White Paper

SGI® NUMAlink™
Industry Leading Interconnect Technology

Table of Contents

1 Introduction

The SGI NUMAlink network is the system interconnect used within SGI® cache-coherent NUMA (ccNUMA) compute servers. It is highly differentiated from other system interconnects to minimize latency, while providing the extremely high bandwidth and reliability required for High Performance Computing. Paramount among these requirements, is the ability to directly access large cache-coherent memory address spaces with high bandwidth and low latency while maintaining a low bit error rate. NUMAlink was designed with these requirements in mind. From the physical layer where high speed, low latency electrical transfer is accomplished to the routing layer where messages are transferred with low overhead and latency, NUMAlink is optimized for high performance computing.

This paper will cover two topics. First, a brief review of the evolution of NUMAlink technology is given. This begins with the first NUMAlink generation named NUMAlink™ 2 through the current product in the market place NUMAlink™ 4 and the next generation NUMAlink under development, NUMAlink™ 5. Second, this paper will provide an overview of the functional distinctions between NUMAlink and other competing interconnects in the industry such as InfiniBand, Myrinet® and Quadrics®. This will serve to underscore the unique capabilities of NUMAlink and to show how it is a clear leader in the industry.

2 Evolution and Technology Innovation Behind SGI NUMAlink

NUMAlink 2 was designed for use in SGI's Origin® 200 and Origin® 2000 family of compute servers. At the physical layer, NUMAlink 2 operates at a switching rate of 400Mb/s across 20 data bits/direction. The NUMAlink 2 bandwidth is 0.8GB/s/direction. A source synchronous clock is utilized to capture data in parallel at the receive end in favor of higher overhead clock encoding/extraction schemes. The SERDES (serializer/deserializer) functionality was optimized for minimal latency. One example is in the serializer section where computed link layer sideband bits can be latched late in the serialization cycle rather than having to wait until the beginning of the next cycle. The physical layer supports up to 24" of FR4 printed circuit board (PCB) interconnect, 3 meters of twin-ax based cables and organic ASIC packages. The NUMAlink 2 physical layer latency including 24" of PCB and 3 meters of cable totals 55 ns. The link layer features the robust 16 bit CRC-CCITT (a standard algorithm) to protect against data errors and all detected errors are corrected through a Go back N Automatic Repeat Request protocol. Efficient 2 level router look-up tables are used to route local and global routes. A 6 port NUMAlink 2 router was designed to achieve processor scalability up to 512 sockets in a globally addressable, single system image. NUMAlink 2 was later used at the foundation for the industry standard GSN interconnect.

NUMAlink 3 was designed for use in SGI's Origin® 300 and Origin® 3000 family of servers and supported the early introduction of SGI's newest Altix® line of servers. NUMAlink 3 leveraged much of the circuit architecture and feature set of NUMAlink 2. At the physical layer the switching rate was doubled to 800Mb/s by paying special attention to the on-chip circuits and off-chip signal integrity effects. The NUMAlink 3 link bandwidth is 1.6GB/s/direction. The NUMAlink 3 physical layer latency including 24" of PCB and 3 meters of cable totals 28 ns. A particularly important improvement was the incorporation of on-chip termination resistors in place of the off-chip resistors on NUMAlink 2 which were a source of signal reflections that affected signal eye opening. The NUMAlink 3 routing layer was enhanced to support up to 2048 processor sockets including headless (i.e. processors removed) memory expansion modules, termed M-bricks. NUMAlink 3 supported a globally addressable memory space of up to 1TB within a single system image. Eight port NUMAlink 3 routers were designed to connect to the higher processor count systems with fewer router-to-router hops. The pin to pin NUMAlink 3 router latency was 26 ns.

NUMAlink 4 is SGI's newest generation interconnect, supporting Altix servers and supercomputers. It operates at the physical layer using an 800Mb/s simultaneous bidirectional scheme in which each wire pair simultaneously transports signals in both directions (40 data bits in each direction at 800Mb/s) as opposed to unidirectional signaling utilized in NUMAlink 2 and NUMAlink 3. This technique effectively doubles the interconnect bandwidth without suffering the higher losses of 1600Mb/s signaling. The NUMAlink 4 links can also operate in NUMAlink 3 mode (20 data bits, 800Mb/s uni-directional) thus enabling backward compatibility with NUMAlink 3, for product upgrades to Altix® systems. The NUMAlink 4 bandwidth is 3.2GB/s/direction. The NUMAlink 4 physical layer latency including 24" of PCB and 3 meters of cable totaled 28 ns. NUMAlink 4 utilizes off-chip precision reference resistors and on-chip FET/resistor circuits to achieve improved termination independent of silicon process, voltage and temperature variations. NUMAlink 4 also utilizes on-chip active equalization circuits to compensate for the interconnect losses and thus allows for increased interconnect reach up to 30" of PCB's (Nelco 4000 class) and up to 7 meters of cabled interconnect. The NUMAlink 4 routing layer was enhanced to support scalability in processor sockets (up to 8K sockets), memory (up to 10's of terabytes), IO (up to 8096 IO nodes), scalable graphics (up to 8096 graphics ports) and reconfigurable computing with FPGAs (up to 8096 NUMA/

Table 1 Summary of SGI NUMAlink interconnect generations

Generation	Signaling Rate	Link Bandwidth per Direction	Physical Layer Latency (inc. 3m cable, 24" pcb)	Products/Year of Introduction
NUMAlink 2	400Mb/s	800MB/s	55 ns	Origin 200 and Origin 2000/1997
NUMAlink 3	800Mb/s	1.6GB/s	28 ns	Origin 300 and Origin 3000/2000
NUMAlink 4	1600Mb/s (simul. Bidir)	3.2GB/s	28 ns	Altix/2004

FPGA nodes). An 8 port count router was designed to interconnect NUMAlink 4 systems with low pin-to-pin latency of just 26 ns.

NUMAlink 5 is currently under development for use in future SGI systems. The NUMAlink 5 physical layer will feature significantly higher unidirectional signaling for increased data bandwidth and even further-reduced latency. The routing layer will feature improvements that further expand multi-paradigm compute system scalability. Multi-paradigm computing is SGI's future vision for attaching a variety of computing elements, such as FPGAs, graphics units, vector processing units, processor-in-memory, etc. directly into the NUMAlink memory fabric.

3 Functional Distinctions Between NUMAlink and Competing Interconnects

A unique feature of NUMAlink is that it is a shared memory, globally addressable system interconnect. All physically distributed system memory is mapped into one global address space. NUMAlink based Altix systems offer memory scalability in the 10's of terabytes and SGI has installed Altix systems with up to 13 terabytes of globally addressable system memory. SGI's next generation NUMAlink 5 based systems will offer petabyte levels of globally addressable system memory. This memory may physically reside in compute, IO, graphics or RASC nodes. No other interconnect offers such globally addressable memory scalability.

An important distinction between the NUMAlink network and competing system interconnects is that NUMAlink is connected into the memory infrastructure of the system, versus being indirectly connected through an IO subsystem chip. As a result, competing system interconnects such as InfiniBand, Myrinet and Quadrics use memory addressing schemes implemented in the IO subsystem, whereas the NUMAlink network allows global memory addressability through processor-issued loads and stores to global shared memory addresses.

There are other unique benefits from the connection of the NUMAlink network into the memory infrastructure. The direct memory interface reduces the one way request-to-memory access time by the transit time through the IO subsystem on each end of the link and furthermore the bandwidth is not affected by limitations by the IO sub system. This is a considerable advantage for SGI's network architecture. InfiniBand, Myrinet and Quadrics interconnect systems typically connect into PCI slots in IO subsystems. The PCI interface adds latency and also limits link bandwidth to that of PCI-X bandwidth of approximately 1GB/s (only one direction at a time). In contrast NUMAlink is incorporated as multiple ports off a SGI HUB chip so there is no requirement for transit through a HUB to PC bridge.

NUMAlink routing layers have been designed and optimized for relatively small message sizes such as cache line transfers. The messages are compact with little overhead, as are the individual micropackets they are made from. For example, each NUMAlink 4 micropacket contains 16 bytes of data and 4 bytes of link level sideband. For a typical NUMAlink 4 cache line transfer, nine micropackets are required of which just one is message overhead in the form of a header that specifies such things as source and destination address. Many competing interconnect architectures offer wide ranging routing layer feature sets that conflict with the goal of efficient routing of cache line transfers. Architectures such as those targeting Data Center or Wide Area Network applications may have much larger headers creating excessive overhead for small message sizes.

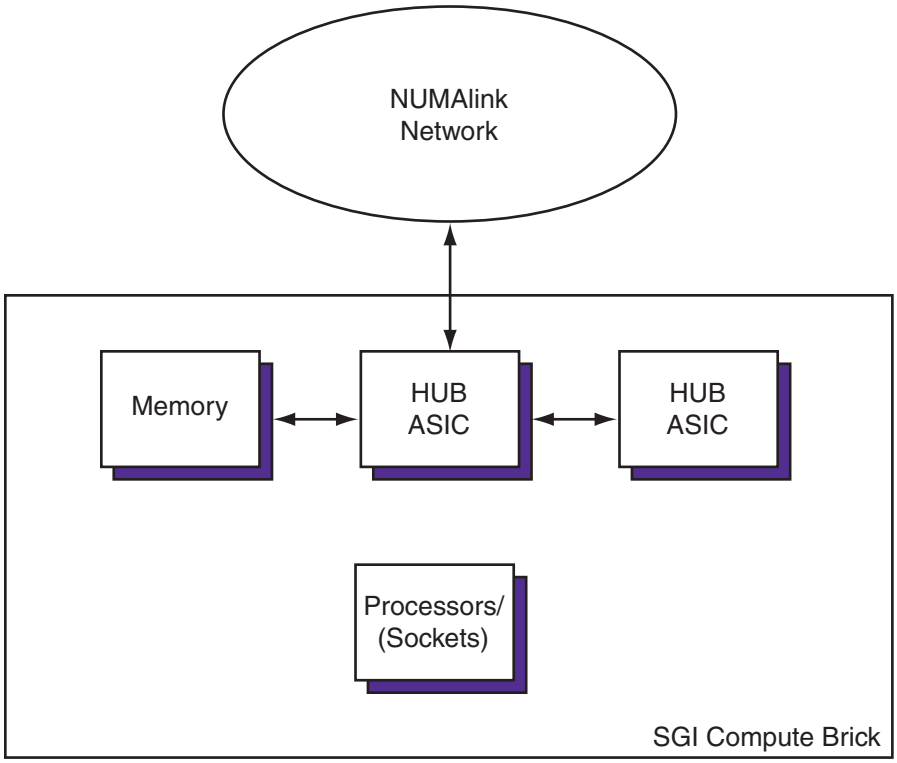

Figure 1 NUMAlink connection into SGI Compute brick

The NUMAlink physical and link layers are optimized for high bandwidth, low latency reliable communication. Data flow at the physical layer is optimized to facilitate transfers quickly after they become available. Clock forwarding schemes offer significant bandwidth-efficient and latency-saving improvements over clock encoded data schemes used in competing interconnects. Advanced equalization techniques are used to support long interconnects as well as to support longer links common to high density PCB designs. Single clock cycle link layer logic ensures very low latency, reliable, data transfer.

The unique advantages of the NUMAlink network enable extremely scalable algorithms and applications to be developed on NUMAlink based computer systems. The ultra low latency of the NUMAlink network permits SGI's Message Passing Interface (MPI) implementation to deliver MPI-1 send-and-receive as well as MPI-2 features like direct put and get communication with latency well below that of other interconnects. SGI's lightweight SHMEM™ message passing library specification is implemented on a number of competing systems and interconnects, but it performs substantially better on systems with memory-connected networks because the library overhead for data transfer requests is extremely low.

The direct memory load-and-store approach to communication yields a number of other benefits. First, the NUMAlink network permits operating system host size scaling to sizes well beyond others in the industry. This in itself promotes code scaling by allowing OpenMP™ and POSIX threads parallel applications to run on more CPUs. Additionally, the remote memory load-and-store capability allows users to code efficient memory sharing approaches using SGI's Global Pointers capability or the global memory allocation function suite provided in the SHMEM and MPI libraries. By use of these constructs, the program may employ generic, flexible, compiler-generated code to reference arbitrary data structures on remote memory without the need for the user to translate the remote references into less intuitive message passing subroutine calls.

4 Performance Results, User Benefits

4.1 Superior Performance Across a Variety of Common Benchmarks and Applications

Beyond its support for breakthrough big-memory applications, a variety of competitive benchmark data demonstrates superior performance of the NUMAlink interconnect on both technical metrics and user applications.

From the perspective of system bandwidth and latency, a recent paper commissioned by Hewlett Packard gives an assessment of commonly available interconnects in the mid-2004 timeframe (Cambridge Consulting, 2004). NUMAlink 3 technology is the clear acknowledged leader in this paper. A compilation of the most recently developed interconnect technologies below in Table 2 demonstrates the continued dominance of the NUMAlink 4 interconnect as well.

In the case of highly parallel system benchmarks, such as Linpack used in the Top 500 ranking, NUMAlink systems exhibit the highest Linpack efficiency compared to other microprocessor based systems and clusters (table 3).

Beyond the Linpack benchmark, another recent study comparing SGI's Altix to Cray® vector and IBM® POWER4™ architectures again showed the value of the NUMAlink interconnect. Presented at Supercomputing 2004, the paper showed very good sustained performance in the Altix compared to Cray and IBM systems costing significantly more. (Oliker et al., 2004).

The performance impact of the NUMAlink design with direct access to very large amounts of global shared memory has been demonstrated recently in a variety of applications. For example, Landmark Graphics employed the SGI shared memory capability to interact with 400GB of seismic data in real time—more than 4 times the previous record (Landmark Graphics, 2004). In the area of CAE, ANSYS recently reported solving a 111 Million Degree of Freedom structural analysis problem for the first time, with time-to-solution of only a few hours (ANSYS Corp., 2004). Finally, SGI has set a world record in database performance (TPC-H benchmark) with an 8 processor Altix accessing 64GB of memory. (See *www.tpc.org*). An SGI database partner, Xcelerix, is taking advantage of this

Table 2 Latency and bandwidth performance of common intereconnect technologies

Technology	Vendor	MPI Latency, µsec, short msg	Bandwidth per Link (Unidirectional, MB/s)
NUMAlink 4 (Altix)	SGI	1	3200
RapidArray (XD1)	Cray	1.8	2000 (1)
QsNet II	Quadrics	2	900 (2)
Infiniband	Voltaire	3.5	830 (3)
High Performance Switch	IBM	5	1000 (4)
Myrinet XP2	Myricom	5.7	495 (5)
SP Switch 2	IBM	18	500 (6)
Ethernet	Various	30	100

1. *http://www.cray.com/products/xd1/index.html#RapidArrayInterconnect*
2. *http://doc.quadrics.com/Quadrics/QuadricsHome.nsf/DisplayPages/81DD13F71CFD762580256EAD0010AA75/$File/Performance.pdf*
3. *http://nowlab.cis.ohio-state.edu/projects/mpi-iba/*
4. *http://publib-b.boulder.ibm.com/Redbooks.nsf/f338d71ccde39f08852568dd006f956d/55258945787efc2e85256db00051980a?OpenDocument*
5. *http://www.myricom.com/myrinet/performance/*
6. *http://www-1.ibm.com/servers/eserver/pseries/hardware/whitepapers/sp_switch_perf.pdf*

Table 3 Comparison of Linpack system efficiencies in the November 2004 Top 500 list

System/Interconnect	Ave. Linpack Efficiency for 256P System, Percent*	Sample size, number of systems on list*
SGI Altix and NUMAlink	84	14
HP Superdome	79	18
Various/Quadrics	75	4
Various/Infiniband	75	3 (one system at 288P)
Various/Myrinet	63	19
Various/Gigabit Ethernet	59	14

*Linpack Rmax/Rpeak for 256P systems listed on November 2004 Top 500 list (see *www.top500.org*).

capability, achieving 2-3 orders of magnitude speedup with in-memory queries compared to disc-based databases. This has been shown for up to 120GB databases containing up to 500 Million records (Xcelerix Corp. 2004).

4.2 Ease of Development, Administration, and Use

The benefits of these systems go beyond application performance. The large system sizes enabled by NUMAlink technology create a development and administration environment that is unmatched in ease-of-use. Scaling of applications across tens or even hundreds of processors can be accomplished with up to 512 processors under the control of one operating system image. SGI systems with NUMAlink interconnect easily handle applications with any type of shared or distributed memory programming paradigm and even hybrid schemes. Larger systems are particularly well-suited for situations where home-grown applications are actively developed, scaled and optimized whereas mid-range systems are ideal for any sort of smaller-scale workload and efficiently run hundreds of commercially developed codes. Overall, the superior performance, flexibility and ease-of-use makes SGI Altix with NUMAlink technology an ideal platform for nearly any high performance computing deployment.

5 Summary

SGI NUMAlink technology has evolved over several generations as a reliable, high bandwidth, low latency interconnect for SGI's ccNUMA compute servers.

NUMAlink technology enables cache-coherent globally addressable memory. This type of direct memory access greatly reduces latency and requires less data transfer while preserving system bandwidth. The ultra low latency, high bandwidth, and cache-coherent addressability provide system software with mechanisms that can be used to craft highly scalable applications. Host size is increased, MPI communication is optimized, and alternative programming styles like SHMEM programming model and direct use of global shared memory are available.

The NUMAlink routing layers have been designed and optimized for efficient (low overhead) routing of small messages such as cache lines. The routing features are necessarily limited in order to achieve the highest performing, tightly coupled systems. SGI systems utilize a network off HUB chip architecture in place of a network off IO sub system to minimize communication latency and avoid bandwidth limitations brought on by the IO sub system. The physical layer features advanced circuit designs to enable high bandwidth, low latency transfers and to enable connections found in high density NUMAlink board designs.

SGI is currently working on the development of the next generation NUMAlink, NUMAlink 5, to be used in the system architecture. The next generation product will build upon SGI's system architecture to deploy multi-paradigm computing, using NUMAlink 5 technology as the foundation for scaling system configurations.

The results from the superior design of NUMAlink systems are apparent in both standard system metrics, system benchmarks and real-world applications. Not only does the NUMAlink technology deliver industry-leading sustained performance, but it enables breakthrough problem-solving and discovery due to its ability to support very large globally addressable memory. This same large system size also results in an efficient, productive development and administration platform. With good performance across both shared and distributed memory programming models, SGI Altix with NUMAlink interconnect constitutes an extremely versatile and elegant platform for any high performance computing deployment.

6.0 References

ANSYS Corp., 2004. “ANSYS Breaks Engineering Simulation Solution Barrier” Press release dated May 25, 2004. Available at *http://www.corporate-ir.net/ireye/ir_site.zhtml?ticker=ANSS&script=410&layout=6&item_id=575478*.

Cambridge Consulting, 2004; “The Optimal Interconnect for High performance Clustered Environments.” Published May 30, 2004 by Cambridge Consulting.

Landmark Graphics, 2004. “Silicon Graphics and Landmark Graphics Announce Breakthrough in Search for New Oil Deposits.” Press release dated October 11, 2004. Available at *http://www.lgc.com/news/pressreleases/20041011-silicon+graphics+and+landmark+graphics+announce+breakthrough+in+search+for+new+oil+deposits.htm*.

Oliker, et al., 2004; “Scientific Computations on Modern Parallel Vector Systems” Oliker, L., A. Canning, J Carter, J. Shalf, S. Ethier. Supercomputing 2004 Conference, November 2004) Available at *www.sc-conference.org/sc2004/schedule/pdfs/pap247.pdf*.

SGI, 2004. “SGI Seizes Lead Over HP in Database Performance for Mid-Sized Servers”. Press Release dated October 15, 2004. Available at: *http://www.sgi.com/company_info/newsroom/press_releases/2004/october/tpc_benchmark.html*.

Xcelerix Corp., 2004. “Xcelerix and SGI Announce Partnership” Press release dated October 26, 2004. Available at *http://xcelerix.com/index.tpl?action=news&newsid=5*.

SGI 2004 "Message Passing Toolkit (MPT) User's Guide". Available at *http://docs.sgi.com*.

Silicon Graphics, Inc.
1500 Crittenden Lane
Mountain View, CA 94043
(650) 960-1980
www.sgi.com

North America +1 800.800.7441
Latin America +55 11.5509.1455
Europe +44 118.912.7500
Japan +81 3.5488.1811
Asia Pacific +1 650.933.3000

3771 [03.2005] J14820