Cahners
MICROPROCESSOR REPORT
www.MPRonline.com
THE INSIDER'S GUIDE TO MICROPROCESSOR HARDWARE

SGI Updates Systems, CPU Plans

Better Architecture, Faster Processors Make Computers More Super

By Peter N. Glaskowsky {8/7/00-01}

SGI has announced new supercomputer systems based on a new nonuniform memory access (NUMA) architecture, NUMA3, that supports more processors and reduces memory latency compared with previous SGI systems. SGI has implemented the NUMA3 architecture in the new NUMAflex modular packaging technology, which will allow SGI's customers to upgrade these new systems over several years. The NUMA3 architecture and NUMAflex technology will be used in two system families, the Origin 3000 server line and the Onyx 3000 visual supercomputer family.

Origin and Onyx systems are distinguished primarily by their respective combinations of processing, display, storage, and interconnection components. Many of these components will be used in common by both families. SGI will offer both MIPS and IA-64 processors for these new systems. MIPS systems will run SGI's IRIX operating system, while the yet unannounced IA-64 machines will run Linux.

Though SGI's overall business is struggling, it remains one of the world's leading suppliers of "big iron"—large multiprocessor systems with shared-memory architectures. The new systems will help SGI preserve its position in this market and should also allow SGI to participate more effectively in the low end of the supercomputer market, where less sophisticated architectures, such as clustering, are used.

NUMA3 Boosts Multiprocessor Performance

Multiprocessor systems that use shared memory can be designed so that all processors have equal access to every memory block—a characteristic known as uniform memory access—or they can be configured to put a portion of the shared memory close to each processor (NUMA). In NUMA systems, each processor can access all memory in the system, but remote memory blocks take longer to access.

Uniform memory access works well with up to about eight processors (see *MPR 8/23/99-06*, "Profusion Lowers Cost of Eight-Way Servers"), but beyond that point the delays caused by requiring all processors to arbitrate for access to memory outweigh the benefits of uniformity. For larger systems, the NUMA architecture is superior, because each processor's memory-access patterns have some characteristic locality of reference—each processor is likely to access some locations more often than others.

NUMA configurations, especially when combined with intelligent virtual-memory management schemes, can scale up to very large numbers of processors. The primary limit to scalability is the delay encountered by memory transactions that must be completed in remote memory blocks. These transactions must cross an interconnection network between memory arrays, and all such networks add some latency to the transaction; some also provide less bandwidth than the connection to local memory.

SGI's NUMA3 architecture puts a small number of processors (currently four MIPS CPUs) plus local memory into each node, along with two interconnection ports. An eight-way system includes two connected nodes, each of which is connected to one I/O subsystem. Larger shared-memory systems use routers with six or eight ports each to provide the necessary connections among processor nodes

and I/O devices. Nodes may also be connected by SGI's proprietary XIO+ interface to create clustered systems.

SGI's Origin 2000 family was designed to scale up to 128 processors, in which configuration the system imposed a worst-case round-trip delay of nearly 1.2µs. The NUMA3 hypercube interconnection scheme in the Origin 3000 cuts this latency figure to just 415ns. Using Express Links, a NUMA3 feature that adds extra interprocessor routing resources, a 1,024-processor system is possible with only 460ns of worst-case latency. Even larger configurations may be possible.

The reduced latency of the Origin 3000 design improves application performance, especially on tasks that produce large amounts of nonlocal memory traffic. SGI says the Origin 3000 is about 20% faster on SPECfp2000 than an Origin 2000 system with an otherwise equivalent configuration. SPECint2000 results are almost unaffected, since the SPECint benchmark fits in cache and local memory. Large databases, on the other hand, could see up to twice the effective performance on Origin 3000 systems, even before considering the benefits of more and faster CPUs.

NUMAflex Improves Configuration Options

Mechanical design is a nontrivial issue in multiprocessor systems, especially those with hundreds of CPUs and hundreds of interprocessor connections. SGI's answer to this problem is NUMAflex, a hierarchy of modules, or "bricks," that perform specific functions. These bricks are installed in standard 19-inch racks, according to the needs of SGI's customers.

A minimal four-processor system can be created from just three modules: a four-CPU processor node called a C-brick; an I-brick I/O module that includes six hot-plug PCI slots, two XIO+ ports, two Fibre Channel interfaces, and a CD-ROM or DVD-ROM drive, plus USB, IEEE-1394, Ethernet, and serial interfaces; and a power bay that can hold up to six N+1-redundant power supplies. All power distribution is handled using 48V DC with local point-of-use regulation and redundant cooling fans in each brick.

Router bricks (R-bricks) are available in two versions—one with six ports for systems with up to 32 processors and one with eight ports that can be used to create systems with up to 512 processors. Larger systems may be created by adding another layer of routing bricks to the system configuration. Figure 1 shows a 128-processor system connected using eight-port R-bricks. Each router port, like the two ports on each CPU module, has a peak bandwidth of 3.2GB/s. Round-trip latency through the router brick is just 45ns.

SGI offers two other I/O bricks, a P-brick with 12 hot-plug PCI slots on six buses plus two XIO+ ports, and an X-brick for further XIO+ expansion. A disk brick (D-brick) that holds Fibre Channel disk drives is also available.

SGI has not forgotten its graphics-hungry customers. An optional G-brick that implements SGI's Infinite Reality 3 or 4 graphics subsystems is available for Onyx 3000 systems. Multiple parallel graphics pipelines are supported in the G-brick, along with optional boards to support multiple streams of uncompressed high-definition digital video. Although even low-end PCs now offer high-performance 3D graphics for gaming and CAD, these systems can't begin to compare with high-end visual supercomputers for complex applications, such as military flight simulators that use multi-terabyte terrain databases. An Onyx 3000 with the Infinite Reality subsystem supports real-time rendering of these enormous datasets to multiple simultaneous high-resolution monitors or HDTV feeds; it will be some time before any PC can do the same.

The NUMAflex architecture will allow SGI to update individual elements of these systems over time. The company plans to introduce IA-64 processor modules when the processors themselves become available, and PCI-X and Infiniband bricks (see *MPR 9/13/99-msb*, "NGIO, Future I/O Merge") are also planned.

Resiliency is another benefit of the new architecture. Any failed brick can be isolated from the rest of the system, replaced, and put back into operation when the system is restarted. Some smaller elements, such as CPUs, PCI cards, and power supplies, can be replaced if they fail, without disturbing other components in the same brick.

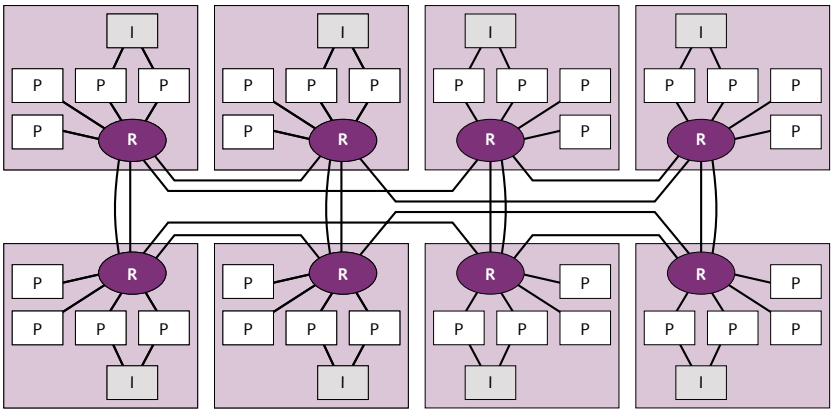

Figure 1. A 128-processor Origin 3000 system can be constructed with eight partitions, each with 16 CPUs on four P-bricks, one I-brick, and one router.

First Systems Rely on MIPS Processors

Although SGI will eventually offer Origin 3000 and Onyx 3000 systems with both MIPS and IA-64 processors, only the MIPS option is available now. Initial CPU options include the R12000, R14000, and R14000A. The next MIPS processor for these systems, the R16000, will have a core derived from the R14000 core, but the chip will feature an integrated L2 cache. The R18000 will use a new core with twice the number of execution units and twice the peak throughput, as measured in instructions per clock period.

SGI's roadmap also includes CPUs code-named N3 and N4. The N3 is expected to appear during the same timeframe as Intel's Madison IA-64 processor.

According to SGI's current plans, the new systems will support Itanium next year and McKinley in 2002, followed by Madison sometime later. SGI's plan to support only Linux on IA-64 is in keeping with its focus on scientific users, who have demonstrated a marked preference for the open-source Linux platform over Microsoft's proprietary Windows 2000.

SGI has announced several specific configurations (see box), but many customers are likely to create their own combinations of processors, I/O, disk, and graphics bricks to meet specific needs. SGI says it has already sold more than $100 million worth of the new systems, including more than 25 systems that each contain 128 processors or more. Customers include NASA's Ames Research Center, the U.S. Army Engineering Research Development Center, Sony Computer Entertainment, and Morgan Stanley Dean Witter.

Price & Availability

The entry-level Origin 3200 system with two processors and 512M of RAM is priced at $50,000. A 128-processor Origin 3800 costs about $3 million, depending on configuration. Systems with up to 128 processors are shipping now. SGI expects 512-CPU systems to be available by 1Q01. For more information, visit *www.sgi.com*.

SGI's biggest challenge in marketing these new machines will be finding enough users that need a high level of performance and can afford the large price tag that goes along with it. SGI's efforts to reduce system cost through modular design will make the new Origin and Onyx machines attractive to a wider audience.