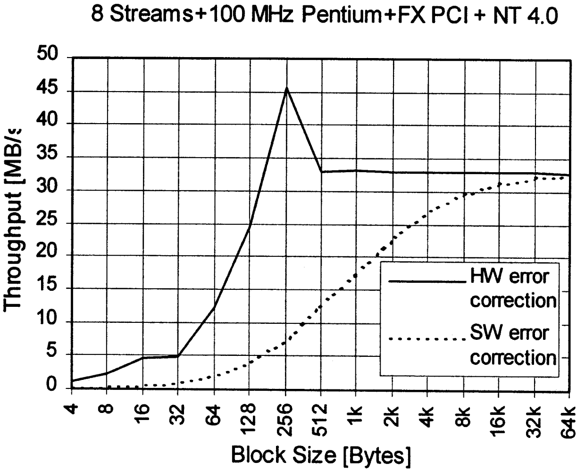


Fig. 13. Throughput for 8-streams remote write.

For completion of the conducted tests, the latency times are given for two, four and eight combined streams in Figs. 14–16, respectively.

6. Summary and conclusions

In this paper, the feasibility of an SCI-based data acquisition system is demonstrated by the achieved throughputs and latencies. The sample DAQ system is based on PC test stands on which

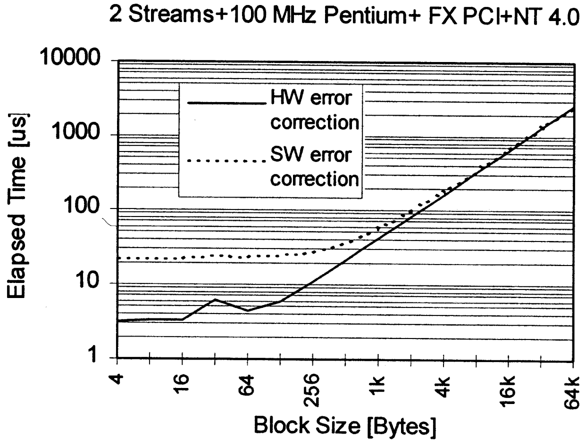


Fig. 14. Latency for 2-steams remote write.

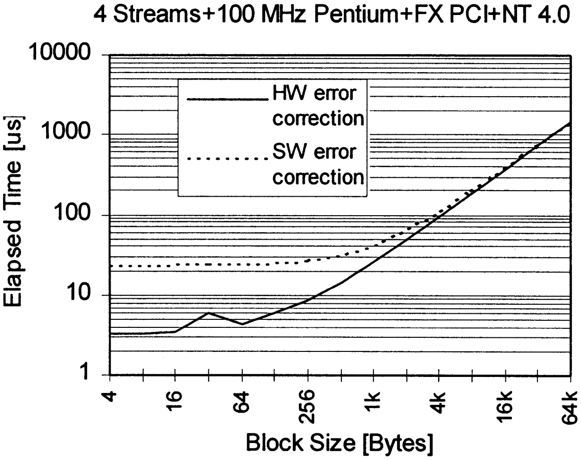


Fig. 15. Latency for 4-streams remote write.

a decent performance was also measured for the case of additional software error checking and correcting. In the future, the test-stands will be upgraded to glass fiber transmission lines and an SCI switch will be included to study its influence. Furthermore, the interplay between the transmissions performance and the higher software levels have to be analyzed. Finally, a programming model must be devised that is suitable for the physicists' needs.

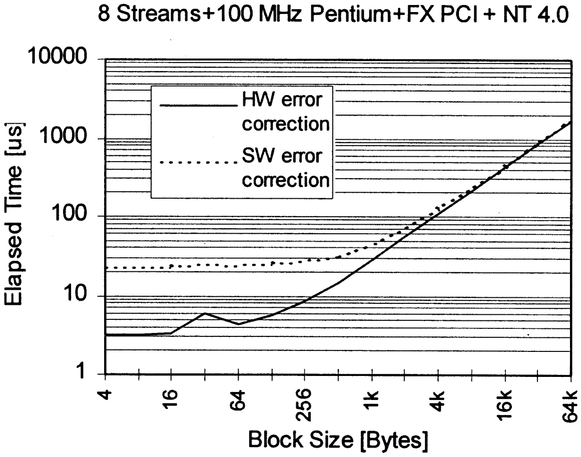


Fig. 16. Latency for 8-streams remote write.

References

[1] A. Bogaerts, R. Divia, H. Müller, J. Renardy, SCI based Data Acquisition Architectures, IEEE Trans. Nucl. Sci. 39, (2) April 1992.

[2] 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.

[3] E.H. Kristiansen, G. Horn, S. Linge, Switches for point-to-point links using OMI/HIC technology, in: Int. Data Acquisition Conference on Event Building and Data Readout, Fermi National Accelerator Laboratory, Batavia, IL, USA, Okt. 1994.

[4] B. Wu, SCI Switches, Int. Data Acquisition Conference on Event Building and Data Readout, Fermi National Accelerator Laboratory, IL, USA, Okt. 1994.

[5] A. Bogaerts, H. Mueller, et al., RD 24 Status Report 1996, http://www.sunshine.cern.ch/RD24/Report96.pdf

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

[7] R. Clarke, Data General, K. Alnes, Dolphin Interconnect Solutions, An SCI Interconnect Chipset and Adapter, Proc. Hot Interconnects Symposium IV, Stanford University, Aug. 15–17, 1996.

[8] H. Richter, M. Liebhart, Performance Optimizations of Switched SCI-Rings, Proceedings 11th Annual International Symposium on High Performance Computing Systems (HPCS'097), 10–12. Juli 1997, Winnipeg, Kanada.

[9] S. Scott, J. Goodman, M. Vermon, Performance of the SCI Ring, Proceedings IEEE ISCA 92, Queensland, May 1992.

[10] 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.

[11] M. Liebhart, A. Bogaerts, E. Brenner, A Study of an SCI Switch Fabric, Proceedings IEEE MASCOTS '97, Haifa, Israel.

[12] D. James, The Scalable Coherent Interface: Scaling to High-Performance Systems, Apple Computer Cupertino, CA.

[13] IEEE Standard for Scalable Coherent Interface (SCI), IEEE Std 1596–1992.

[14] Dolphin, PCI/SCI Cluster Adapter Specification, Dolphin Interconnect Solutions, Oslo, Norway, 1996.

[15] Dolphin, Link Controller LC-I Specification, Dolphin Interconnect Solutions, Oslo, Norway, 1995.

[16] Dolphin, A Backside Link (B-link) for Scalable Coherent Interface (SCI) Nodes, Dolphin Interconnect Solutions, Oslo, Norway, 1994.

[17] H. Richter, Interconnection Networks for Parallel and Distributed Systems, Textbook in German, 320 p., Spektrum Akademischer Verlag, Heidelberg, Germany, 1997.