SiliconGraphics
Computer Systems

Origin™ ccNUMA Servers

True Scalability with a Difference

White Paper

Origin™ ccNUMA Servers

True Scalability with a Difference

As reinforced by its increasingly visible implementation industry-wide, cache-coherent nonuniform memory access (ccNUMA) architecture is overtaking symmetric multiprocessing (SMP) as the preferred architecture for high-performance multiprocessing servers. Silicon Graphics was an enthusiastic and early adopter of the ccNUMA model and holds the current leadership position with the most robust and successful implementation of the architecture.

With several years of practical experience over late-coming vendors, over 30,000 installed systems, and documented customer improvement encompassing the areas of performance, scalability, flexibility and bandwidth utilization, Silicon Graphics proudly delivers award-winning Origin server products today and is poised to lead the industry with its next-generation products tomorrow.

Why ccNUMA over SMP?

Most SMP systems do not scale well beyond four or eight processors because the memory bandwidth does not scale when adding additional CPUs. After a certain point, adding processors to an SMP system creates a memory bottleneck. ccNUMA systems have the ability to vary memory access workloads to meet increasing throughput demands from additional processors; therefore, memory bandwidth scales upward as CPUs are added.

Like SMP systems such as the earlier Silicon Graphics® POWER CHALLENGE™ server series, which has a large cache-coherent memory, ccNUMA systems rely on shared memory between the processing modules. The difference is that in SMP systems, memory accesses have equal latencies. By definition, ccNUMA architectures have memory access times that vary. Silicon Graphics ccNUMA implementation was designed to make the best use of local memory access, and most memory accesses occur between nearby processing modules; high-latency memory accesses between far-apart modules occur less frequently.

In October 1996, Silicon Graphics introduced the Origin line of servers based on the Silicon Graphics implementation of ccNUMA. The basic processing unit of an Origin server is a central node board that provides all the interconnections the system needs to scale upward by adding node boards. Each node board contains two 64-bit MIPS® RISC microprocessors, each with 4MB of cache, up to 4GB of error-correcting memory, the corresponding directory memory, a direct connection to the I/O system, and an ASIC called the Hub.

The Silicon Graphics ccNUMA architecture was designed to provide a distributed memory system that supports very low memory latency without creating memory bottlenecks. Therefore, a primary design consideration was to keep as many memory calls as possible local to the Origin node board. The Hub ASIC provides each processor and I/O device with transparent access to all distributed memory.

A significant difference between ccNUMA and SMP architecture is that ccNUMA lacks a centralized system bus. Instead of a system bus, each node board in a Silicon Graphics Origin server uses the Hub ASIC to connect local memory with the two processors on the node. The Hub Crossbar, with a latency of less than 300 nanoseconds, provides the processors with access to local memory.

Differentiating Features of the Silicon Graphics ccNUMA Implementation

The reduced latency of the Silicon Graphics ccNUMA architecture provides the extra performance that can help customers tackle a wide variety of large and growing business problems. The modular, low-cost implementation of ccNUMA in the Origin2000™ series of servers has more flexibility, higher scalability, better bandwidth utilization, and greater availability than SMP systems and other ccNUMA servers available today.

Flexibility

Each node of the Origin2000 server is a two-processor CPU board. A module consists of up to four nodes. Customers can deploy and redeploy these modules as computing requirements change and increase. Adding to the number of processors working on a business problem is as simple as adding node boards and rebooting, whereas exhaustive integration is necessary to reconfigure both hardware and software in a clustered environment. As application requirements change, Origin2000 servers can also be split apart to function as smaller systems for geographically dispersed distributed applications.

The base configuration of an Origin2000 server already has the infrastructure it needs to support additional processors. A deskside system is configured with two to 16 processors, and an Origin rack can be scaled from two to 128 or more processors as a single system image. If a customer needs to upgrade a system repeatedly to meet growing performance requirements, upgrading an Origin2000 server from, for example, eight to 32 CPUs and then 32 to 128 CPUs is far less disruptive than having two "forklift upgrades" to new server architectures and different operating system variants. Even if a customer quadruples the number of processors, the Origin2000 server continues to use the same system memory, disk arrays, firmware, and operating system.

Higher Scalability

Many business problems get larger and more complex as time passes. For example, more data warehouse data accumulates each month, and more subscribers compete for an Internet service provider's bandwidth resources as the ISP's subscription services become more popular. The scalability of Origin2000's implementation of ccNUMA helps customers respond to business problems that require more computing resources as they get larger. Clustered SMP servers require additional infrastructure investments each time an organization purchases another SMP server, but not so with Origin servers. Customers can increase the performance of an Origin2000 server by purchasing additional modules and connecting them to the existing system without having to upgrade the software or networking.

There is also scalability within and between Origin2000 modules. The Origin2000 server uses the high-speed CrayLink® interconnect to move data between local and remote memory that is physically distributed on different nodes. The CrayLink interconnect is a patented fabric of cascading crossbar switches routing data to resources on an as-needed basis, which improves efficiency in larger distributed applications.

Better Bandwidth Utilization

Adding an Origin2000 processor module to an existing server adds bandwidth and I/O to the overall system and provides both I/O and performance gains no matter how large the system gets. SMP architectures do not continue to scale in this way. They encounter bottlenecks as they grow beyond a certain size because interprocess communications and contention for resources between processor modules begin to take more and more of the system's time.

Clustering SMP systems does not increase the efficiency of their bandwidth utilization because memory access across clusters also slows these systems. Therefore, adding resources to an already large clustered SMP system does not increase the performance proportional to the size of the new addition, because the processors in the cluster are all competing for the same resources and I/O through a high-speed bus. As more processors in the same SMP cluster share this bus, the bandwidth per processor goes down. Although ccNUMA systems can take advantage of performance gains from parallelized versions of SMP software (Oracle® Parallel Server, for example), it is not necessary to cluster these systems to attain the full performance advantages of the additional processors.

The Silicon Graphics "busless" switch-based ccNUMA architecture is designed to keep as many memory calls as possible local to the node. In the Silicon Graphics ccNUMA architecture each two-processor node uses the Hub Crossbar switch, which connects to another crossbar called the router. The nodes connect to each other via these routers, creating an interconnect switching system with multiple paths. Because of the global addressability of all memory and all I/O, the overall transfer of data in and out of the machine occurs much more

efficiently than in bus-based architectures. In instances when memory access is not local to the node, the IRIX® 6.5 operating system has the intelligence to invoke a command that migrates data in the system to a location that improves performance.

Greater Availability

Customers want assurance that business-critical systems will keep running even if individual parts fail or single processes die. Many require that their applications remain up and running even during routine system maintenance. The interconnect fabric within Origin2000 isolates each module so that a failure in one module does not affect other modules. In a multiple-partition system, the isolation of modules also makes it possible to power down one module for parts swaps and system maintenance without affecting the other module.

Origin2000 has safeguards against environmental causes of failure that may affect only one module. Each module has its own power source and is designed to function independently in the event of localized failures. The Silicon Graphics multimodule system controller (MMSC) provides individual control of the power and cooling systems for each module.

The ease with which an Origin server can be partitioned adds another dimension of reliability and fault tolerance. In partitioned Origin2000 servers, failures, whether in software, hardware, or power supply, affect operations only in the one partition rather than in the entire system. Applications running in a partitioned Origin server do not share memory latency or bandwidth with other applications on the server and remain unaffected by any problem that may plague an application in another partitioned area of the server.

ccNUMA as Architecture of Choice

ccNUMA is becoming the architecture of choice for multiprocessing systems because it has better price/performance, demonstrates better scalability in performance, and runs applications with less administrative overhead than comparably priced clustered SMP servers. ccNUMA systems can run, without modification, the shared memory processing codes that were originally created for symmetric multiprocessing systems. In fact, these shared memory-processing codes may perform better on ccNUMA systems.

Software costs are also lower for ccNUMA than for clustered SMP systems. Clusters require specialized operating systems, applications, and filesystems. Consequently, fewer applications have been rewritten for clustering SMP systems, and the license costs for them reflect the additional development efforts for specialized versions. Although it is not necessary to cluster SMP systems to attain the highest performance, ccNUMA systems do run well in clustered environments. If your company has already made a substantial investment in clustered software for SMP systems, ccNUMA systems can run these applications using the same cluster codes without modification.

ccNUMA is not just for high-end systems. Although it's true that ccNUMA architecture can scale to hundreds of processors, high-end scalability was not the only design goal for the Silicon Graphics ccNUMA implementation. The system architects who designed the Silicon Graphics ccNUMA architecture made special provisions for small systems as well by achieving a granularity that allows small configurations to be priced competitively with clusters of small SMP systems. Therefore, a ccNUMA server with a low CPU count can have a low initial price because it is not burdened with the additional infrastructure that SMP servers require in order to scale. The ccNUMA system design adds infrastructure as the system scales, allowing for a lower entry price and a "pay as you grow" investment model. Because of the extra infrastructure that SMP servers require, most lines of SMP servers have significantly higher initial prices, thereby making the price for a small configuration of a highly scalable cluster less competitive.

ccNUMA: Solving the World's Toughest Problems

Two years after introduction, Silicon Graphics Origin servers are widely deployed at companies where they are used to tackle tough problems that require more bandwidth, greater scalability, and more flexible distribution of resources than other commercially available servers offer. Some industries and innovative customer solutions are detailed below.

Helping ISPs Serve More Subscribers

Many Internet service providers, including some of the world's largest Web-hosting providers, use Origin servers. Hiway Technologies, a rapidly growing hosting service, uses entry-level Origin200® servers for the heart of its 100,000-site network. Origin200 servers are more scalable than comparably priced servers that rely on clustering to achieve scalability.

The superior scalability of the Silicon Graphics ccNUMA architecture is helpful for businesses such as ISPs, which must respond quickly to spikes in demand. ISPs often need to scale up their bandwidth as soon as more users subscribe to their service. One way that ISPs respond to this problem is to add mirror machines (other Web servers with similar content) and then use a proxy server to do load balancing between the machines. Thus, as demand increases, there may be multiple servers sharing the same data and running the same URL.

Mirror machines provide extra capacity at a high price. Adding a server means adding an operating system and a proxy server. Silicon Graphics Origin servers provide a better way to scale up quickly. Adding a node to an Origin server instead of adding another server is a more cost-effective way to add capacity. The extra capacity shares the same data store and does not require additional software. When an ISP has added so many nodes that it becomes necessary to add a proxy server, Origin200 can be partitioned to function as the proxy in addition to being a Web server.

Analyzing Complex Data More Efficiently

Automotive and aerospace manufacturers use Origin2000 servers to solve complex problems such as determining how much elasticity is needed when an airplane skin is stretched over the fuselage. This is a complex modeling problem that requires immense processing for numerical analysis of very large amounts of constantly changing data. For problems such as this, engineers simulate and model the fuselage and the outer surface of an airplane and then run simulations of different problems against these models. Many time-to-solution analysis codes use very large data sets, and dividing a large problem into components to be solved by a cluster is not an easy task. Such large problems are better served by a single system image system having access to many processors and a single global memory.

If this simulation problem is split between more than one computer in a cluster, the clustered computers can reduce the total analysis time required to solve the problem. But as more CPUs are added to a cluster, the communication load between the CPUs starts to increase. At some point adding another CPU will not reduce the time significantly. In fact, copying and moving the model data created during analysis and simulation may create a bottleneck.

Silicon Graphics Origin2000 can tackle a large analysis problem in a way that does not create a data transfer bottleneck. Manufacturers who use Origin2000 servers for compute-intensive simulation problems can simply add new nodes and parcel the analysis to more nodes. The data stays in the same place and remains accessible to all the nodes, so there is no longer a bottleneck.

Rendering for Realism and Special Effects

Just as product manufacturers utilize rendering to solve the complex mathematical analysis necessary to visualization, creators of animation for the movie industry need fast rendering to provide realistic images for feature-length films. DreamWorks/Pacific Data Images (PDI) used Origin200 servers to provide the compute

power necessary to render the entire feature-length movie *Antz*. Special challenges included the visually complex ant colony and large crowd scenes with thousands of individually realized characters.

DreamWorks/PDI used more than 270 dual-CPU Origin200 servers to manage and manipulate the massive amounts of image data. Two large Origin2000 servers were used to serve all of the render data to the animators' desktop machines. The production required over 3.2TB of storage. "The groundbreaking animation seen in *Antz* required unprecedented levels of computing power," said Carl Rosendahl, president of PDI and executive producer of *Antz*. Low-cost Origin200 servers working together can achieve more linear scalability than comparably priced SMP servers working in a clustered environment and without the administrative overhead of clustering software.

High-Bandwidth Processing for Genetic Sequencing

At Monsanto Corporation in Saint Louis, Missouri, researchers in the genomics group use Origin2000 servers to process complex genetic sequences. By studying the genetic content of various strains of commonly cultivated plants, including corn, soybeans, cotton, and potatoes, researchers can better understand the genetic structure of plants and cultivate strains that are more nutritious, pest-resistant, drought-tolerant, and easier to cultivate than existing varieties.

To process gene sequences, an application requires high bandwidth, fast cache, and large data stores. "We selected Silicon Graphics servers as the best all-around solution for our bioinformatics applications. On the software side, we needed the best visualization tools. For our high-performance computational platform, we needed servers with very fast cache to move the genetic data to multiple processors as quickly as possible and very high bandwidth to process this vast amount of sequencing data without creating bottlenecks. And with our large and growing data set, we also needed a scalable solution," said Harry Harlow, director of bioinformatics at Monsanto.

Applying New Data for Accurate Decision Support

Origin2000 servers help QVC, a cable shopping network, to look at purchasing trends and to determine the best ways to market services to the vendors whose products they advertise. QVC processes an immense number of transactions 24 hours a day. Every day, data captured from these transactions feed decision support applications that can help QVC detect purchasing trends and provide better service to its 5 million customers.

There is a nonstop flow of data into QVC's decision support system. Therefore, the server used in this application needs to perform two separate tasks. It must process the incoming data, and it must also analyze the data. The Origin2000 ccNUMA architecture is ideal for complex, nonstop strategic business analysis because the architecture is flexible enough to support data entry and data analysis within the same partitioned server. Instead of having one system for data entry and another to support the analysis process, the same Origin2000 server supports both tasks. In addition, the Origin2000 server's superior data handling capabilities allow for very fast loading and data sorting. Its large distributed shared memory system design is also a benefit to QVC because it can support the retailer's large and rapidly growing store of data.

Data Access for Future Telecom

Origin servers excel at handling big problems that continually grow in complexity. In the future, telecommunications companies could deploy Origin servers to meet the demand for portable phone numbers that follow people wherever they go. The challenge lies in allowing customers to change local phone companies yet retain their assigned phone number.

To do so, massive distributed computers, handling thousands of calls instantaneously, must look up data concerning a customer's phone number and personal information, poll local companies to determine which one controls the number, and provide routing and switching instructions so that the call will go to the right destination.

Fast call identification and access are not the only data-intensive problems that phone companies must solve. Additionally, outgoing calls need to provide return routing and caller ID information to other switching systems. Add 911 emergency information and other caller ID info that is sent between phone companies, and the vast amount of data transfer alone could bring down servers that cannot scale upward.

The ability of Origin servers to scale both as a single machine and as a cluster make them well-suited to the telecommunications industry's efforts to provide local number portability. Working together in clusters, multiple Origin servers could be used to look up caller information and then transfer calls through the switching systems. As the workload shifts at various times of day, individual Origin servers can be repartitioned on the fly to bring optimal processing power to the application areas where they are most needed.

World's Most Powerful Computer as Nuclear Safeguard

Los Alamos National Laboratory (LANL) in Los Alamos, New Mexico, has installed the world's most powerful computer system. This computer, code-named Blue Mountain, also comprises the largest single installation of Origin2000 servers in the world. Blue Mountain has 48 128-processor single-system image (SSI) Origin2000 servers for a total of 6144 processors. The final configuration also includes 1.5TB of memory and 76TB of Fibre Channel disk.

Together with Lawrence Livermore and Sandia National Labs, LANL is a participant in the Accelerated Systems Computing Initiative (ASCI), a government program utilizing high-performance simulations to replace live nuclear testing in an effort to safeguard our defense weapons stockpile. Blue Mountain runs the Stockpile Stewardship application to formulate the simulations that replace nuclear testing. In addition to maintaining the United States' nuclear stockpile, Blue Mountain will also be deployed on other difficult problems, including modeling the global climate, wildfires, and traffic and simulating the worldwide spread and control of influenza and other health risks. All of these applications involve tough problems whose solutions will make the earth a safer and more habitable place.

The world's largest and most powerful computer, Blue Mountain, uses the Silicon Graphics implementation of ccNUMA for some of the same reasons that businesses are using Origin2000 servers to solve their toughest business problems. Origin2000 servers deliver more CPU power in less space than a cluster of symmetric multiprocessing computers. Composed of Origin2000 SSI modules, Blue Mountain can take advantage of shared storage and memory, enhanced scalability in all dimensions, and performance benefits at the highest levels of utilization. In addition, the ccNUMA architecture provides a powerful well-integrated system with much less system management overhead and latency than a very large SMP cluster would create.

Silicon Graphics and ccNUMA: Ahead of the Market

Whether announced publicly or hidden in a research lab, most major server vendors are now incorporating ccNUMA concepts in their designs for their next generation of multiprocessing servers. While the rest of the industry is migrating to ccNUMA-like architectures and away from multiprocessing architectures that are based on one centralized system bus, the Silicon Graphics busless ccNUMA design has more than two years' time-to-market advantage and has actual applications and solutions experience over ccNUMA designs that are still in development.

SiliconGraphics
Computer Systems

Corporate Office
2011 N. Shoreline Boulevard
Mountain View, CA 94043
(650) 960-1980
www.sgi.com

U.S. 1(800) 800-7441
Europe (44) 118-925.75.00
Asia Pacific (81) 3-54.88.18.11
Latin America 1(650) 933.46.37

Canada 1(905) 625-4747
Australia/New Zealand (61) 2.9879.95.00
SAARC/India (91) 11-621.13.55
Sub-Saharan Africa (27) 11.884.41.47

2189 (1\99)