White Paper

SGI on the Grid

Unique Capabilities for Grid Computing

1.0 Introduction

1.1 What is Grid Computing?

Grid computing evolved to provide scientists and engineers quick and transparent access to critical computing resources and advanced instrumentation. A grid user can directly access computers, software, data, and other critical resources with little concern for the physical location of those resources. A grid enhances the ability to solve difficult problems, enables collaboration across organizations, and increases the utilization of expensive resources.

While distributed computation—using multiple computer systems to solve a single large problem—is an important aspect of grid computing, distributed access to unique capabilities is the driving force behind much of the ongoing grid development. Particularly in Europe, a great deal of effort has gone into creating grids for resource sharing. Connection to a grid increases system utilization, ensuring maximum return on investment for advanced computing resources.

It is important to recognize that "the grid" is really a model for organizing networked computing and not a single distinct entity. For example, the COSMOS Project at Cambridge University is studying the formation of galaxies after the Big Bang, using a dedicated cosmology grid. Cambridge hosts the largest cosmology supercomputer in the U.K. to tackle capability problems. Other universities have smaller servers or other system architectures connected to the grid to tackle capacity problems and other problems best suited to a particular architecture. Similar small grids are being organized all over the world.

While the current focus of the grid is scientific and government-funded use, the idea of grid computing is also starting to appeal to private industry. Many companies are exploring the deployment of intra-, extra-, and inter-grids to cost-effectively meet their computational needs. Intra-grids—sometimes referred to as enterprise grids—are designed to meet the computational needs within an organization. Extra-grids, or partner grids, tie in trusted partners and suppliers to enhance collaboration, while inter-grids might span several commercial organizations to facilitate broader sharing of resources. Since inter-grids may use the Internet as a transport medium, they are sometimes called Internet grids.

High-speed networking is a critical facilitator of grid computing in all these scenarios. The availability of high-speed backbones linking major research facilities—such as SuperJANET in the U.K., NSFnet in the U.S., and CANARIE in Canada—provide the bandwidth to make grid computing possible over wide areas. Standard LAN technology provides enough bandwidth for more localized grids or for within grid-attached computing centers.

The problems tackled by grid computing can be broken into two classes:

- Capacity computing, in which a large number of small jobs are run simultaneously. Each job requires a small number of processors. Peak loads may require use of a large number of individual systems. Large problems that are easily partitioned into many smaller jobs also fall into this category. The self-contained nature of capacity jobs typically requires computing ability but places fewer demands on system I/O resources and internal system bandwidth.
- Capability computing, in which a huge number of processors, large shared memory, and multiple I/O channels are required to tackle the most difficult computing problems. Capability problems often stretch the limits of the processor interconnect as much as they do the compute ability.

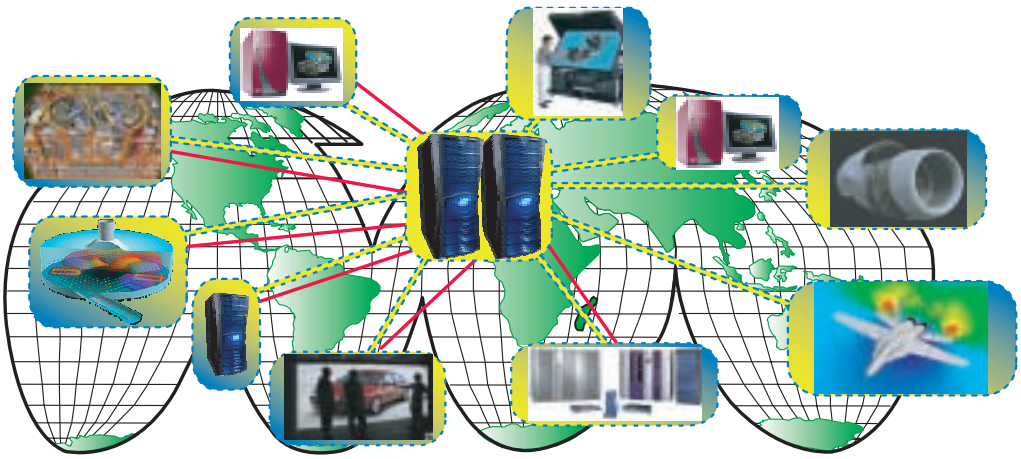

Fig. 1. Grid computing creates an infrastructure of shared resources that can be accessed transparently from any point on the grid.

A grid can make a diversity of resources available to tackle both classes of problems. Grids are by nature heterogeneous, offering a range of resources from multiple vendors such as large commodity clusters, large single system image (SSI) machines, smaller servers, and advanced visualization systems.

Because grids emphasize distributed resource sharing—often crossing organizational boundaries—issues such as scheduling, accounting, access control, and security are critical. As a result, the Globus Project™ (*www.globus.org*) released Globus Toolkit™, a suite of open source services for the grid. Globus Toolkit, developed by USC's Information Sciences Institute (ISI), the University of Chicago's Distributed Systems Laboratory, and Argonne National Laboratory, is becoming a de facto standard for grid middleware, providing many of the basic protocols and services required for reliable and secure operation of heterogeneous grid systems.

Middleware like the Globus Toolkit solves some of the challenges of grid computing, but others remain:

- System and application scalability remains an important issue. The largest capability problems consistently push the limits of available technologies. Different problems may require different programming models for the best solution.
- Grid users must often share data between heterogeneous systems with high performance and over significant distances. Technologies that can do this while ensuring data integrity and security are needed.
- The grid is only as secure as its weakest link. Secure middleware is of little value if grid-attached systems are not themselves secure.
- Visualization provides unique benefits for many scientific and engineering problems, but typical visualization systems only support local viewing. A solution to bring remote visualization to the grid would dramatically expand the use and benefit of advanced visualization.

1.2 SGI on the Grid

As a recognized leader in high-performance computing and advanced visualization, SGI has been involved in grid computing from the beginning. SGI has a long history of achievement in grid computing:

- The first public demonstration of grid technology was at SuperComputing '97 and powered exclusively by SGI® systems. The demonstration was led by Argonne and USC.
- Development of the Globus Toolkit was done entirely on SGI systems using the SGI® IRIX® operating system and associated development tools.
- The NASA Information Power Grid—NASA's distributed high-performance computing grid—is powered entirely by SGI and includes the largest SSI, shared-memory system ever created.
- SGI systems are used in almost all of the major grid installations in Europe, North America, Japan, and Australia.

SGI is a member of the Global Grid Forum and is actively involved in its Working Groups and Research Groups. In June 2002 SGI announced an alliance with Platform Computing, a leader in distributed and grid computing software solutions. Under the terms of the agreement, SGI and Platform will work together to deploy Platform™ Grid Suite and Platform Globus™ with SGI grid solutions, including the SGI® Origin® server family, the SGI® Onyx® 3000 series, the SGI® Onyx® 300 visualization system, OpenGL Vizserver™ computing solution, and SGI® CXFS™ shared filesystem. The partners will also collaborate to develop new grid solutions that integrate data, compute, and visualization resources.

SGI faces and solves many of the important challenges of grid computing—such as scheduling, bandwidth, data distribution, and coherency—every day. SGI computer systems, storage systems, networking technologies, and software provide unique capabilities for grid computing:

- SGI offers the largest available SSI, shared-memory supercomputers with proven scalability to tackle the most intractable capability problems. Smaller configurations are well suited for capacity computing.
- With CXFS, SGI provides high-performance, shared access to data over storage area networks (SANs). This data can be accessed by systems running IRIX, Solaris™, or Windows NT® with support planned for additional platforms.
- Trusted IRIX™ system software provides a B1 level of security, with safeguards against internal and external threats that exceed the protections available from other UNIX® operating systems. Trusted IRIX is ideal for cross-organizational grid computing where security is extremely important.
- With Visual Area Networking (VAN) technology, SGI brings the power of advanced, remote visualization to the grid. Grid users can view the output of advanced visualization systems anywhere on the grid. Multiple grid users can collaborate by viewing and interacting with the same image at the same time in different locations.

The remainder of this paper discusses these technologies and their applicability to grid computing.

2.0 The Origin Platform: Shared-Memory Supercomputing

The SGI Origin platform is based on the patented SGI® NUMAflex™ architecture, providing an SSI and shared

memory for Origin configurations ranging from 2 to 512 processors. In an Origin system, 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. Programming many types of capability problems is simpler with shared-memory programming models and performance is substantially better. SSI makes Origin systems easier to manage than clusters of comparable size.

The patented SGI NUMAflex architecture makes it possible to scale the number of processors in Origin systems 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).

Fig.2. NUMAflex

A modular high-performance computing architecture concept composed of any number of the following elements:

- C-brick—Up to 400GB of high-performance disk storage per brick
- D-brick—Up to four processors and 8GB of memory per brick
- P-brick—PCI expansion
- X-brick—XIO™ expansion
- G-brick—Advanced graphics and visualization
- R-brick—Routing and exceptional bandwidth between processors and memory in different C-bricks
- I-brick—Base I/O providing system disk, CD-rom, and network and I/O ports

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 systems provide an extremely small physical footprint relative to other comparable systems because of the efficient modularity of NUMAflex.

2.1 The SGI® Origin® 3000 Series for Capability Computing

Origin 3000 series systems are SGI's premier supercomputing platforms, supporting up to 512 processors, 1TB of memory, and 716GB per second of system bandwidth. All Origin systems support shared-memory programming models and also support the massively parallel interface model, the programming model used on clusters.

The shared-memory programming models available on SGI systems provide a big advantage for many classes of computing problems. For example, NASA has turned increasingly to SGI in recent years to provide production supercomputers. For NASA workloads such as computational fluid dynamics simulations of advanced aircraft and spacecraft or climate simulations, shared-memory, Origin family systems based on SSI have proved superior to large, clustered supercomputers. SGI systems provide far greater performance on NASA projects with less programming effort and at significantly less total expense. Jobs that previously took months to execute on NASA's Cray C-90 vector supercomputers can now be completed in a matter of hours. NASA scientists now focus more attention on science and less on computer programming.

2.2 SGI® Origin® 300 for Capacity Computing

SGI Origin 300 offers the same capabilities as the Origin 3000 series in a smaller package. Scaling from 2 to 32 processors and 32GB of system memory, the Origin 300 server is an ideal platform for tackling capacity problems or smaller capability problems.

At Cardiff University, both academic and commercial users from a wide variety of disciplines, including engineering, physics, earth science, bioscience, and chemistry, access grid resources for HPC and visualization. A 32-processor Onyx 300 system with three graphics pipes is used for capability problems, while four 8-processor Origin 300 servers are available to handle capacity problems. Remote visualization on the grid is enabled using SGI OpenGL Vizserver. [See section 5, "Visual Area Networking," for details.]

3.0 Powering Heterogeneous SANs

For large capability problems, data I/O is just as important as computing capabilities. SGI offers storage area network solutions and industry-leading software to meet the most demanding storage needs. SGI has more experience deploying high-performance SAN technology than any other system vendor. SGI was the first system vendor to ship Fibre Channel storage and the first to ship a complete SAN fabric. SGI also developed CXFS, the first shared filesystem for the SAN, and more recently was the first system vendor to begin shipping 2Gb-per-second Fibre Channel technology. Using SAN technology, SGI has demonstrated I/O throughput in excess of 7GB per second to a single system.

In addition to enabling shared access to computing resources, grids are, in many cases, intended to provide shared access to important data. Storage solutions from SGI can help make that goal a reality.

3.1 CXFS: High-Performance Shared Data Access

Shared, fast, and reliable data access is a critical component of successful grid computing. Without it, data either has to be copied over the grid to local storage each time it is needed—creating potential issues with data security and consistency—or slow network file-sharing technologies like NFS must be used. In either case, valuable time is wasted waiting for data.

CXFS provides high-performance shared data access to the same files and filesystems to any system on a SAN. Since a SAN can span tens of kilometers, CXFS is an ideal complement to grid computing when used in local grids such as campus environments.

For greater distances, latencies are too long to make the use of CXFS practical in most cases, so data must still be copied. This is less of a problem than it may seem, since most grids consist of centers of localized activity. If each node in a large grid has a local copy of a given data set, that data can be shared locally using CXFS, minimizing the number of copies that exist while optimizing I/O performance for each node.

CXFS provides the ability to make data accessible where it's needed for capability problems without copying. Capacity problems that require shared access to the same data set also benefit immediately from the high-performance data sharing provided by CXFS. For its grid-computing environment, Cardiff University installed a SAN with 2TB of RAID storage and CXFS. This storage system provides direct high-speed data access and data sharing for its 32-processor capability system and four 8-processor capacity systems.

SGI designed CXFS for applications where shared data access is critical and local area networks (LANs) simply cannot provide the necessary bandwidth. CXFS gives all SAN-connected systems simultaneous high-speed access to the same filesystems and files. A single system can have multiple connections, making it possible to achieve data rates of multiple gigabytes per second.

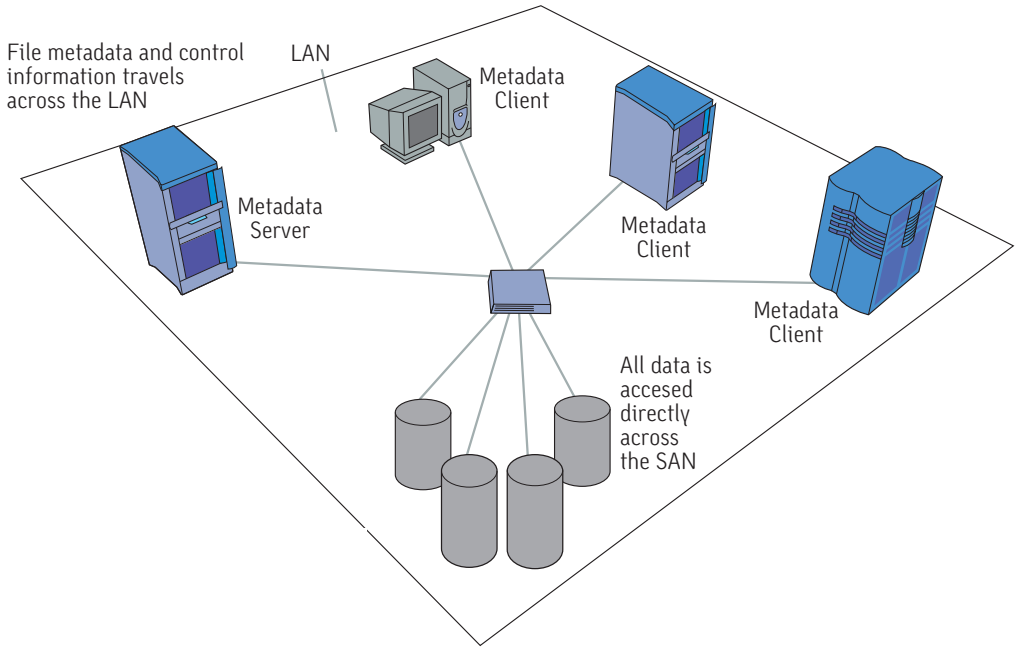

Fig. 3. CXFS is an ideal storage solution for intra-grids or grid-attached HPC centers, since it enables data sharing at SAN speeds.

One system on the SAN acts as a metadata server, controlling file permissions and mediating shared access. Unlike network file sharing, where all data goes through the file server (which often becomes a bottleneck), once the metadata server grants access, systems with CXFS read and write data directly over the SAN to and from disk.

Should a metadata server fail, a designated backup metadata server automatically takes over management of the CXFS filesystem. This feature—in combination with fully redundant SAN configurations and RAID storage—delivers extremely high availability along with exceptional performance. Even if failures occur, CXFS ensures that a path to access data is always available.

3.2 Data Migration Facility (DMF)

Existing grids are already managing huge quantities of data. Since grids maximize the utilization of computing resources, their potential to generate new data and consume storage is very high, making storage capacity and storage management critical issues. SGI® Data Migration Facility is a hierarchical storage management (HSM) system that is in use as an adjunct to online storage at a large number of HPC facilities. DMF is the industry's leading HSM product. With DMF, a storage pool of nearly unlimited capacity can be created, making it an ideal solution for storage-hungry grid-attached HPC centers. Within such a center, multiple systems can use DMF to transparently migrate unused data to tape. That data can be recalled for access by local systems or for transfer to other grid-attached locations.

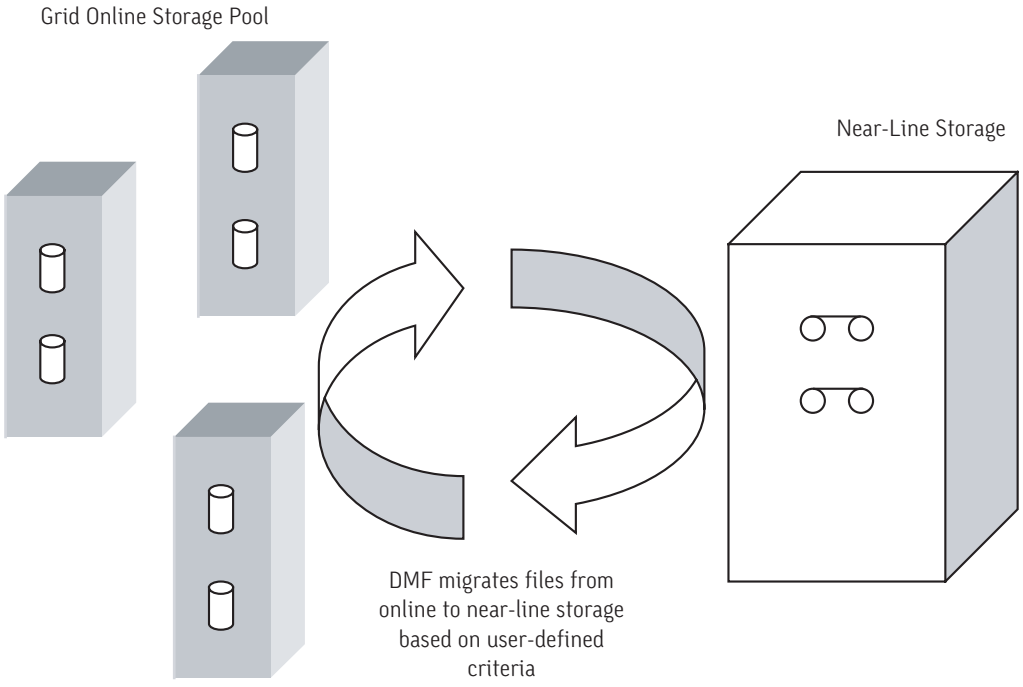

Fig. 4. DMF provides hierarchical storage management for grid-based storage, creating a virtual storage pool with capacity far in excess of available online storage.

DMF automatically migrates data from online storage to tape-based storage according to user-defined criteria. Files are automatically recalled to online storage as they are accessed without user or system administrator intervention. DMF allows access to a nearly unlimited pool of data without concern for the medium on which that data is stored. By comparison, manual data archiving solutions eat up valuable time determining which data to archive and then copying that data to tape. Restoring data from tape for later use is similarly time consuming. DMF frees users to focus on the scientific or technical problem under study by eliminating the need for manual data management.

The COSMOS Project at Cambridge University uses 1.6TB of SAN-based RAID storage with CXFS for shared data access. A 4TB tape library on the back end is managed by DMF, creating a large virtual storage pool that significantly reduces data management tasks, allowing cosmologists to spend more time thinking about the universe and less time worrying about data storage.

4.0 Security for the Grid

Because grid computing may involve the sharing of critical data between systems in different organizations, security can be extremely important. As grid technology becomes more widely adopted by private industry, the need for security will increase even more. The security of a grid is only as good as its weakest link. Secure middleware does little good if the systems on the grid are themselves insecure. Therefore, operating system security is a key element in overall grid security.

SGI has an impressive history of trusted systems expertise based on its long involvement with the federal marketplace. Trusted IRIX is a secure version of the SGI IRIX operating system that adds security features to standard IRIX without diminishing any of the other capabilities of IRIX.

The Trusted IRIX operating environment is a separately purchased add-on product that was developed to conform to the standard Trusted Computer Security Evaluation Criteria (TCSEC) B3 feature set, but with the assurance of security at the B1 level. A standard installation of IRIX is a C2-level deployment. Standard IRIX features include:

- Identification and authentication
- Capability-based privilege mechanism
- Superuser-based privilege mechanism
- Discretionary access control
- Object access control list
- Hardware object scrubbing
- Activity audit trailing

Trusted IRIX confers the additional feature of mandatory access controls (also known as mandatory object sensitivity/integrity). The B1 rating requires several features not present in standard UNIX systems and requires review and modification of existing codes. Added features include improved user identification and authentication procedures, audit records of all system activity, and more stringent access controls on files and devices. Because of these enhanced security features, Trusted IRIX is an appropriate solution for grids that require high levels of security.

5.0 Visual Area Networking (VAN)

Simulation and visualization are critical tools for the understanding of difficult problems in science and engineering. As a recognized leader in advanced visualization, SGI is continually exploring new ways to bring the benefits of visualization to the widest possible audience. With OpenGL Vizserver, SGI's flagship product for VAN, SGI now has the capability to deliver the visualization power of advanced SGI® Onyx® family systems to client systems anywhere on the grid. This means that visualization can be used without having to physically move to a system that provides advanced visualization (which in many cases requires travel) and without the need to copy data for visualization on a local system.

OpenGL Vizserver is a client/server application that uses an SGI Onyx family system as a visualization server. Graphics are rendered using the advanced capabilities of the Onyx family system, compressed, and delivered over the grid to virtually any client display. Since the client displays pre-rendered graphics, it requires no graphics acceleration. OpenGL Vizserver can be used for single sessions or collaborative sessions in which multiple grid users receive and interact with the same visualization stream.

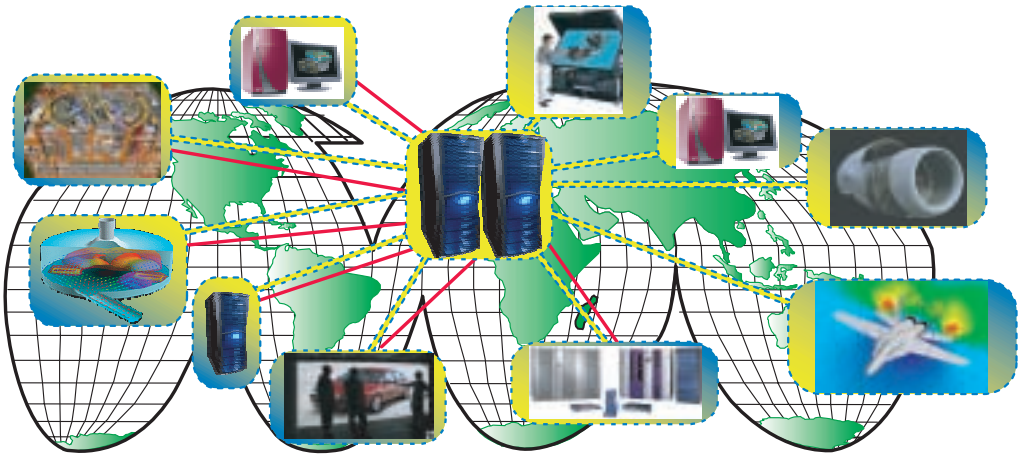

Fig. 5. Visual Area Networking on the grid: Users anywhere on the grid can view advanced visualization output created by an SGI Onyx family system acting as a grid visualization server. No data is required at the user's location.

OpenGL Vizserver contains a number of grid-enabling features including compression, authentication, reservations, and accounting. The features of OpenGL Vizserver are summarized in table 1.

Table 1. OpenGL Vizserver Features

Supported Visual Servers	Silicon Graphics® Onyx2® with InfiniteReality® graphics, SGI Onyx 300 with InfiniteReality3™ graphics, Onyx 3000 series with InfiniteReality3 or InfinitePerformance™ graphics
Supported Clients	Silicon Graphics® workstations (IRIX® 6.5.5+), Sun™ workstations (Solaris™ 2.6+), workstations with Intel® Pentium® III or better and Red Hat® Linux® 6.2, Microsoft® Windows NT® 4.0, Windows® 2000, or Windows® XP
Compression	4:1, 8:1, 16:1, 32:1, or compression API
Frame Spoiling	
Authentication API	Allows integration in a wide variety of authentication environments
Reservation System/Reservation API	Users can reserve time on visualization server; API allows integration with existing calendar systems
Dynamic Pipe Allocation	Waiting sessions are initiated whenever a pipe becomes dormant
Accounting	A full per user usage log for planning, billing, etc.

Advanced visualization systems are usually a fixed and limited resource. The combination of OpenGL Vizserver and the grid can greatly increase the utilization—and thus the benefits—of visualization.

For more information on Visual Area Networking in grid environments, see the companion white paper titled "SGI on the Grid: Visual Area Networking in Grid Environments."

6.0 Conclusion

Grid computing is the next step in the evolution of networked computing. SGI has been involved with networked computing and grid computing from the beginning, and is already powering many of today's largest grids. SGI offers advanced technologies for HPC, advanced visualization, data management, and security that make the promise of grid computing a reality.

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

3307 (07/18/2002) J14013