

A Comparative Throughput Analysis of Scalable Coherent Interface and Myrinet

S. Millich, A. George, and S. Oral

*HCS Research Lab, ECE Dept., University of Florida, Gainesville, FL 32611
{millich, george, oral}@hcs.ufl.edu*

Abstract

It has become increasingly popular to construct large parallel computers by connecting many inexpensive nodes built with commercial-off-the-shelf (COTS) parts. These clusters can be built at a much lower cost than traditional supercomputers of comparable performance. A key decision that will greatly affect the overall performance of the cluster is the method used to connect the nodes together. Choosing the best interconnect and topology is not at all trivial since performance and cost will change as the system size is scaled. This paper presents throughput models used for the analysis and comparison of performance in two leading System Area Networks (SANs), Myrinet and Scalable Coherent Interface (SCI). First, analytical models for throughput are developed by determining the theoretical bandwidth of all internal buses and links that are part of the interconnect architecture. Then, experiments are conducted to measure the actual bandwidth available at each of these components, and the models are calibrated so they accurately represent the experimental results. Finally, the models are used to compare the maximum throughput of Myrinet and SCI systems with respect to system size and overall dollar cost.

1. Introduction

Not long ago, substantial computing power was reserved only for those who could afford a supercomputer. These high-performance systems have always been very expensive due to the high design cost and relatively small market for them. Today however, powerful computer clusters can be built for a fraction of the cost of traditional supercomputers by combining inexpensive, mass-produced PCs with a high-performance System Area Network (SAN).

Selection of the SAN becomes an important decision that will greatly affect the overall performance of the cluster. First, the required system bandwidth and acceptable level of latency must be determined. Latency and throughput depend on the interconnect technology, system size and topology. The additional performance offered by one network should be weighed against the extra money it will cost. Choosing the best interconnect can pose a significant challenge when attempting to

assemble a high-performance cluster, where performance, scalability, and cost must be taken into account.

A method for directly comparing different SANs and topologies is needed to make cluster design better and more efficient, resulting in better performance for each dollar spent. Also, the deficiencies of current SANs need to be exposed, so that next-generation products can be improved. With models of system performance in terms of throughput, latency, and cost, educated choices can be made easily, and the above needs can be met. This paper presents a throughput analysis and comparison of SCI and Myrinet, two of the most widely used high-performance SANs.

A number of papers have appeared in the literature investigating SCI and Myrinet performance. For instance with SCI, Ibel et al. [1] provided a throughput performance analysis of an SCI-based cluster. Omang and Parady [2] used throughput measurements to examine the scalability of SCI rings. Horn [3] applied an architecturally motivated approach to develop a throughput model for a single ring. This model was used to show the scalability of the SCI ring for different PCI bandwidth capabilities. The study did not include any other topologies and the effect they had on throughput. Bugge [4] examined all-to-all communication on multicubes to find the theoretical bandwidth limits for second-generation PCI-SCI adapters. The architecture was analyzed in depth, but no experiments were performed to show actual performance or to verify the approach.

The simulative performance analysis performed by Sarwar and George [5] presented analytical derivations for average paths taken by SCI request and response packets. These analytical expressions were used for verification of simulative results, but no validations were made using experimental data.

Gonzalez et al. [6] developed analytical models of SCI shared-memory latency from an architectural perspective. The models were used to project the performance of multi-dimensional SCI topologies. Experimental data was used to derive and validate the models. Though latency is an important piece of the overall performance of a system, it is hard to accurately and fairly compare different SANs without also considering bandwidth and dollar cost.

SCI and Myrinet were compared at several different levels by Kurmann and Stricker [7] in determining the

performance characteristics of simple optimized remote load/store operations, optimized message-passing libraries, and also a connection-oriented TCP/IP LAN networking protocol. Low-level and MPI performance of Myrinet was examined by Hsieh et al. [8] and compared to GigaNet's hardware implementation of the Virtual Interface Architecture (VIA) known as cLAN.

Several Myrinet-based systems studies have also appeared in the literature. For instance, Brightwell and Plimpton analyzed the performance and scalability of two large clusters [9], while Bal et al. did the same on the distributed ASCI supercomputer [10]. In addition, other research has focused on simulative analysis of Myrinet, such as George and VanLoon [11] that presented a high-fidelity, event-driven model for performance analysis of Myrinet SANs. Their simulation was verified analytically and validated through comparison to experimental testbed results.

By contrast, the research herein focuses on the experimental analysis of throughput on SCI versus Myrinet. Analytical models are developed and calibrated from the testbed, after which they serve to provide for performance projections with systems of various sizes and topologies.

2. Overview

Currently, two of the more prominent interconnects for high-performance clusters are SCI and Myrinet. This section provides a brief overview of each SAN, followed by an overview of the analysis that will be performed in Sections 3 and 4.

2.1. Scalable Coherent Interface

The SCI standard describes a packet-based protocol using unidirectional links that provides participating nodes with a shared-memory view of the system [12]. It specifies transactions for reading and writing to a shared address space, and features a detailed specification of a distributed, directory-based, cache-coherence protocol.

There are many advantages of using SCI to fulfill the high-speed networking demands of cluster computing. SCI is well suited to support finer-grained parallel computations because of its low-latency performance. Typical systems can achieve single-digit microsecond latency performance for small messages. SCI offers support for both the shared-memory and message-passing paradigms, unlike most competing systems [13].

For the rest of this study, we will be considering Dolphin Interconnect's implementation of SCI using Scali's software platform. Both hardware and software can be purchased together as the Dolphin/Scali Wulfkit. The 64-bit, 66 MHz PCI-SCI adapters has a link data rate of 5.33 Gbits/s in current systems.

2.2. Myrinet

Second-generation Myrinet connects computing nodes through full-duplex 1.28 Gb/s (160MB/s) point-to-point links, and low-latency, cut-through switches [14]. A Myrinet interface to a host computer nominally has one port. The ports of Myrinet interfaces are the only points where new packets are injected into the network, and the only points at which they are properly consumed. A Myrinet switch is a multiple-port component that switches packets from the incoming channel of a port to the outgoing channel of another port selected by a source route defined in the packet header.

Any way of linking together interfaces and switches is allowed. The network topology can be viewed as an undirected graph. It can contain cycles (necessary for multiple-path redundancy) and can include unpowered host interfaces and unused switch ports [15].

Myrinet packets may be of any length, and thus can encapsulate other types of packets without an adaptation layer. Each packet is identified by type, so that Myrinet can carry packets of many different protocols simultaneously.

2.3. Analytical analysis

Since SCI and Myrinet support a range of topologies and are very different architecturally, a fair comparison of achievable throughput is not straightforward. By looking at the architectures of both, we can determine how the available bandwidth from a node will be restricted by the different components of each interconnect.

Of course, the effective bandwidth of the links limits the throughput of both SCI and Myrinet systems. For the distributed switching of the SCI torus, the internal bus that connects the multiple rings also can restrict the available bandwidth. Myrinet switches are non-blocking crossbars that provide full bandwidth between all available ports. But for larger system sizes that require more than one crossbar switch, the Myrinet links connecting these switches become the main limiting factor on throughput.

Packet structure and efficiency of the links and buses are also important issues. Packet overhead causes the effective data rate to be less than the gross bandwidth of a network. After considering all the architectural elements that contribute to available bandwidth restrictions, we can model the maximum throughput for systems of various size and topology.

3. SCI analysis

In this section, we analyze the architecture of Dolphin's SCI to form a throughput model. Experiments are conducted to measure the actual maximum throughput of the SCI links and buses so that the throughput model can be calibrated.

3.1. Analytical investigation

The block diagram given in Figure 1 shows the basic structure of an SCI NIC. Each link controller (LC-3) handles the interface to an incoming and an outgoing SCI link. The board shown could be used in either a 1D or 2D topology since it has two link controllers. Adapters supporting a higher number of dimensions would have an additional LC-3 chip connected to the B-Link bus for each additional dimension. Notice that when any packet switches dimensions (rings), it must be transferred from one LC-3 to another through the B-Link.

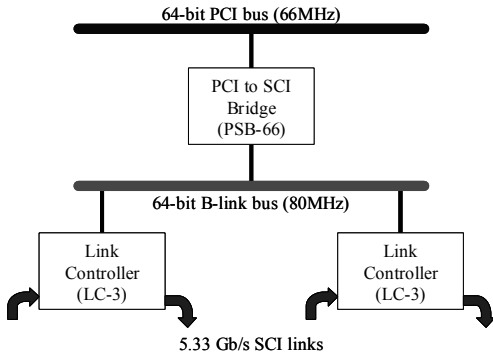


Figure 1. Block diagram of the PCI-SCI adapter

Bugge has analytically determined the available bandwidth for multicube topologies using Dolphin's 32-bit, 33MHz PCI-SCI adapters [4]. He looked at all-to-all communication and calculated the total number of packets that must pass through each link and bus as a function of system size. We use the same approach to examine the available bandwidth of the next-generation SCI hardware.

The total number of packets that traverse a B-Link bus is called the *hot-B-link*. This variable represents all traffic generated by or destined for a node, as well as any traffic that changes dimensions at that node. Similarly, the total number of packets that must flow through a single SCI link is called the *hot-link*. All traffic passing through a node, including packets that are forwarded along the same ring, comprise the hot-link. The hot-link and hot-B-link have been introduced and derived in detail by Bugge [4] for multicubes, so only the final equations are repeated in this paper, in Eqs. 1 and 2, respectively.

Consider a regular r -ary f -cube, where each node is connected to f dimensions, with r nodes in each ring. The total number of nodes is $N = r^f$. During an all-to-all

communication, each of the N nodes sends a message to the remaining $N-1$ nodes, resulting in $N(N-1)$ packets being sent. The total number of packets that must pass through each SCI link is found to be

$$hot-link_{SCI} = N \left(\frac{r-1}{2} \right) \quad (1)$$

The total number of packets that cross each B-Link is

$$hot-B-link_{SCI} = 2(N-1) + \sum_{d=2}^f (d-1) \frac{f!(r-1)^d}{d!(f-d)!} \quad (2)$$

Starting with the gross bandwidths of the buses and links given in Table 1, we will work our way towards the overall available bandwidth of the SCI adapter.

Table 1. Gross bandwidth of SCI links and buses

Bus	Width (bits)	Frequency (MHz)	Gross B/W (MB/s)
PCI	64	66	533
B-Link	64	80	640
SCI link	16	166 (DDR)*	667

* uses rising and falling clock edges

First the effective bandwidths are calculated by multiplying the gross bandwidth by the corresponding packet or cycle efficiency as given below in Eq. 3.

$$B_{eff} = B_{gross} \times Efficiency \quad (3)$$

Efficiency is the ratio of the sizes of data payload versus payload plus overhead. The necessary overhead bytes and cycles are well described in [4] for Dolphin's second-generation adapters. The 64-bit, 66MHz PCI-SCI adapter supports 128-byte payloads, which results in an efficiency on the B-Link and SCI links of 66.7%. Therefore, the effective bandwidth of an SCI link is 444MB/s or 66.7% that of its gross bandwidth (667MB/s), while the effective bandwidth of a B-Link is slightly lower at 427MB/s.

Now that the *hot-link* and *hot-B-link* traffic and effective bandwidth are known, the bandwidth that is available to a node is the $N-1$ packets it issues, divided by the traffic, multiplied by the effective bandwidth. Eqs. 4 and 5 formulate the available bandwidth to a node as follows:

$$B_{SCI-link} = \left(\frac{N-1}{hot-link_{SCI}} \right) B_{eff} \quad (4)$$

$$B_{B-link} = \left(\frac{N-1}{hot-B-link_{SCI}} \right) B_{eff} \quad (5)$$

The effect of the *hot-link* and *hot-B-link* traffic on a node's bandwidth is shown separately in Figures 2 and 3.

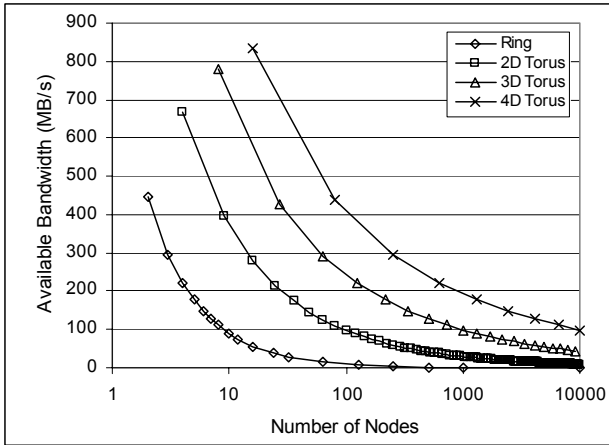


Figure 2. SCI available bandwidth per node limited by the SCI links ($B_{SCI-link}$)

The limited scalability of a simple ring can clearly be seen in Figure 2. Available bandwidth is restricted sharply by the SCI links as system size increases. The restriction is alleviated some for higher dimensional topologies since they have fewer nodes per ringlet for a given system size. Conversely, a higher number of dimensions increases the restriction imposed by the B-Link, as shown in Figure 3. Higher dimensional topologies require more dimension switching, which in turn increases traffic over the B-Link.

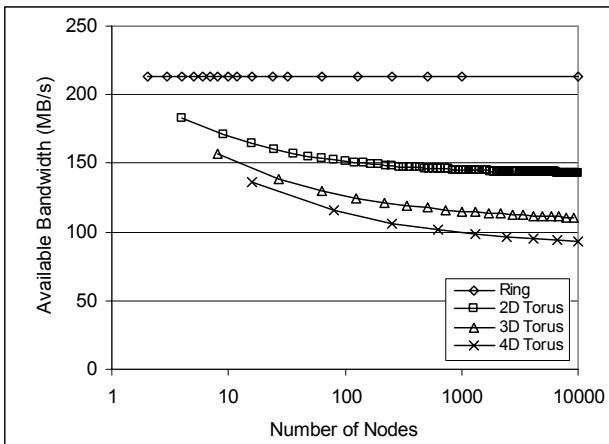


Figure 3. SCI available bandwidth per node limited by the B-Link bus (B_{B-link})

The overall bandwidth available to a node is rendered as the minimum of the available bandwidths of the SCI link and B-Link. Using Figure 4, the best dimension can be determined for various size systems.

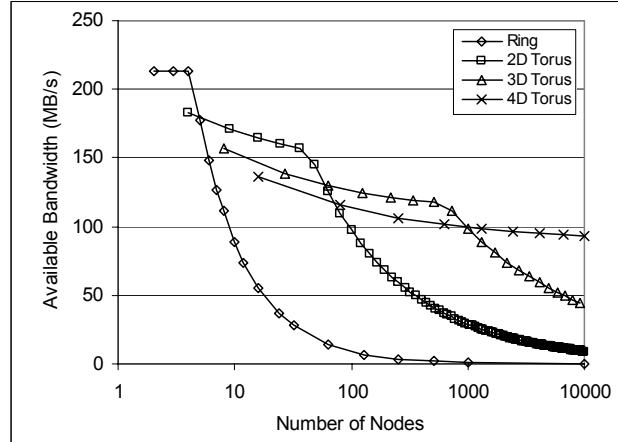


Figure 4. SCI available bandwidth per node limited by SCI links and the B-Link bus

The results indicate that a simple ring offers the highest bandwidth for small systems of five nodes or less, after which point the 2D torus becomes the best option. With the bandwidth that is currently supplied by the B-Link, switching to a 3D torus topology is only beneficial for systems with more than 64 nodes.

This throughput model is currently based on the maximum theoretical bandwidth of the links and buses. Many different factors can contribute to reduced throughput when using the actual hardware. In the following section, we will discuss experiments that are designed to expose the practical maximum throughputs of the SCI link and the B-Link bus.

3.2. Experiments and results

Measurements in the following experiments were made with a modified version of *mpptest*, an MPI benchmark that is distributed with the popular MPI implementation, MPICH [17]. Several modifications were necessary in order to make accurate measurements. Most notably, the bisection bandwidth test had to be modified so that the throughput of each pair of nodes is added together, rather than simply multiplying the measured throughput of a single pair by the total number of pairs. The same source code was compiled with ScaMPI [16] libraries for the SCI tests.

All experiments were conducted on the same computers with Red Hat 7.2 and kernel version 2.4.7-10smp. Each node in the testbed consists of the following hardware: dual 1GHz Intel Pentium-III processors, ServerSet III LE chipset and 133MHz system bus, 256MB PC133 SDRAM, and a Dolphin D335 64-bit, 66MHz PCI-SCI interface adapter with a daughter card for 2D topologies.

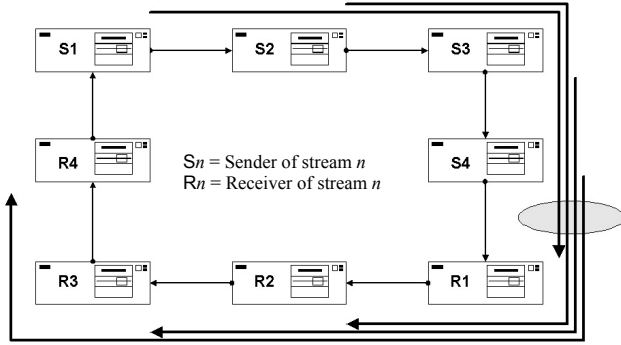


Figure 5. SCI link saturation test; all messages pass through same link to reach their destinations

To experimentally measure the maximum throughput possible for an SCI link, we need to send enough packets across a single link so that it becomes saturated. By grouping senders and receivers as shown in Figure 5, enough traffic will be generated to saturate the link between the last sender (S4) and first receiver (R1). The experimental maximum throughput of an SCI link is equal to the aggregate throughput of all the nodes. A standard ring topology is used to eliminate the possibility of the B-Link being saturated first by packets switching dimensions. Results of the test are shown in Figure 6. Notice that the aggregate throughput of two senders is twice that of a single sender, but when the third is added, the link is saturated and throughput is limited. Adding another sender makes almost no difference to aggregate throughput, which is measured to be 390 MB/s or about 88% of the 444 MB/s theoretical value of effective bandwidth for a single link.

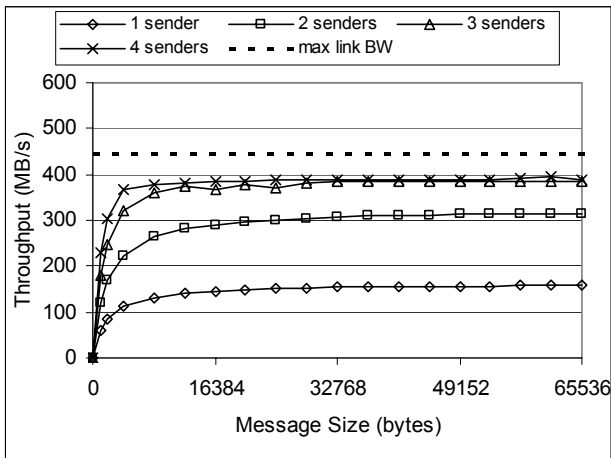


Figure 6. SCI link saturation test; throughput when all messages cross same link on the ring

The next experiment is designed to saturate a single B-Link bus and again measure the maximum throughput. Using the same approach as before, senders and receivers are chosen such that all request and response packets will

be switched between dimensions, from one ring to another, by the same node and thus the same B-Link.

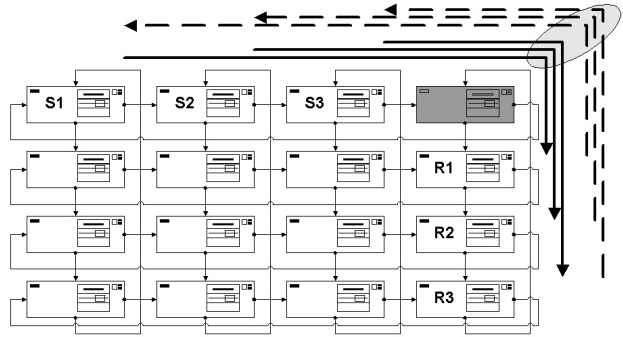


Figure 7. SCI B-Link saturation test; all messages switch dimensions through B-Link of same node

In Figure 7, every packet switches rings through the node in the upper-right corner. As traffic increases, this node's B-Link becomes saturated and aggregate throughput is limited as shown in Figure 8. The maximum throughput for the B-Link is measured at 360 MB/s, or approximately 84% of the 427 MB/s value of effective bandwidth that is theoretically available.

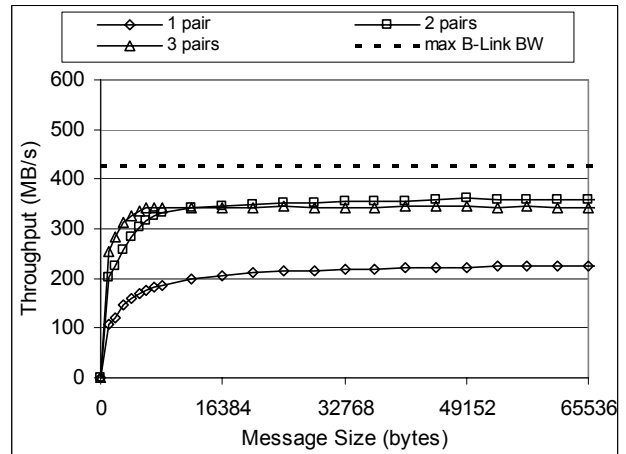


Figure 8. SCI B-Link saturation test; throughput when all messages switch rings through same B-Link

4. Myrinet analysis

In this section, the basic characteristics of Myrinet will be analyzed to support the development of a throughput model. Afterwards, the maximum link throughput will be determined experimentally and compared with the theoretical limit.

4.1. Analytical investigation

We again look at all-to-all communication and calculate the total number of packets that must pass

through each link, for several Myrinet topologies. As any possible way of connecting Myrinet switches and adapters is allowed [14], there are numerous possible Myrinet configurations. We have narrowed down the scope to several different topologies. The first is a minimally connected network, which is the least-expensive method of connection, but provides poor bandwidth between switches. Using 16-port switches, each switch is connected to another switch with one link, forming a linear array of switches. On the other end of the spectrum, we consider Myricom's recommended topology, a Clos network, built with Clos64 "network in a box" enclosures. Each Clos64 contains sixteen 16-port crossbar switches, and will connect 64 nodes to each other and to another Clos64 while still providing full link bandwidth to all. The last topology considered is a compromise between the previous two, where the switches are connected together to form a ring with multiple links between each adjacent pairs of switches.

Table 2. Gross bandwidth of Myrinet links and buses

Bus	Width (bits)	Frequency (MHz)	Gross BW (MB/s)
PCI	64	66	533
Myrinet link	8	80 (DDR)*	160
Myrinet RAM	64	133	1067
* uses rising and falling clock edges			

Each of the Myrinet nodes is connected to a switch by a dedicated, full-duplex link. Therefore, all nodes have the full link bandwidth available to them via a port on the non-blocking crossbar switch to which they are directly connected. The available bandwidth per node is restricted, though, when multiple switches are connected together with too few links. We must calculate the total number of packets crossing these links to find out the maximum throughput per node during all-to-all communication.

First, we assume a fairly symmetric network, where all switches are connected to the same number of nodes and switches. Let N be the total number of nodes and S be the total number of switches. For each switch, let j be the number of ports connected to switches and let k be the number of ports connected to nodes. A general form for the total number of packets passing through each switch-to-switch link is

$$hot-link_{Myri} = \frac{N(N-k)(h_{sw2sw})}{(j \cdot S)} \quad (6)$$

where h_{sw2sw} is the average number of hops between switches. For the minimally connected topology, the average hops between switches is

$$h_{sw2sw_{mconn}} = \begin{cases} \frac{\sum_{i=0}^{S-1} 2i(S-i)}{S(S-1)}, & S = 2 \\ \frac{\sum_{i=0}^{S-1} 2i(S-i)}{(S-1)^2}, & S \geq 3, S \in \mathbb{Z}^+ \end{cases} \quad (7)$$

The Clos64 interconnect provides full bisection bandwidth. So, for calculating the hot-link, the average hops is considered to be 1 for this Myrinet topology. For the ring of switches, the average hops between switches is

$$h_{sw2sw_{ring}} = \frac{\sum_{i=0}^{S-1} \left\lceil \frac{i}{2} \right\rceil}{S-1}, \quad S \geq 2, S \in \mathbb{Z}^+ \quad (8)$$

Overhead in a Myrinet packet is dependent on the number of hops the packet must make, rather than depending on the message size, as is the case with SCI. Therefore, packet efficiency ranges from poor for small messages to very good for large messages. Some example efficiencies are shown in Table 3.

Table 3. Myrinet packet overhead

Packet overhead	Number of bytes
Packet type	4
CRC-8	1
Source Route (1 byte/switch on route)	$h_{sw2sw} + 1$
Cumulative overhead for any size message	$h_{sw2sw} + 6$

Based on the values given in Table 3, for Myrinet packets the efficiency is calculated as

$$Efficiency = \left(\frac{M}{M + (h_{sw2sw} + 6)} \right) \quad (9)$$

where M is the message size in bytes.

The available amount of switch-to-switch bandwidth is found as the number of packets sent by a node that must travel switch-to-switch, divided by the hot-link, multiplied by the effective bandwidth.

$$B_{sw2sw} = \left(\frac{N-k}{hot-link_{Myri}} \right) B_{eff} \quad (10)$$

Notice that Eqs. 6 and 10 are only counting the $N-k$ packets that are destined for another switch. The remaining $k-1$ packets to the nodes on the same switch

still have full link bandwidth from source to destination as given below in Eq. 11.

$$B_{sw} = \left(\frac{k-1}{hot-link_{Myri}} \right) B_{eff} \quad (11)$$

Thus, the overall available bandwidth per node is

$$B_{avail} = \left[\frac{(N-k)B_{sw2sw} + (k-1)B_{sw}}{N-1} \right] \quad (12)$$

The bandwidth available with switch-to-switch communications on Myrinet systems is shown in Figure 9. Here, the available bandwidth decreases as the number of nodes per switch increases for all cases, except of course for the Clos64 where full bisection bandwidth is always maintained. For the minimally connected configuration and the switch rings, it is noted that the increases in switch-to-switch bandwidth occur when an additional switch is added, but not all ports are connected to nodes yet.

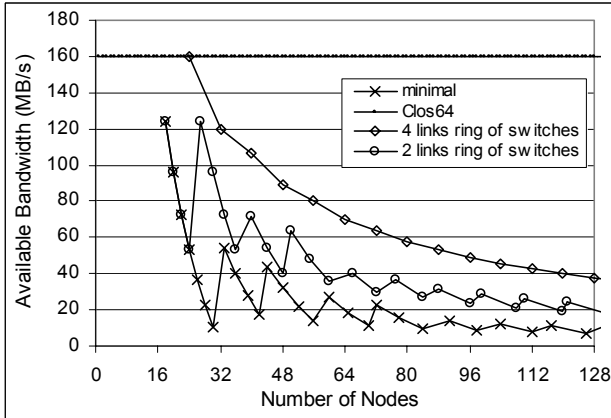


Figure 9. Myrinet switch-to-switch available bandwidth per node

Figure 10 provides data on the overall bandwidth that is available for each of the Myrinet configurations. Here, switch-to-switch bandwidth is averaged with the same-switch bandwidth to find the overall available bandwidth per node.

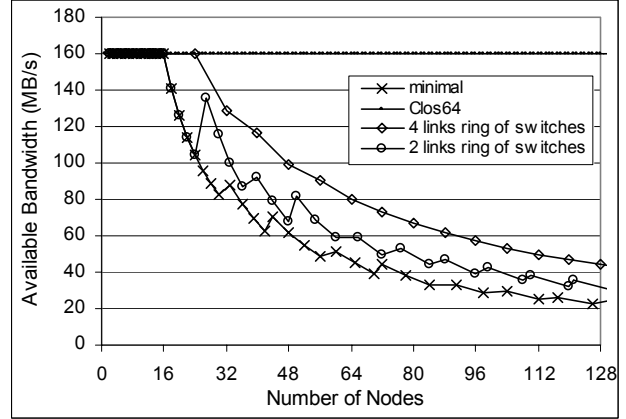


Figure 10. Myrinet overall available bandwidth per node

4.2. Experiments and results

A Myrinet experiment is conducted to obtain the practical saturation point of a Myrinet link. The same test program that was used for the SCI experiments is also used for the Myrinet experiments. The source code was compiled with MPICH-GM 1.2.1..7b [17] libraries for operation over the Myrinet 1280 (M2L-PCI64A-2) 64-bit, 66MHz PCI host interfaces. The rest of the Myrinet testbed consists of the same hardware, operating system, and kernel used in the SCI experiments.

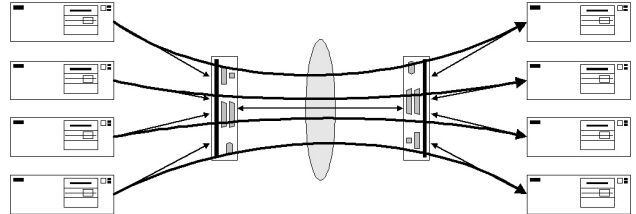


Figure 11. Myrinet link saturation test; two switches connected by only one link

Throughput is measured for four senders that are connected through two Myrinet switches as shown in Figure 11. A single link between the two switches provides minimal connectivity. Packets from the four senders saturate this link. Results of the throughput measurements are shown in Figure 12.

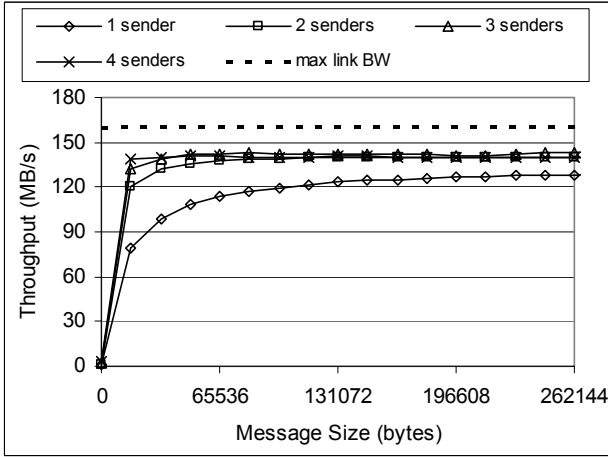


Figure 12. Myrinet link saturation test; two switches connected by only one link

A maximum throughput of 144 MB/s is achieved when all messages are forced through one link. This measured value is 90% of the theoretical peak of 160 MB/s (i.e. the 1.28 Gb/s base data rate of the network).

5. Projections and comparison

The measured bandwidths of SCI and Myrinet links are substituted for the theoretical values in the analytical models for available bandwidth per node. The calibrated models more closely represent the actual throughput that will be seen on a real system. The chart in Figure 13 shows the calibrated available bandwidth per node for SCI and Myrinet systems.

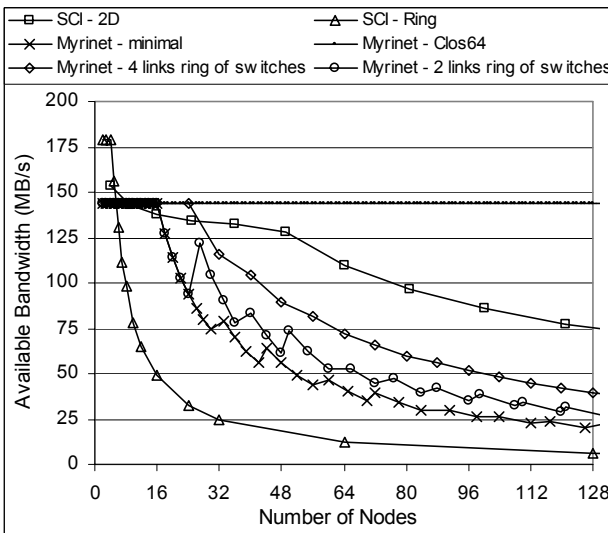


Figure 13. Projected available bandwidth per node

As can be seen from Figure 13, for systems of four nodes or less, a simple SCI ring provides the highest throughput. However, as more nodes are added to the ring, the available bandwidth decreases exponentially.

SCI then offers the most bandwidth when used in 2D torus topology. After nine nodes, SCI throughput falls below what is offered by a Myrinet Clos64 system, which remains the leader for larger systems due to its high bisection bandwidth. After the Clos64, 2D SCI provides the next highest bandwidth for systems of 25 nodes or greater.

It's important to remember, when looking at this chart, that Myrinet switches provide full link bandwidth to all nodes connected to the same switch. If a parallel program only needs to communicate with processors connected to the same switch, then each node could achieve up to the full Myrinet link bandwidth when executed on any Myrinet system, regardless of switch topology. Of course, when a program must communicate with all other nodes in the system, throughput will be limited to approximately the switch-to-switch bandwidth shown in Figure 9.

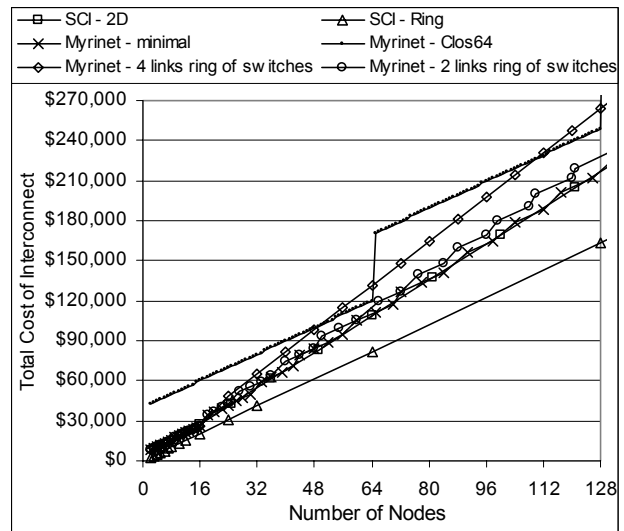


Figure 14. Total dollar cost versus system size

Bandwidth is not the only consideration when choosing a SAN. Figure 14 shows the impact of the costs for the SANs discussed so far using advertised prices sampled at the time of this research. An SCI ring is the least expensive, but is not a good option past about 8 nodes because of its severely limited bandwidth. The total cost for an SCI 2D torus is nearly identical to the cost of a minimally connected Myrinet system. However, as shown previously, the 2D torus provides considerably better bandwidth for systems of all sizes when compared lowest-cost Myrinet systems with only minimal connectivity.

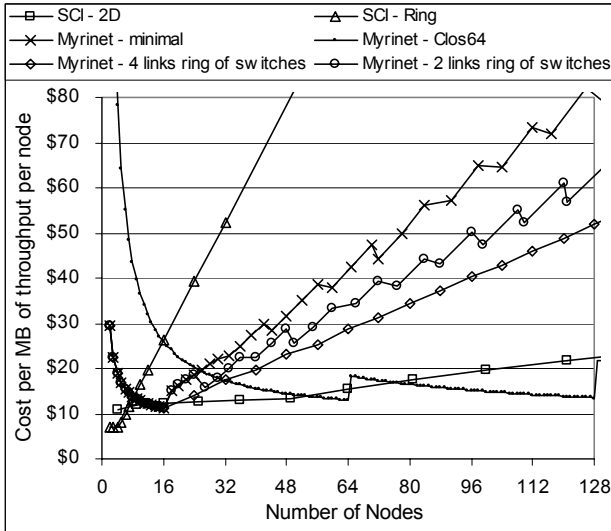


Figure 15. Cost effectiveness in per-node throughput

The Myrinet Clos64 option is near the top of the price range. For systems of 64 or 128 nodes, the Clos64 costs little more than the other options, and provides more bandwidth. For small systems, the initial price is very high. But the Clos64 may still be a good choice if future upgrades are planned. Figure 15 shows how many dollars each megabyte of available bandwidth per node will cost.

In general, SCI 2D torus or Myrinet Clos64 systems provide the most available bandwidth for the money. SCI is more cost effective for smaller systems up to approximately 64 nodes. Myrinet takes over for larger systems of approximately 64 nodes or more.

As shown earlier in Figure 4, a 3D torus starts to outperform a 2D torus at about 64 nodes or greater. Therefore, three-dimensional SCI would likely be very competitive with Myrinet Clos64 for moderately large systems, assuming the cost of a 3D SCI adapter is not significantly more than that of a 2D adapter.

6. Conclusions

In this paper, the throughput and the overall dollar costs of two high-performance SANs have been analyzed. Experiments to measure the actual link speed have been conducted on both interconnects. With the use of the calibrated models, different size and topology Myrinet and SCI systems were compared in terms of available bandwidth and price. Although these are not the only important issues to consider when choosing an interconnect, they are among the key factors.

The results of this research help support a fair and accurate comparison and decision when choosing either SCI or Myrinet as a system interconnect for a cluster. Insight on the best topology options for systems of various sizes is provided, helping designers to decide if a Clos64 switch is worth the high initial price, for example.

This work can also be helpful when considering the different upgrades paths for current systems, such as how the throughput will be affected if a system connected with an SCI ring is rewired as a 2D torus.

This research could be continued in several possible directions. An obvious option is to examine and model latency as a function of system size and topology. The resulting comparisons based on latency, throughput, and cost could be very valuable when choosing an interconnect. Also, the models and experiments can be updated to include new technology, such as Myrinet 2000 and the three-dimensional SCI Wulfskit. Another direction is to expand the models with some more exotic topologies, especially for large-scale systems. Recently, very large Myrinet systems have been built by connecting many 16-port switches in a three-dimensional torus configuration. There could also be significant advantages for irregular topologies that are partitioned into groups, offering very high throughput between nodes of the same group, but much less between nodes of different groups. Such a system could provide very good performance and reduced cost as long as work is split up well to take advantage of different degrees of parallelism.

7. Acknowledgements

The support provided for this research by the U.S. Department of Defense is acknowledged and appreciated, as is equipment support provided by Dolphin Interconnect and Scali (SCI), Sandia National Labs (Myrinet), and Nortel Networks.

8. References

- [1] M. Ibel, K. Schauer, C. Scheiman, M. Weis, "High-Performance Cluster Computing Using Scalable Coherent Interface," *IEEE Communications*, Aug. 1996, pp. 52-63.
- [2] K. Omang, B. Parady, "Scalability of SCI Workstation Clusters, a Preliminary Study," *Proc. of 11th International Parallel Processing Symposium (IPPS'97)*, Apr. 1997, pp. 750-755.
- [3] G. Horn, "Scalability of SCI Ringlets," in: H. Hellwagner, A. Reinefeld (Eds.), *SCI: Scalable Coherent Interface, LNCS State-of-the-Art Survey*, Springer, Berlin, 1999, pp. 151-165.
- [4] H. Bugge, "Affordable Scalability using Multicubes", in: H. Hellwagner, A. Reinefeld (Eds.), *SCI: Scalable Coherent Interface, LNCS State-of-the-Art Survey*, Springer, Berlin, 1999, pp. 167-174.
- [5] M. Sarwar and A. George, "Simulative Performance Analysis of Distributed Switching Fabrics for SCI-based Systems," *Microprocessors and Microsystems*, Vol. 24, No. 1, Mar. 2000, pp. 1-11.
- [6] D. Gonzalez, A. George, and M. Chidester, "Performance Modeling and Evaluation of Topologies for Low-Latency SCI Systems," *Microprocessor and*

- Microsystems*, Vol. 25, No. 7, Oct. 2001, pp. 343-356.
- [7] C. Kurmann and T. Stricker, "A Comparison of Two Gigabit SAN/LAN Technologies: Scalable Coherent Interface versus Myrinet," *Proc. of SCI Europe '98 Conference, EMMSEC'98*, Sept. 1998.
 - [8] J. Hsieh, T. Leng, V. Mashayekhi, and R. Rooholamini, "Architectural and Performance Evaluation of GigaNet and Myrinet Interconnects on Clusters of Small-Scale SMP Servers," *Proc. of IEEE Supercomputing (SC'2000)*, Dallas, USA, 2000.
 - [9] R. Brightwell and S. Plimpton, "Scalability and Performance of Two Large Linux Clusters," *Journal of Parallel and Distributed Computing - Special Issue on Cluster and Network-Based Computing*, Vol. 61, No. 11, Nov. 2001, pp. 1546-1569.
 - [10] H. Bal et al., "The distributed ASCI supercomputer project," *ACM Special Interest Group, Operating Systems Review*, Vol. 34, No. 4, Oct. 2000, pp 76-96.
 - [11] A. George and R. VanLoon, "High-Fidelity Modeling and Simulation of Myrinet System Area Networks," *International Journal of Modeling and Simulation*, Vol. 21, No. 1, Jan. 2001, pp. 40-50.
 - [12] IEEE, SCI: Scalable Coherent Interface, IEEE Approved Standard 1596-1992, 1993.
 - [13] Scali Computer AS, Scali System Guide Version 2.1, White Paper, Scali Computer AS, 2000.
 - [14] Myrinet-on-VME Protocol Specification, ANSI/VITA 26-1998, 1998.
 - [15] N. Boden, D. Cohen, R. Felderman, A. Kulawik, C. Seitz, J. Seizovic, and W. Su, "Myrinet: A Gigabit-per-Second Local Area Network," *IEEE Micro*, Vol. 15, No. 1, 1995, pp. 26-36.
 - [16] Scali Computer AS, ScaMPI User's Guide Version 2.1, White Paper, Scali Computer AS, 2000.
 - [17] W. Gropp and E. Lusk, User's Guide for mpich, a Portable Implementation of MPI Version 1.2.0, 1999.