Success Story

SGI and NASA: Extending the Boundaries of Supercomputing

The NASA Advanced Supercomputing (NAS) division, which is part of Ames Research Center, provides supercomputing capabilities for many of NASA's most important projects. From modeling the aerodynamics of potential space shuttle replacements to predicting Earth's climate far into the future, these critical projects encompass some of the most challenging computational problems ever undertaken.

The goal of NAS is to increase the use of supercomputers for important research by making them readily available, reliable, and easy to program and use, enabling scientists to focus on research, not computing. NAS is also bringing the benefits of supercomputing to areas of research where it has previously been underutilized.

According to Bill Feiereisen, NAS division chief, "NASA researchers depend on NAS to provide advanced supercomputing capabilities. Because reducing the time required to execute critical software is essential to continued progress, getting the most from each budget dollar spent on hardware and programming is critically important."

In recent years NAS has turned increasingly to SGI to provide production supercomputers to meet its computational needs. For NASA workloads, shared-memory, single system image (SSI) SGI® Origin® family systems have proved superior to large, clustered supercomputers, providing far greater performance on NASA projects with less programming effort and at significantly less total expense. SGI and NASA collaborated to extend the limits of SGI Origin family systems, culminating in the delivery of the first 1,024-processor SGI® Origin® 3800 system in July 2001.

"The 1,024-processor Origin 3800 system allows us to accomplish in hours jobs that literally took months or even years on previous systems. Because of the demonstrated linear scaling of NASA software on the SGI Origin family platform, we expect to achieve even greater things in the future. The computing capability offered by the Origin family platform will fundamentally change the way NASA carries out many of its most critical projects," said Jim Taft, co-director, Terascale applications group.

Replacing the "Gold Standard"

For years, NASA's gold standard for supercomputing was the Cray C-90, a vector supercomputer rated at 16 gigaflops. (A gigaflop is one billion floating-point operations per second.) The Cray C-90 demonstrated about

25% efficiency running NASA workloads, delivering about 4 GFLOPS when executing OVERFLOW, one of NASA's most important software programs. When it became clear that follow-ons to the C-90 were not available, NAS began exploring other supercomputing solutions.

Over a period of years, a variety of clustered supercomputing solutions were tried from a number of different vendors, but—for NASA workloads—none of these came close to the performance of a Cray C-90, let alone exceeded it. Even though these systems looked great on paper, with theoretical performances of hundreds of gigaflops, they achieved efficiencies of only a few percent on important NASA software and failed to compete with the C-90s already in place. Because of low efficiency, clustered systems large enough to achieve acceptable performance on NASA workloads would be prohibitively expensive if they could be built at all.

NAS purchased its first SGI Origin family system—a 128-processor SGI® Origin® 2000 server—in 1996 and was immediately impressed with the results. After just a few months spent porting code, the Origin 2000 system exceeded the performance of a Cray C-90 on OVERFLOW. More important, NASA software scaled linearly on the Origin 2000 architecture. Performance increased predictably as processors were added.

NASA immediately entered into a memorandum of understanding with SGI to collaborate on the development of larger systems. A 256-processor Origin 2000 server soon followed. That system was fully operational within a few weeks of installation and could run OVERFLOW at 20 GFLOPS—five times the speed of a C-90. This system was later followed by a 512-processor Origin 2000 system, and, most recently, the 1,024-processor SGI Origin 3800, which runs OVERFLOW 40 times faster than the C-90. Despite this impressive performance, the 1,024-processor Origin 3800 system cost NASA less than half as much as the C-90 did in the early 1990s.

SGI facilitated the purchase of the Origin family systems at NAS with special financing arrangements, enabling NAS to obtain equipment and begin important projects while spreading repayment across multiple funding appropriations.

SGI also established a special funding contract for NASA, simplifying the procurement process for all parties.

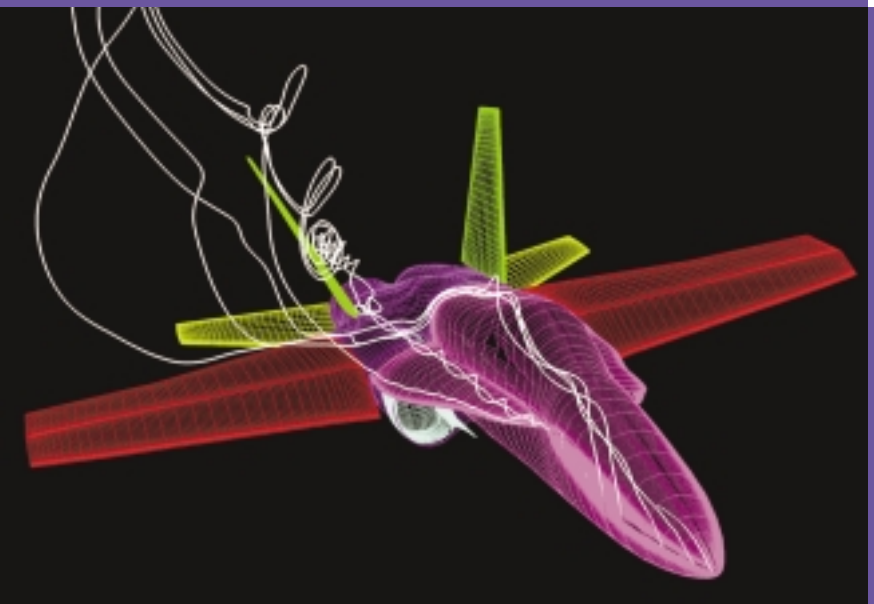
After a few months of operation, the Origin 3800 system is already achieving processing efficiencies approaching those seen with the C-90, proving it is possible to achieve high efficiencies on nonvector machines, something that has generally eluded application designers in the supercomputing world.

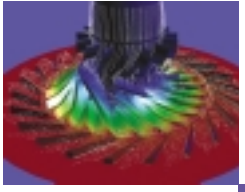
Building a Virtual Wind Tunnel

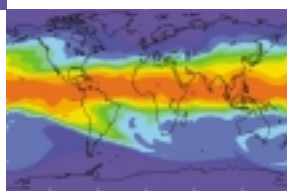
Historically, 20% of C-90 CPU cycles at NAS have been dedicated to running OVERFLOW and other important computational fluid dynamics (CFD) software. CFD is used to model the aerodynamics of advanced aircraft and space vehicles passing through Earth's atmosphere and is critical in the design and simulation of rocket motors.

The goal at NAS is to provide sufficient computing power to enable scientists and engineers to test their designs in a "virtual wind tunnel." Traditional wind tunnel testing is time consuming and expensive—requiring the construction or alteration of physical models for each test—and even the most powerful wind tunnel cannot simulate the conditions of the launch and reentry of space vehicles.

Until the arrival of the 1,024-processor SGI Origin 3800 system, the computing power needed to make the virtual wind tunnel a reality was unavailable. Complete modeling of one aircraft configuration during landing required up to a year on a C-90, imposing serious limits on the use of simulation. The Origin 3800 system can run the same configuration in a matter of hours, so, for the first time, scientists can routinely use simulation to validate their designs under







varying conditions. “The Origin architecture has created a revolution in computational fluid dynamics at NASA and will fundamentally change the way aircraft are designed in the future,” said Taft.

The ability to do advanced simulation has already proved its value in NASA’s mission to design a new reusable launch vehicle to replace the space shuttle. During simulation of the X-37 drone, designed to be dropped from the space shuttle to test reentry, a serious flaw was discovered that would have led to catastrophic failure. Millions of dollars and months of time were saved because of the advanced capabilities enabled by the Origin 3800 architecture.

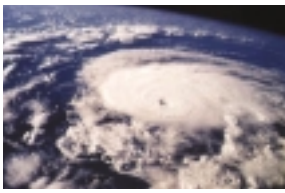
Predicting Climate Change

In the study of climate change, NASA is reaping similar performance benefits from the 1,024-processor SGI Origin 3800 server. After a three-week porting effort on the climate model of NASA’s next-generation climate modeling system, the model ran five times faster than on any other system ever tested. The Origin 3800 architecture again demonstrates superior scalability as more processors are used to execute the climate model. By comparison, clustered systems often reach a peak and then slow down as more processors are added.

With the success in enhancing the atmospheric model performance, NASA has moved on to ingest more satellite observations into the data assimilation system. The use of the 1,024-processor Origin 3800 system with state-of-the-art assimilation will help improve our understanding of the cause and consequence of the change in Earth’s climate system.

NUMAflex™: The Architectural Advantage

The 1,024-processor SGI Origin 3800 server is the largest system ever built with an SSI and shared-memory. All processors access all system memory directly. This is in sharp contrast to clustered solutions in which a separate instance of the operating system is needed for every few processors and each processor has direct access to only a subset of total memory.



The patented SGI® NUMAflex architecture makes it possible to scale the number of processors well beyond the level that has been possible in other shared-memory designs. Each processing node has up to four processors and a local pool of up to 8GB of memory. Instead of the traditional backplane design, NUMAflex uses crossbar switches and high-speed cabling, allowing each node direct access to the memory in other nodes with a relatively slight increase in latency versus accesses to local memory (hence the designation NUMA—nonuniform memory access).

NUMAflex uses standard, modular building blocks called bricks that allow systems to scale independently in different dimensions over time, providing unprecedented levels of flexibility, resiliency, and investment protection. Various types of bricks can be added as needed to tailor a system to the exact capabilities required by the application. As an added advantage, SGI® Origin® 3000 series systems provide an extremely small physical footprint relative to other comparable systems because of the efficient modularity of NUMAflex.

“The 1,024-processor Origin 3800 system allows us to accomplish in hours jobs that literally took months or even years on previous systems.”

—Jim Taft, Co-Director, Terascale Applications Group

MLP: A Simpler Programming Model

SGI supports a variety of programming interfaces for use on Origin family servers, allowing scientists and engineers to choose the best method for tackling a particular problem. SGI designs each programming interface to take maximum advantage of the unique features of the shared-memory Origin architecture, making SGI the leader in shared-memory programming.

NASA programs the 1,024-processor SGI Origin 3800 system using Multi-Level Parallelism (MLP), a programming method developed at NAS specifically for use with shared-memory NUMA architectures. MLP uses both coarse-grained parallelism in the form of UNIX® forked processes and fine-grained parallelism in the form of lightweight OpenMP™ threads. All communication between processes occurs via shared memory rather than message passing.

The message passing interface (MPI) is the standard programming model for clusters in which all communication occurs via messages passed between processors. (MPI is supported on SGI Origin family servers and has been implemented to take advantage of underlying shared memory to provide low latency and high bandwidth.) MPI is an appropriate programming model for various highly parallel problems, but for many of the problems tackled by NASA, MLP offers significant advantages over MPI.

For example, scientists have recently become interested in small protein molecules being studied for their pharmaceutical applications. Modeling the dynamics of these proteins usually involves fewer than 50,000 atoms. On a 1,000-processor machine this means that each processor is responsible for no more than 50 individual atoms.

The performance of MPI on clustered systems when simulating such a system quickly degrades because of the large number of messages that must be passed to calculate interactions between each atom and other atoms not resident on the local processor. (Each message has latency typically in excess of five microseconds.) Because of the submicrosecond latencies of shared-memory access on Origin family servers, the Origin 3800 platform with MLP runs small atom count problems faster than any other system on the planet.

For NASA, MLP has proved significantly easier to program than MPI, in which the programmer is typically responsible for analyzing and decomposing the problem into fine-grained parallel threads. With MLP, the compiler is capable of generating most of the parallel code with limited user directives. MLP also permits simple and highly efficient dynamic load balancing. Work can be shifted quickly and automatically between CPUs in response to load, providing better efficiency and therefore better performance.

NASA has also found that its legacy programs are much easier to port to MLP than to MPI. Relatively few code changes are necessary to port vector applications to MLP, and applications are far easier to debug than MPI programs. This has translated into a cost reduction of millions of dollars for NASA, adding to the price/performance benefit of the SGI Origin 3000 series architecture. For NASA, MLP and Origin family systems have proved a big win in application performance, programming simplicity, and overall expense.

A Future So Bright

Although NASA and SGI are excited by the new computing capabilities incorporated in the 1,024-processor SGI Origin 3800 system, they are far from ready to rest on their laurels. Plans are already in the works to scale the SGI NUMAflex architecture to even higher processor counts without sacrificing the benefits of SSI and shared memory.

Corporate Office
1600 Amphitheatre Pkwy.
Mountain View, CA 94043
(650) 960-1980
www.sgi.com

North America 1(800) 800-7441
Latin America (52) 5267-1387
Europe (44) 118.925.75.00
Japan (81) 3.5488.1811
Asia Pacific (65) 771.0290

3220 (04/02) JI3314