

.....
Pentum Group, Inc.

Dolphin SCI & Sun Remote Shared Memory Speed Oracle9i RAC



*Addressing business IT needs for
availability and scalability forward*

B. K. Parady
bparady@pentum.com



Dolphin SCI & Sun Remote Shared Memory Speed Oracle9i RAC

*Addressing business IT requirements for
availability and scalability forward*

TABLE OF CONTENTS

EXECUTIVE SUMMARY 3

INTRODUCTION 4

FEATURES OF RSM AND SCI 5

- Description of Sun Cluster 5
- Description of RSM 6
- Features of SCI-PCI hardware 7
- Description of Oracle 9i and 9i RAC 8

BENEFITS OF RSM OVER SUN PCI-SCI 11

- Basic latency and bandwidth measurements 11
- Oracle performance profile results from Sun 13
- Results observed by Ericsson using their Alzato NDB cluster database 14

CONCLUSION 15

- About Pentum Group, Inc. 17
- Trademark Notices 17

Executive Summary

Corporate IT must maintain system availability and respond flexibly to rapid growth and change in their enterprise operations. As firms expand and alter their operations, either through growth, expanding new business and IT processes, entering new business areas, or through mergers and acquisitions, IT systems, in particular database systems, must change with them. Since database systems are used for critical business processes, anticipating growth requires a strategy for evolution that allows uninterrupted operations through system upgrades, as well as scalability to meet new QOS requirements, and all of this while seeking the best protection and return on IT investment.

The Sun, Oracle, and Dolphin Interconnect partnership offers a complete system for Oracle9i Real Application Clusters (RAC) providing both scalability and availability in a single, easy to manage database configuration. Real Application Clusters software and the collection of hardware known as a cluster unite the processing power of each component to create a robust computing environment.

In the Hurwitz Group study¹ on total cost of ownership, 60% of firms identified scalability or availability as a principal infrastructure concern. Over three years, the study projected 18% savings in the total cost of ownership through best practices in scalability and availability. This paper examines the business advantages that the Dolphin/Sun PCI-SCI (Scalable Coherent Interface) interconnect delivers to Oracle9i Real Application Clusters and how the business investment is protected into the future.

Sun Cluster 3.1 software for remote shared memory (RSM) using Dolphin PCI-SCI is tuned for top performance of Oracle9i RAC. This in turn provides improved service for distributed applications running Sun Clusters software and provides excellent flexibility and capacity for system growth. The RSM API offered with Solaris 9 provides application developers with a means to bypass of the TCP/IP stack while providing tight, reliable, direct access to the high-speed, high-bandwidth and low-latency Dolphin SCI interconnect for fast messaging in Sun Clusters.

By fully integrating and engineering Oracle9i RAC Solaris-based cluster systems, Dolphin, Sun and Oracle assure that enterprise operations are available, flexible and scalable well into the future.

¹ “Oracle9i Real Application Clusters,” *An Oracle Business White Paper, August 2002.*

Introduction

Combining the reliability of Solaris and Oracle software with Sun and Dolphin hardware creates a solid technology base for Sun's Oracle9i RAC clusters. Key to the software technology is support for remote shared memory on the Dolphin SCI interconnects.

With this technology base, the benefit of performance can be measured and quantified. Current customer satisfaction and future performance, capacity, reliability and availability are assured through Sun PCI-SCI based clusters.

Before focusing on various aspects of the Oracle9i RAC and Dolphin database cluster, the definition of a cluster deserves discussion. A cluster can be made up of multiple single-processor systems or nodes, or multiple multiprocessor systems or nodes that share memory only locally (SMPs). Every node has local memory as well as a copy of the operating system, a database instance or process, and application software. Figure 1 shows a cluster at its simplest.

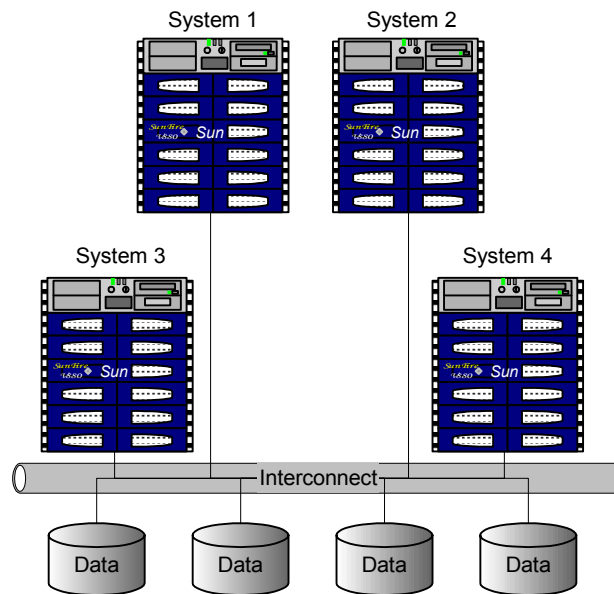


Figure 1: Simple representation of a database cluster

Features of RSM and SCI

There is a lengthy and broad technical background supporting Oracle 9iRAC on Sun clusters, RSM, and Dolphin SCI interconnect technology.

Description of Sun Cluster

Sun Cluster 3.1 is Sun's next-generation clustering technology, focusing on integrated availability, scalability, and manageability of the service delivery platform. Sun Cluster 3.1 provides a single management view for all these services, enabling the cluster to be administered as a whole. The cluster can scale in capacity, while incremental management costs are minimized. System administration tasks such as data backup and restore, installation of patches, software upgrades, and adding new hardware can be done without interrupting service delivery.

Sun Cluster 3.1 software enables applications to scale both by using larger servers and by using multiples of them concurrently. Individual cluster nodes still run separate copies of Solaris, providing fault isolation.

The high availability delivered by Sun Cluster 3.1 is provided through a combination of well-managed hardware and software components. Availability is assured by configuring the cluster without any single points of failure throughout the system, including public networks, cluster interconnect, storage and software service components. Data integrity is maintained in the cluster by allowing only fully operational nodes to deliver services. The condition of each individual node, along with its hardware and software components, is constantly monitored. Failing or failed nodes are prevented from delivering services and accessing data. Failed software components can be restarted within the cluster, and failed nodes may be returned to the cluster following repair.

The Sun Cluster extensions to the Solaris operating system are designed to provide a continuously available platform, enabling highly available or continuously available services. In contrast, fault-tolerant hardware systems also provide constant access to data and applications, but at greater cost because more specialized hardware is required. However, fault tolerant systems typically are not designed to deal adequately with software failures.

Failover is the process by which the cluster automatically relocates an application from a failed node to a healthy one. **Scalability** enables a service to meet increasing load requirements, while delivering the same quality of service. A scalable application service employs the multiple nodes in a cluster by running multiple instances of the same application service concurrently on multiple nodes. An example is Web service, where individual instances can process client requests independently of each other. In cases where application data sets need to be consistent across the cluster, such as a general instance of an Oracle database, the application could be made into a scalable service, but the application must maintain data synchronization among its instances. This is exactly what Oracle9i RAC does.

Sun Cluster 3.1 enables applications to be implemented either as a failover or a scalable service. A failover application provides high availability; a scalable application provides both

high availability as well as increased performance. Both failover and scalable applications can run on the same cluster concurrently.

Description of RSM

The remote shared memory implementation in Sun Cluster 3.1 allows a user process on a given cluster node direct access to the memory on another cluster node. The access to another node’s memory is done using memory based interconnect hardware; in this case Sun Dolphin PCI-SCI. The advantage of RSM operations is that they reduce latency time with an OS bypass implementing zero copy of data. The performance advantages for SCI message passing have been well quantified.² Remote shared memory operations are not available as part of any Ethernet or Internet protocol.

Although applications write to memory on a remote node, the implementation of RSM employs a message passing programming model. These reliable hardware/software mechanisms optimize operations such as lock management, checkpointing or transaction processing. The performance of these functions is directly dependent on the implementation of efficient barriers, which greatly benefit from the inherent low latency of RSM using SCI.³

Applications can speed up communication between nodes by making low-level calls to access the interconnect. E.g., SCI can support native load and store operations, while other hardware may support only block load and store operations. SCI supports both remote direct memory access (RDMA) and parallel input/output (PIO) while some other hardware may support either PIO or RDMA. The basic software model is shown below in Figure 2.

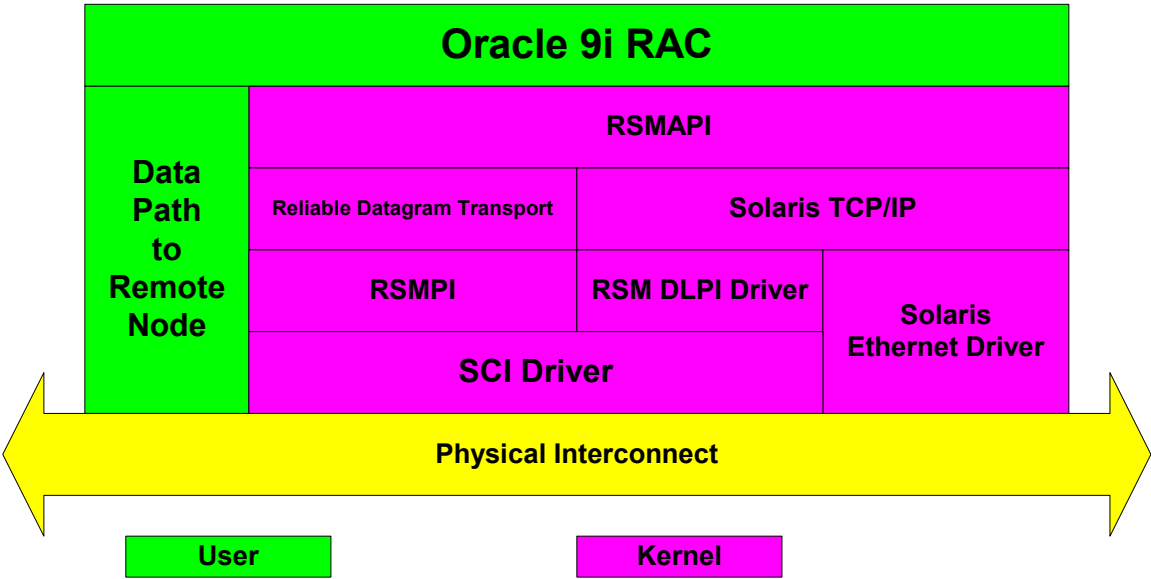


Figure 2: SCI Software Stack

² Joachim Worrigen, F. Seifert & T. Bemmerl, “Efficient Asynchronous Message Passing via SCI with Zero-Copying,” http://www.tu-chemnitz.de/informatik/RA/papers/p2001/zero-copy_sci-mpich.pdf
³ Sohrab Modi, Sun Cluster 3.0 Remote Shared Memory,” at Sun Super-G 2001.

The Sun Cluster Interconnect framework is shown below in Figure 3. Applications needing remote shared memory in Sun Cluster communicate over Solaris RSMAPI. A reliable datagram transport (RDT) driver has been created specifically to optimize Oracle9i RAC over RSM in Sun Cluster. Oracle9i RAC uses RDT to address remote nodes and quickly process port send and port query messages.

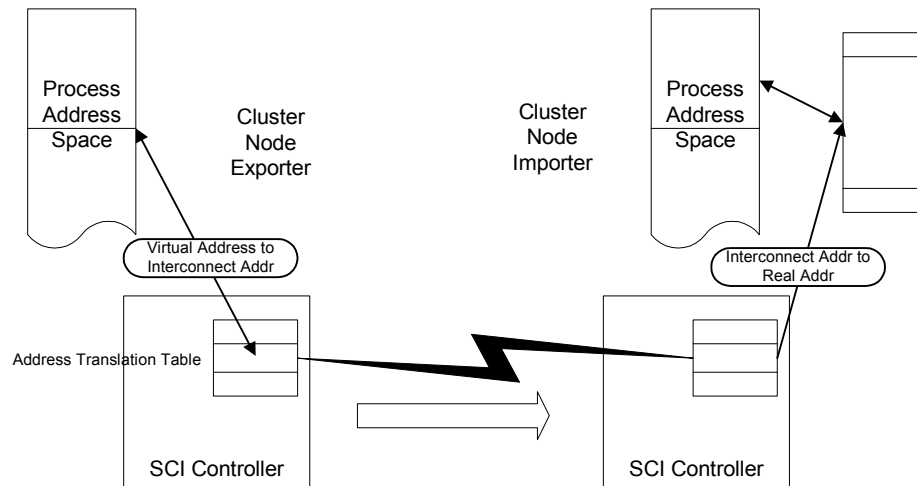


Figure 3: SCI node-to-node memory data transfer

Only reliable cluster interconnects such as SCI can bypass the OS and speed memory data transfers across system boundaries. TCP/IP connections through Ethernet cannot provide this OS bypass to deliver data at near memory access speeds. TCP/IP connections force data handling to use OS services and require software checking for reliable transfer.

Features of SCI-PCI hardware

The Dolphin PCI-SCI cards used by Sun are based on and compliant with ANSI/IEEE 1596-1992 Scalable Coherent Interface. SCI emerged from the effort to define a superbus in the 1980s. SCI defines a distributed solution to hardware and protocols providing a shared-memory view of buses. Although the SCI protocol support for coherent shared memory is not used in the PCI-SCI implementation, the hardware, which is designed for fast memory transactions, yields a superbly conceived cluster interconnect.⁴

SCI's original goals were high performance, high throughput, low latency, low CPU overhead, and scalability. SCI was designed as an *open distributed bus*. Conceptually it is made up of point-to-point links supporting many concurrent data transfers using sophisticated signaling technology. It can be configured into complex networks, from simple ring

⁴ http://www.bode.cs.tum.edu/~gerndt/home/Teaching/scalable_shared_memory_systems/Kapitel8.pdf

topologies to complex rings of rings and multidimensional tori. Other important aspects of SCI are the built-in transactions to read, write, and lock memory locations, which are exploited in the PCI-SCI implementation.

SCI offers unique performance advantages to database systems because it supports both short, 64 bit messages and longer 128 byte messages. That is because in a database system there is a mixture of low latency, short messages (<512 byte) for lock management, and larger, high bandwidth transfers (> 4KB) for data movement and cache management. The ability of SCI to deliver microsecond short messages, and gigabit-to-gigabyte bandwidths, makes it an ideal choice for database cluster interconnections.

Perhaps most important to IT professionals is that reliability is designed into the hardware. There is 16-bit CRC (cyclic redundancy check) based error detection at the hardware level that assures reliable transport.

Clustering two or more database servers requires a private interconnect between them for message passing. Some existing Sun high-end cluster customers may prefer to stay with SCI-SBus due to limited availability of PCI slots. Most customers, however, will want to upgrade to PCI-SCI to gain the advantages of RSM, higher performance and support under Sun Cluster 3.1. It is also worth noting that PCI-SCI is now supported for Workgroup servers, which was not the case for Sbus-SCI.

PCI-SCI consists of a PCI-64 adapter card, a choice of cables in three lengths and a four port SCI switch, soon to be an eight port SCI switch. The SCI switch is required when building a cluster of three or more nodes. For more details please visit the Dolphin website:

<http://www.dolphinics.com/products/hardware/pci64.html>

PCI-SCI offers high availability operation with hot-pluggable connections and redundant PCI-SCI cards. It supports both direct memory access (DMA) and programmed remote memory access (RMA).

Description of Oracle 9i and 9i RAC

Oracle Real Application Clusters harnesses the processing power of multiple interconnected computers. With the increased functionality of Real Application Clusters, all systems and applications can efficiently exploit clustered environments.

A cluster database comprises two or more nodes that are linked by an interconnect. The interconnect serves as the communication path between each node in the cluster database. Each Oracle instance uses the interconnect for the messaging that synchronizes each instance's use of shared resources. Oracle also uses the interconnect to transmit data blocks that the multiple instances share. The principal shared resources accessed by all nodes are the data files. The cluster and its interconnect are linked to the storage devices, or shared disk subsystem, by a storage area network.

In Real Application Clusters environments, all database activities can concurrently execute transactions against a shared database. Real Application Clusters coordinates each active instance's access to the shared data to provide data consistency and data integrity. Dividing a large task into subtasks or running a number of simultaneous tasks and distributing them among multiple nodes has the advantage of finishing the task faster than if the entire task

were processed on a single node. Cluster processing also improves performance for larger workloads and accommodates rapidly changing user populations.

With Real Application Clusters, applications scale to meet increasing data processing demands without the need to change the application code. Data warehouse applications that access read-only data are prime candidates for Real Application Clusters. Among many other applications, RAC successfully manages Online Transaction Processing (OLTP) systems and hybrid systems that combine the characteristics of both read-write and read-only applications. Real Application Clusters also serves as an important component of robust high-availability solutions, tolerating failures with little or no downtime.

The Benefits of Real Application Clusters as listed by Oracle are:

- Lower overall cost of ownership
- Expanded scalability
- High availability
- Transparency

Oracle9i RAC uses a high-speed interprocess communication (IPC) component for internode communications. The IPC defines the protocols and interfaces required for Real Application Clusters environments to transfer messages between instances. Communication in this interface is based on messages. The IPC is based on an asynchronous, queued messaging model. RAC supports user-mode and memory-mapped IPCs. These types of IPCs substantially reduce CPU consumption and IPC latency. These IPCs are not available with Ethernet or Internet protocols.

The scalability of Oracle9i RAC can be strongly dependent on latency, depending on the application. Real Application Clusters synchronizes between nodes, and if the synchronization is slow, then at some point it will limit the scalability of the application. In 9iRAC, scalability of synchronization events is determined by the time required to access physical storage. In order to access physical storage, there must be locked exclusive access to data. Thus, the upper limit for the number of transactions per unit time that can be satisfied is directly proportional to the interconnect latency.

Real Application Clusters also requires that all nodes have simultaneous access to the shared disks to give the instances concurrent access to the database. The implementation of the shared disk subsystem is based on the operating system: either a cluster file system or placing the files directly on raw devices. Cluster file systems greatly simplify the installation and administration of Real Application Clusters.

Oracle9i RAC offers major enhancements in scalability, availability and manageability. The most significant technology breakthrough is the complete implementation of CacheFusion that enables highly scalable applications to be built without any concern for data partitioning. CacheFusion stores data on each cluster node system memory when possible, to reduce the need to access the disk before a data transfer. Requests for data are often satisfied by direct memory access of data from nodes, which is faster than disk by orders of magnitude.

CacheFusion resolves data block read/read, read/write and write/write conflicts among RAC nodes. Data block ping (the repeated request for data on a remote node because of locking)⁵ is resolved through high-performance interconnect networks, bypassing much slower physical disk operations typically used in previous releases. Close to linear scalability of database performance can be achieved when adding nodes to the cluster, provided that a high-performance interconnect is available.

Applications no longer need to be partitioned according to data access patterns to avoid or reduce data block ping. A scalable application on a single-node Oracle server will be just as scalable on a multi-node RAC. Other performance features include dynamic lock re-mastering among nodes to further reduce interconnect traffic for data ping and provide a more efficient inter-node messaging mechanism.

⁵ <http://gwynne.cs.ualberta.ca/~oracle/817doc/paraserv.817/a76968/pslkgdtl.htm#6405>

Benefits of RSM over Sun PCI-SCI

As discussed in the previous section, the performance of the cluster interconnect (both hardware interconnect and software application interface) is key to satisfying the requirement for scalability and performance. Achieving high levels of scalability requires optimizing the right type of performance. This section focuses on measured performance and examines the data.

Basic latency and bandwidth measurements

Measurements of interconnect latency and bandwidth offer the best insight into total and relative benefits of the various available interconnects. For the foreseeable future SCI will continue to be the best choice for low-cost, high-performance cluster interconnect.

Table 1 displays latencies and bandwidth data for various interconnects. The values were derived from standard application-level message passing MPI API calls using Sun Cluster 3.1. Although Oracle9i RAC does not use MPI, the table offers insight into the benefits to be seen by Oracle from high-speed cluster interconnects.

Table 1: Sun Cluster Interconnect Performance

Interconnect	Gigabit Ethernet	PCI-SCI (Sun D320)	SunFire Link
Driver	DLPI	RSM	RSM
H/W Latency	60 μ s	3.6 μ s	1.7 μ s
MPI Latency	100 μ s	9 μ s	3.7 μ s
1-way Bandwidth	60 MB/s	120 MB/s	1.0 GB/s
2-way Bandwidth	110 MB/s	200 MB/s	1.0 GB/s
Connection Type	PCI 64 bit 66 MHz	PCI 64 bit 33 MHz	Proprietary

Note that the RSM (Remote Shared Memory) driver is not available for Gigabit Ethernet.

These results demonstrate excellent system-to-system interconnect latencies for SCI. SCI cluster interconnects offer latencies near that of Sun's top proprietary interconnect at costs near that of Gigabit Ethernet, making PCI-SCI an excellent performance/value choice. Bandwidth for SCI also exceeds that for GBE, and with the support of dual SCI NICs for failover and performance, SCI offers excellent bandwidth capacity. Deploying a second PCI-

SCI interconnect on a separate system PCI bus offers twice the interconnect hardware bandwidth and maintains excellent point-to-point latencies.

Looking forward, the SCI link hardware supports much higher bandwidths than the current PCI bus can support.

Table 2: Dolphin SCI hardware performance capability

Dolphin Component	Latency	SCI Link Speed	Bandwidth
Dolphin Sun D320 Measured	3.60 μ s	2x400 MB/s	120 MB/s 1-way 200 MB/s 2-way
Dolphin D331 66/64 PCI ⁶ Measured	1.46 μ s	2x667 MB/s	326 MB/s 1-way 385 MB/s 2-way
Dolphin D336 66/64 PCI ⁷ 3D mesh	1.46 μ s	6x667 MB/s	326 MB/s 1-way 385 MB/s 2-way
Dolphin MS8X D535 8 port SCI switch ⁸	180 ns	16x667 MB/s	3.2 GB/s
Dolphin LC-3 Link chip ⁹	70 ns	2x667 MB/s	1.33 GB/s

With the large bandwidth available to SCI through the Dolphin link chips, SCI can support future cluster interconnect bandwidth requirements beyond 10 Gbits/s. This preserves the hardware and software technology investment in the cluster because SCI can handle increases in future IO bus bandwidths, thus providing future state-of-the-art cluster performance.

⁶ <http://www.dolphinics.com/products/hardware/pci64.html>

⁷ http://www.cse.clrc.ac.uk/disco/mew/Talks/Lochsen_Dolphin.pdf

⁸ <http://www.dolphinics.com/products/hardware/ms8x.html>

⁹ <http://www.dolphinics.com/products/hardware/lc3.html>

Oracle performance profile results from Sun

Oracle makes available a series of test probes of its database to ascertain performance characteristics of the database runtime environment. CRTest, which exercises the interprocessor communications code path, was used to run the following tests. There are also additional test probes available with the Statspack tools which were not used.¹⁰

Sun Microsystems used the Oracle Statspack tools to determine an Oracle9i RAC database benchmark on a two node cluster using E6800 servers, each configured with eight 750 MHz SPARC III processors. An average 4KB message size was observed in the running of the benchmark, and the following table reflects latencies and bandwidths for message sizes of that length.

Table 3: Relative Performance, 4KB message length

Interconnect	Latency, relative	Bandwidth, relative
Sun SCI-PCI	.71 (-29%)	1.33 (+33%)
Gigabit Ethernet	1.00	1.00

While Table 1 showed a greater difference in raw performance between SCI and GBE, the difference here is more indicative of the actual performance advantage a real system would see from SCI at the database application level. Additionally, SCI has much lower CPU utilization, which adds up to a strong cost advantage in that fewer processors are needed to support the bandwidth provided by a fully loaded interconnect. The much lower CPU utilization of SCI is due to the support for direct memory transfers that entirely bypass the operating system.

SCI offers better performance, better future performance, and cost advantages compared with the alternatives. Because of the excellent, low internode latency provided by SCI, cluster scalability is assured. This is good news to customers who may eventually need increases in capacity far beyond their starting configurations.

¹⁰ <http://www.oracle.com/oramag/oracle/00-Mar/index.html?o20tun.html>

Results observed by Ericsson using their Alzato NDB cluster database

Ericsson's real-time NDB Cluster database benefits from a fast memory-mapped network interconnect, SCI. While Ericsson's measurements were obtained using a different database than Oracle9i RAC, they offer direct evidence of the SCI performance gains for a database application, as compared to Gigabit Ethernet.

Ericsson observed major benefits in performance; network redundancy/failover and scalability for the Alzato database cluster using SCI.¹¹ SCI offered very high performance, low latency and high bandwidth. Failover is managed at the application/user level, not at the driver level, thereby trapping transfer errors earlier.

The performance comparison is between SCI and TCP/IP as the communication mechanism in the NDB Cluster. Results are shown in Table 4, which shows the performance advantage of the SCI interconnect.

Table 4: Latency for various message lengths

Data packet	Cost of TCP/IP μ s	Cost of SCI μ s
100 Bytes	40	6
1 Kbyte	130	33
10 Kbytes	1030	303

Scalability: Ericsson found that TCP/IP does not scale for clusters with more than 4 to 8 nodes, when used for their database applications.

Failover Performance: Of most interest to IT specialists is the fact that with TCP/IP links, which are OS and driver dependent, total failover times were in the range of 100,000 μ s. Compare this to the time to complete a failover (dropping a link or switch in a redundant configuration) using the SCI interconnect, which was completed in the range of 100-200 μ s.

¹¹ Johan Andersson, **NDB Cluster with SCI**, Alzato, Ericsson Business Innovation, 2002.

Conclusion

This paper reviews the architecture and performance of PCI-SCI for Sun Cluster and indicates the best practices available for scalability, availability, latency and throughput when using Oracle 9iRAC with RSM. Here are some of the conclusions that can be drawn:

General:

- For clusters of intermediate and large Sun servers requiring bandwidths under 2 GB/s, SCI offers the best latencies, bandwidth and by far the best price/performance.
- For high-end systems requiring higher bandwidth, SunFire Link is suggested.
- For smaller systems such as the SunfireV1280 or F6800 or smaller, SCI is the clear choice.
- SCI applied to clustering is more cost effective than GBE, when CPU utilization and total cost of ownership are considered.
- For running RSM on larger, multinode clusters, greater than 4 nodes, the Dolphin D535 switch, which is currently being qualified by Sun, makes SCI the only possible choice in the near term.

Better Database Latencies: To reduce database latencies, the choice of SCI over GBE is a clear one. Explicitly, for the 100 μ s latencies seen with GBE, there would be on the order of 100,000 additional clock cycles (1 GHz processors) consumed for every critical exchange (program locks and synchronizations) between cluster nodes. This could seriously affect database performance and CPU load since, for 100 μ s, the cycles needed to take a context switch would approach that for a polled lock.

Remote Checkpointing: The same reasoning applies for remote checkpointing where it is imperative to release the System State from checkpointing as soon as possible. For those systems where ultra-high reliability, real-time or near real-time response and failover is needed (such as network routing clusters or wireless routers), SCI is the clear choice for medium to small Sun configurations.

All in all, though GBE looks less expensive, it can increase costs substantially as clustered systems will be limited in scalability by the much higher latencies of GBE.

Maximum Reliability: To minimize the risk from any single point of failure, dual SCI interface cards and dual switches are recommended. Such a configuration is fully supported by Sun Cluster 3.1. Deploying dual interfaces and switches has the added benefit of providing double the bandwidth and performance for day-to-day operations while maintaining excellent latency. The following diagram in Figure 4 shows a typical four-server, fully redundant configuration.

With Sun Cluster 3.1 expanding the role of RSM to include GFS (Global File System), and broader support for applications, it makes sense to put as much state-of-the-art performance as

possible into your cluster system in support of RSM. Dolphin SCI is best of breed in latency, and offers excellent bandwidth and proven technology at modest cost.

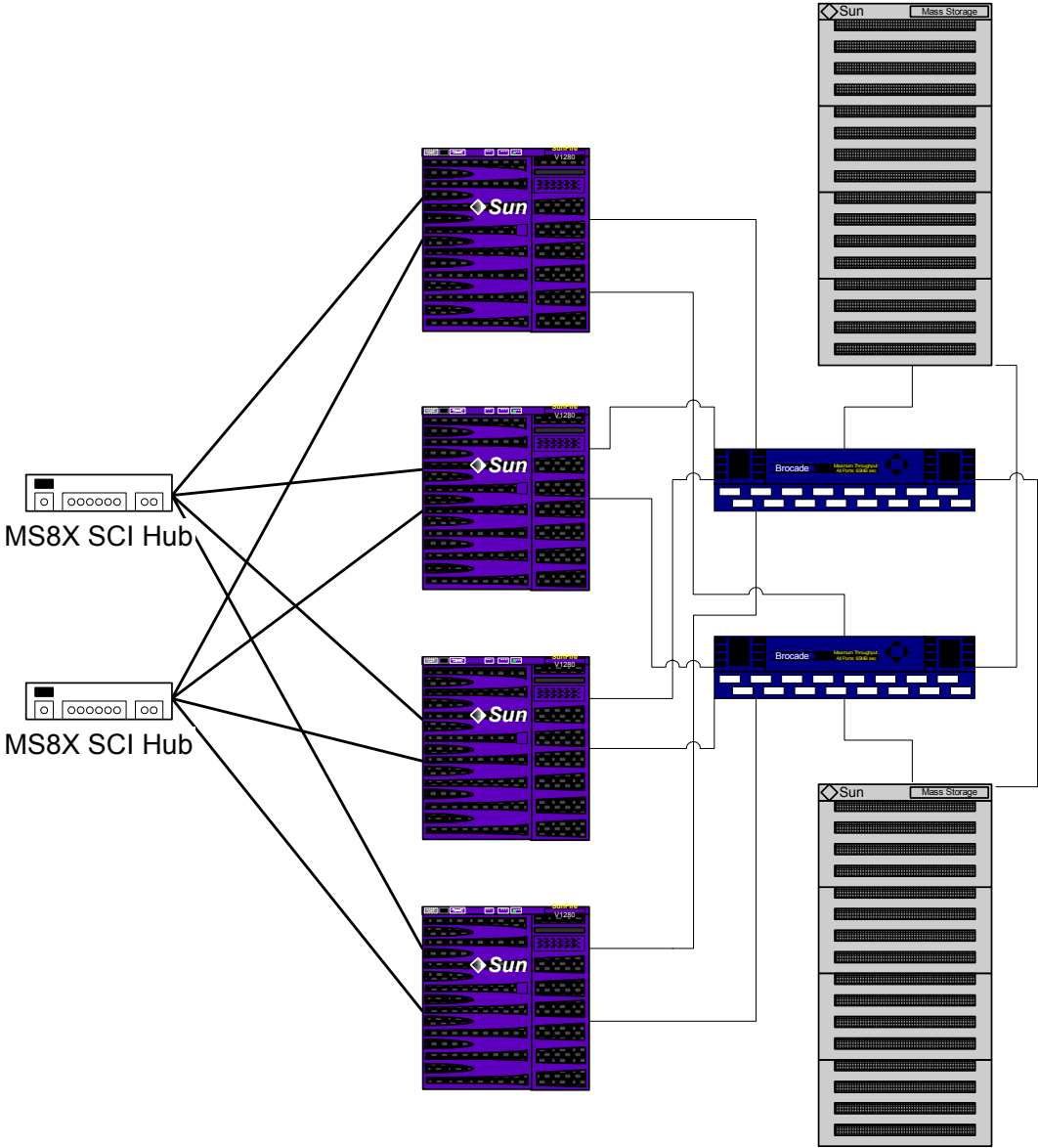


Figure 4: Recommended SCI & SAN failover cluster configuration

The ability to increase the speed of moving database locks across a cluster interconnect is a needed step forward. The increased efficiency of the Sun RSM and Oracle9i RAC combination will provide faster database operations to Oracle users in IT environments implementing Sun Cluster 3.1.

The Sun and Dolphin strategic alliance assures existing and future Oracle9i RAC clients of the commitment to and availability of current and emerging leading-edge technologies essential to the success of their operations. Sun and Dolphin Interconnect are ready to provide proven and cost-effective connectivity packages to current and new clients wishing to establish or expand database capabilities.

About Pentum Group, Inc.

Pentum Group, Inc. provides core technology and senior technologists to help clearly define the impact of technology and its ability to improve and accelerate critical organization missions.

Dr. Bodo Parady (CTO of Pentum Group, Inc.) is a 25-year veteran in computer system technology development. He currently leads Pentum Group and its clients in the development of new markets for high technology, both in ultra high-speed networks and in product marketing for clients.

Trademark Notices

Sun, Solaris and Sun ClusterOS are trademarks of Sun Microsystems, Inc. in the United States and other countries.

Oracle is a registered trademark and Oracle9i is a trademark or registered trademark of Oracle Corporation.

Dolphin is a trademark of Dolphin Interconnect Solutions, AS.

All other product names are for informational purposes only and may be trademarks or registered trademarks of their respective companies.

© 2003 Pentum Group, Inc. All rights reserved.