4.6 Performance Optimization of SCI-Rings

by Harald Richter

4.6.1 Introduction

The Scalable Coherent Interface (SCI) defines a high-speed interconnect that provides for a distributed shared memory with optional cache coherence [IEE92]. Up to 64 K nodes can be closely coupled in one or more rings that are concatenated via SCI switches. However, it has become appearent that data-packet overload in a single SCI receiver leads to poor throughput in the whole ring because excessive retry traffic limits the overall bandwidth. To cope with this, a retry delay is proposed to be added to the SCI protocol that shows significant improvement.

4.6.2 The Ring Saturation Problem

Simulations of SCI systems have revealed that the data throughput of an SCI ring is sensible to traffic overload in any receiving node contained in the ring. Generating too many packets per time unit effects a full receiver's input fifo, thus resulting in negative echo packets transmitted back to the sender. This leads to transaction retries which occupy much of the system's bandwidth. The data throughput decreases, while the retry- and negative echo traffic increase dramatically. So, it is important not to overload any SCI receiving node in order to achieve a decent ring performance.

The degradation caused by the retry traffic occurs if, on a mid-range time scale, there is a mismatch between the packet-generation rate of a sender and the packet-acceptance rate of the receiver. Speed fluctuations on a short-range scale are balanced by the input buffers of the receiver, provided that they are not full. However, in a SCI system that consists of multiple rings connected by switches, for example, the situation always arises that more than one sender in a ring wants to communicate with one or more receivers in another ring. Then, the switch port which is the gate to the other ring becomes a bottleneck, because it can serve only one sender at a time. The switch port will issue negative echos at high rates that finally effect that all other ongoing communications in the ring are deprived from their bandwidth.

The following investigations address that performance degradation problem. It will be shown that the undesired effect of the retry traffic can be overcome by adding retry delay times between negatively acknowledged requests or responses and their subsequent retry transmissions. Positively acknowledged packets are not affected by the proposed delay.

4.6.3 Retry Delay Times

For the investigations described herein, we used SCILab [WB95] [BW94], a simulation tool for SCI-based networks that was developed at CERN. It is suited to study the influence of link speed, FIFO sizes, bypass delays and topological network structures under various load

conditions. Effects due to bandwidth allocation, queue acceptance and switch delays are taken into account. Large systems, consisting of multiple rings that are connected by switches, can be explored in principle. Additionally, other simulations were carried out on SCINET, our own SCI network simulator, that are backing the SCILab results, although showing slightly different values. The results gained with SCINET will be published later for better comparison.

To observe saturation on a ring, we have to configure one of the ring's nodes as a slow responder to which transactions are continuously directed by several fast requesters. In Dolphin's implementation of a 4-port SCI switch [Dol95], for instance, the switch-internal bandwidth is significantly lower than the accumulated ports' bandwidths, so that each port must be considered a slow responder in its ring (Figure 4.9).

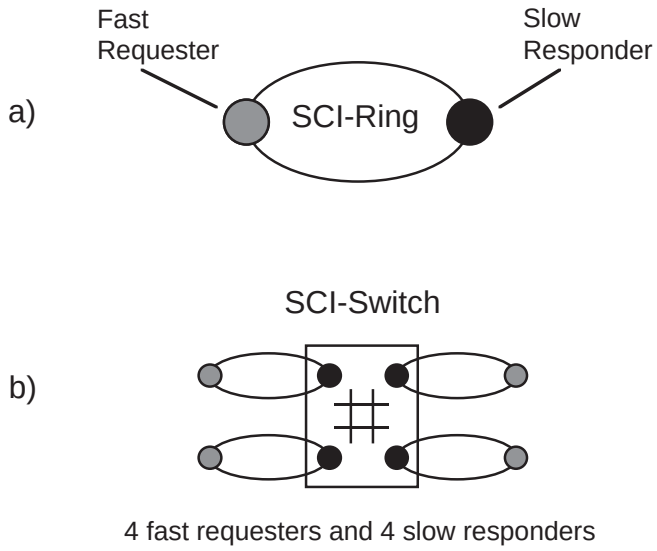

Figure 4.9: Simulation Configurations: Single SCI Ring (a) and Multiple Rings with Switch (b)

In our simulations, we used the SCI `dmove64` command as a transaction request. Although this command requires no response, this is not a methodical restriction since from the retry point-of-view there is no difference between a retried request or retried response packet. The advantage of this is that only request packets, request retries and request echoes have to be monitored. The transmission rate was chosen to be 500 MB/s for both, SCI-rings and switch-internal interconnects. Thus, the transfer bandwidth of a four-port switch is exactly one fourth of the summed bandwidths of the rings connected to it. All other parameters for the simulation were taken from Dolphin [Dol96a]: 40 ns bypass delay, 70 ns internal bus-to-SCI delay, 120 ns SCI-to-internal bus delay. Furthermore, propagation delays due to the speed of light ($\approx$4 ns/m) were added. In each ring, the fast requesters are modelled as data generators that send packets at random times to their slow responder. The simulations also take into account that each SCI packet that has to transit a ring, requires an encapsulation overhead of 8 bytes due to the protocol of the switch's internal bus.

In Figure 4.10 on page 130, for an example of three fast nodes and one slow node, the simulated throughput of each ring is depicted, versus various initial delay time values. It is

important to notice that on the x-axis the initial value D_{start} for the retry delay-time D is varied, not D itself. The delay time is obtained according to one of the following strategies:

- Constant delay time: Always the same time, given by the initial value, passes before a specific packet is retried: $D = D_{start}$.
- Linear increase of the retry delay: For each new delay time, the initial value is added to the previous delay time: $D = r \cdot D_{start}$, where r indicates the number of rejections of a specific packet. If a packet is rejected the first time, we have $r = 1$.
- Exponential increase of the retry delay: For each new delay time, the previous delay is multiplied by two: $D = D_{start} \cdot 2^{r-1}$, where r is the number of rejections.

The linear and the exponential delay time strategies are adaptive, since they take into account how often a specific packet has been rejected so far.

In the simulation results in Figure 4.10 , the total throughput of one SCI ring is shown that consists of three sending nodes and one switch port (caption title “raw”). Additionally, the fraction of the bandwidth is depicted that is occupied by the retry traffic (“retry”). Both curves are parametrized by three different retry delay strategies. It can be seen that the retry traffic decreases asymptotically to 0 as the initial value D_{start} of the delay is increased. At the same time, the ring throughput approaches its theoretical limit of about 500 MB/s (the switch’s bandwidth) for an infinite delay. Simulations also show that the exponential increase performs best, followed by the linear increase, since it reaches the maximum throughput already for small inital values D_{start}. The factor of improvement is ≈2 for large delay times compared to the case where no delay is present.

The same improvement can be observed when the slow responder is heavily overloaded. This was simulated by increasing the ratio of fast requesters to slow responders from 3:1 to 9:1 (Figure 4.11 on page 131). If a minimum delay time is present, the throughput doubles from 0.22 GB/s to ≈0.42 GB/s.

In Figure 4.12 on page 132 the corresponding retry traffic is depicted. It shows that the exponential strategy allows a decent throughput also in the case of heavy overload by rapidly minimizing retry traffic already for small initial values D_{start}.

By comparing the asymptotic values of Figure 4.11 on page 131 of ≈425 MB/s with the value of ≈500 MB/s for the 4-node system, it becomes apparent that the SCI ring has roughly retained its throughput, even if the number of nodes has increased. This demonstrates the benefit of the retry-delay mechanisms that are not provided with the original SCI standard.

The simulation results of Figure 4.10 – Figure 4.12 on page 132 can be explained as follows: If a packet suffers from being rejected once, it its avantageous for the ring throughput to delay that packet as long as possible, to allow newly generated data to occupy the full bandwidth (“looser-gets-nothing” principle). The exponential backoff scheme has proven best, because it quickly provides large delays even for small values of D_{start}.

On the other side, maximizing throughput means that the delay for packet delivery is also maximized, which is untolerable for many applications, especially from the real-time sector.

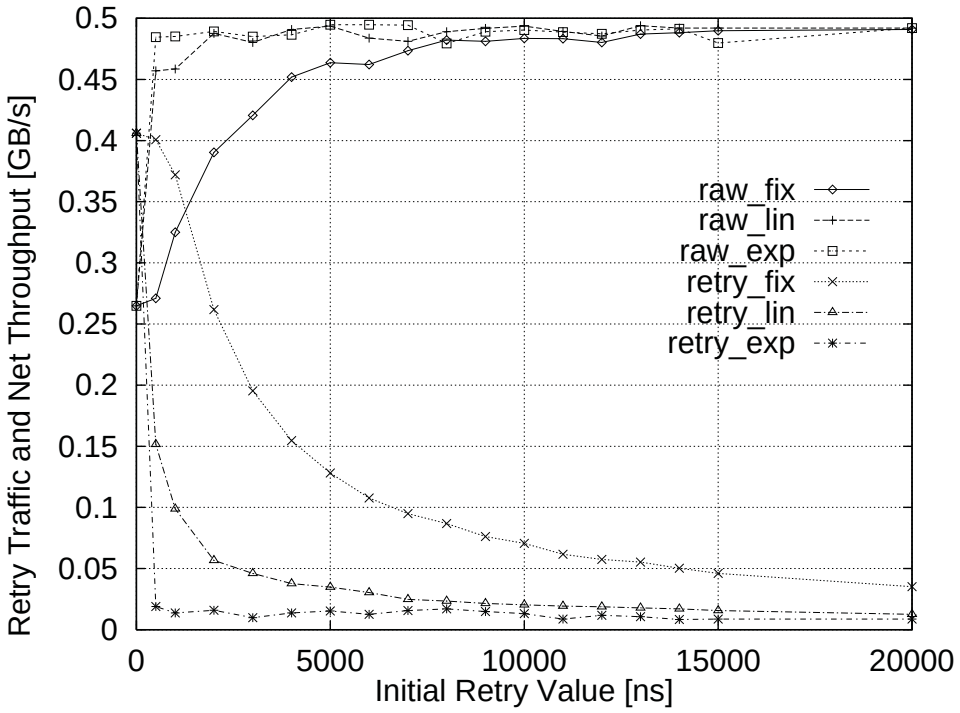

Figure 4.10: Throughput of 1 SCI Ring with 3 Fast Nodes as Senders and 1 Switch Port as Slow Node

Also for microprocessors with several hundreds of MHz clock frequency, the costs for accessing remote data would be very high. Additionally for reliable transmissions, the optional SCI time-out and aging mechanisms have to be employed by higher protocol layers. Because of these factors, an upper limit for the applicability of the proposed looser-gets-nothing strategy is imposed.

Figure 4.10 and Figure 4.12 on page 132 indirectly reflect the retry count r, because the high retry traffic that results from small values D_{start} is caused by repeatedly retrying the same packets. This means that for the linear and exponential strategy, r decreases strongly non-linear if D_{start} is increased.

In the previous simulations, it was assumed that the retry delay is performed by the transaction requester, but this is not mandatory. It is also possible that the busy responder delays the sending of a negative acknowledge. In both cases, a hardware counter/timer is required to implement the delay. For the responder-controlled delay, additional buffer must be provided to store the echo packet until it is transmitted. The requester instead must store its sent packet anyway, until a positive acknowledge is received. In the latter case, no additional buffers are required. Therefore, it is better to implement the delay in the requester.

4.6.4 Conclusions

It was shown that the system performance of switched SCI rings can be significantly improved, if a delay time for the retransmission of rejected SCI packets is introduced. Simulations indicate

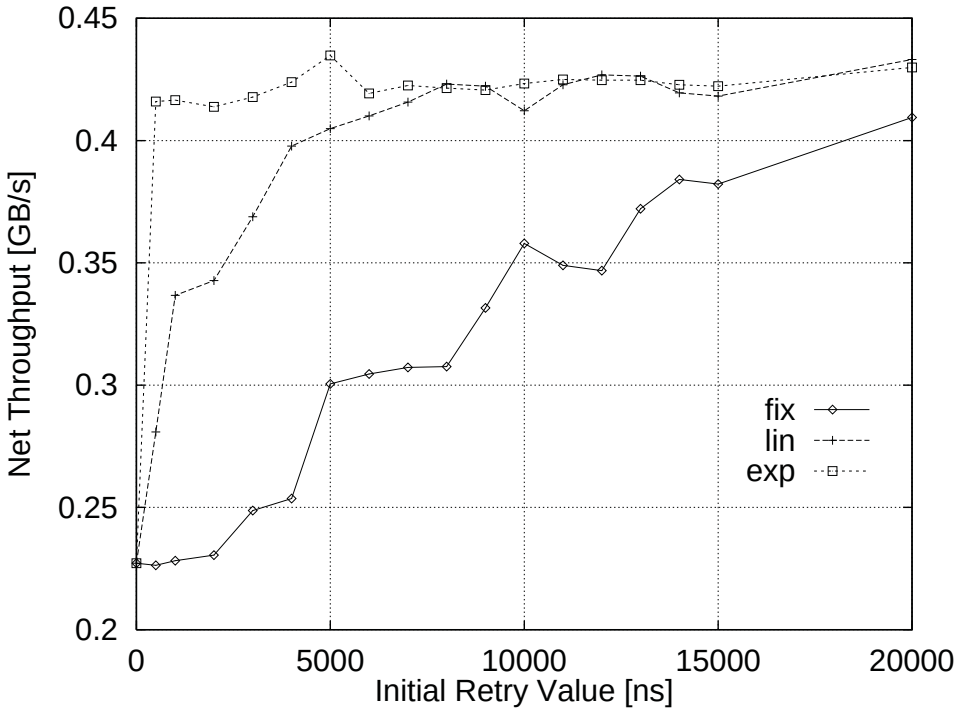

Figure 4.11: Throughput of 1 SCI Ring with 9 Fast Nodes and 1 Slow Node

that the throughput increases and the retry traffic decreases both in a strongly non-linear manner if the initial delay time is increased. A strategy with exponentially increasing delay proves best compared to a fixed or linearly increased time. The effects found can be exlained by the "looser-gets-nothing" principle that says that packets that have been rejected once get the lowest transmission priority compared to newly generated ones.

4.6.5 References

[BW94] *A. Bogaerts and B. Wu*, editors. "SCI Simulations with SCILab". European SCI Workshop, Oslo, Norway (1994).

[Dol94] *Dolphin*. "A Backside Link (B-Link) for Scalable Coherent Interface (SCI) Nodes". Oslo, Norway (1994).

[Dol95] *Dolphin*. "4-way SCI Cluster Switch". Oslo, Norway (1995).

[Dol96a] *Dolphin*. "Dolphin ICS, private communication" (1996).

[Dol96b] *Dolphin*. "Link Controller LC II Specification". Oslo, Norway (1996).

[IEE92] *IEEE*. "Standard for Scalable Coherent Interface SCI". IEEE Std 1596-1992 (1992).

[JG92] *R. Johnson and J. Goodman*, editors. "Synthesizing General Topologies from Rings". Proceedings of ICPP'92 (August 1992).

[Joh93] *R. Johnson*. "Extending The Scalable Coherent Interface For Large-Scale Shared-Memory Multiprocessor, PhD thesis". University of Wisconsin-Madison (1993).

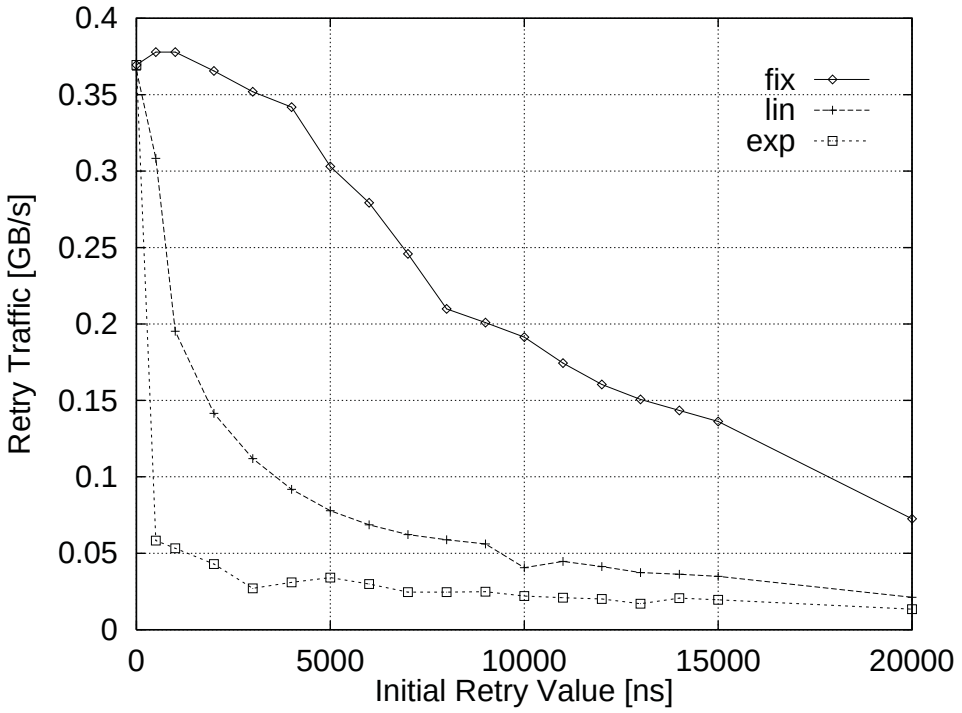

Figure 4.12: Retry Portion of the Throughput of 1 SCI Ring with 9 Fast Nodes and 1 Slow Node

[LBB97] *M. Liebhart, A. Bogaerts, and E. Brenner*, editors. "A Study of an SCI Switch Fabric", Haifa, Israel (1997). Proceedings IEEE MASCOTS'97.

[LVA82] *T. Lang, M. Valero, and I. Alegre*. "Bandwidth of Crossbar and Multiple-Bus Connections for Multiprocessors". IEEE Computer (December 1982).

[OP96] *K. Omang and 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.

[Ric97] *H. Richter*. "Interconnection Networks for Parallel and Distributed Systems". Spektrum-Verlag, Heidelberg, Germany (1997). Textbook in German, 320 p.

[RL97] *H. Richter and M. Liebhart*. Performance Optimizations of Switched SCI-Rings. pages 463–472, Winnipeg, Kanada (1997). Proceedings 11th Annual International Symposium on High Performance Computing Systems (HPCS'97). 10.-12. Juli.

[RO97] *H. Richter and M. Ohlenroth*. Data Acquisition with the SCINET, a Scalable-Coherent-Interface Network. Garching (1997). IPP Proceedings of the Technical Committee Meeting on Data Acquisition and Management for Fusion Research of the International Atomic Energy Agengy. July 22.-24.

[Rob93] *T. Robertazzi*. "Performance Evaluation of High Speed Switching Fabrics and Networks". IEEE Computer Soc. Pr. (1993).

[SY94] *I. Scherson and A. Youssef*. "Interconnection Networks for High-Performance Parallel Computers". IEEE Computer Soc. Pr. (1994).