sgi®
INNOVATION
FOR RESULTS™

White Paper

SGI® Altix® ICE Integrated Blade Platform:
A New Era in High Performance Computing

Table of Contents

1.0 Executive Summary

The SGI® Altix® ICE platform is a high performance computing (HPC) system designed for the most demanding scale-out workloads. This innovative new product line from SGI is based on a new blade architecture specifically designed by SGI to meet the unique needs of the HPC market. SGI Altix ICE forges a new approach to high performance computing (HPC). It provides breakthrough scalability, manageability, reliability, and price/performance without the compromises inherent with either traditional cluster or symmetric multi-processing (SMP) systems.

While small-node clusters can provide real value and flexibility at the small-to-medium scale, they are not designed for large-scale deployments where they falter on several critical issues: price/performance, data center resource efficiency, reliability, and manageability. On the other hand, SMP systems, known for manageability, scalability, and functionality, address processing requirements in more challenging HPC environments, but are not able to take advantage of the cost economics of commodity components. SGI Altix ICE forges a new path between these two approaches and offers an ideal solution for high performance, scale-out computing. SGI Altix ICE provides a manageable, reliable, efficient, and highly scalable environment at an unprecedented price/performance level.

2.0 The Challenges of Scaling Today's Computing Environment

The growth of high performance computing over the past several years has been fueled by the rapid adoption of clusters, driven primarily by their relatively low up-front cost and seeming ease of scalability. As technical organizations attempt to scale their cluster environments to address ever more complex problems, however, they find themselves hitting a wall on both price/performance and scalability. While traditional small-node clusters prove useful and cost effective for small and even medium size application workloads, users need to weigh the initial advantages of quick deployment and extensibility against the increasing complexity and cost that goes hand-in-hand with scaling these types of systems past a certain limit.

This increasing complexity of HPC workloads has created an ever widening gap in user productivity versus performance. SGI Altix ICE was designed to close this gap, and was purpose-built to efficiently handle true HPC applications and large scale-out workloads. When evaluating a potential HPC solution, users should carefully consider bottom-line effectiveness in four key areas:

- Price/performance
- Data center capacity – power/cooling/space requirements
- Reliability
- Complexity and manageability

As clusters grow in size, the hardware and software costs for provisioning the nodes, deploying the applications, and managing, monitoring, and tuning the entire system to achieve anticipated high performance goals can escalate dramatically, virtually eliminating the up-front cost-savings and anticipated simplicity offered by commodity cluster solutions. A different approach is needed to address the needs of larger scale, production-ready scale-out computing.

3.0 Introducing SGI Altix ICE:
The Benefits of Clusters Without the Compromises

SGI Altix ICE is based on an innovative new board design, resulting from collaborative efforts by SGI and Intel. Designed for density, this new board enables each SGI Altix ICE blade to accommodate up to eight Intel® Xeon® processor cores. The SGI Altix ICE rack can support four blade enclosures, scaling to 512 processor cores per rack, and 1000s of nodes per system.

SGI Altix ICE uses dual high speed 4X DDR InfiniBand interconnects, integrated into blade enclosures. Each blade enclosure includes two switch blades, thus eliminating external switches altogether. The system can easily scale to thousand of nodes, reducing system cost and improving performance for many workloads.

In addition to its tightly integrated, yet highly scalable, hardware architecture, SGI Altix ICE ships with a complete, standards-based software solution stack, for maximum out-of-the-box functionality at a competitive price point.

SGI Altix ICE offers a robust solution to the challenges of scaling in high performance computing. Its wealth of capabilities, offered at a competitive price point, make it a revolutionary alternative to the hidden costs of cluster-based scale-out systems. The SGI Altix ICE platform design deftly handles the critical issues inherent to cluster-based computing by offering unprecedented capabilities in the areas of:

- Price/performance value
- Power/cooling/space efficiencies
- Reliability and serviceability
- Simplified management and scalability

3.1 Price/Performance Value

The SGI Altix ICE blade provides exactly the hardware necessary to get the job done. Everything on board serves to boost performance; there's nothing extraneous. The integrated blade approach enables the SGI Altix ICE platform to focus all resources exclusively on delivering maximum performance for size and cost. The attention to paring away the non-essential results in a highly dense platform, supporting up to 512 Intel Xeon processor cores and up to 6 teraflops per rack. With its unmatched performance density and low node cost, SGI Altix ICE offers an aggressive price/performance value proposition, especially in large-scale configurations.

SGI Altix ICE includes software enhancements to address a key performance issue often encountered in parallel systems: operating system synchronization. An SGI-engineered software mechanism synchronizes operating system overhead – OS jitter and noise – to improve performance significantly on parallel workloads.

The SGI Altix ICE compute network also contributes significant performance boost. The dual-plane network runs on 4X DDR InfiniBand in a hypercube topology, providing maximum band-width for both MPI and I/O traffic. With its high speed and low latency characteristics, InfiniBand enhances the performance of the network, and thus the system as a whole.

	Gigabit Ethernet	DDR InfiniBand
Bandwidth (MPI)	85 MBps	1860 MBps
Latency	30 uSec	1.1uSec
Link Rate (Bidirectional)	1 Gbps	20GBps

Table 1. Interconnect Performance Characteristics

To further boost throughput, management communications run across a separate GigE administrative network, segregated entirely from the InfiniBand compute network.

The basic SGI Altix ICE unit is the cable-free blade enclosure, which supports up to 16 blades (128 processor cores). A single SGI Altix ICE rack accomodates up to four blade enclosures; multiple racks can be easily networked together to support larger configurations. The onboard interconnect, in combination with integrated InfiniBand switches, serves to minimize the cabling, as well as the hops and latency, even for very large installations. For example, a single rack, 128 socket (256 or 512 core) installation requires a total of only 24 InfiniBand cables. Its maximum hop count is just three and its MPI latency is 3,446 nS.

The SGI Altix ICE value proposition extends far beyond the price/performance story to encompass virtually every aspect of its design and implementation. The sections that follow describe other SGI Altix ICE key strengths.

3.2 Power/Cooling/Space Efficiencies

The design of the SGI Altix ICE platform takes into account the environmental constraints of today's data centers and offers high efficiency solutions to meet the challenges.

Today's data centers are in a state of crisis. AFCOM's Data Center Institute predicts power failures and power availability will halt IT operations at more than 90% of companies over the next 5-years (AFCOM, 2006). Built years ago for a very different computing environment, many data centers today suffer from limited power, cooling, and space capacity. The environmental inefficiencies of traditional clusters (low density, combined with high power and cooling requirements), barely noticeable on a small scale, become prohibitive when scaling larger installations.

SGI has long been a technological leader in solutions optimizing power and cooling efficiency, mostly recently with its innovative SGI® Altix® 4700 server platform, where it introduced both a power architecture featuring 90% efficiency power supplies and field-proven water-based cooling system. These innovations have been further optimized and applied to SGI Altix ICE. Together, they represent a major advance in overcoming the constraints of the data center.

SGI Altix ICE utilizes 90% efficiency redundant power supplies that transform AC voltage directly to 12VDC. These high efficiency power supplies are combined with other high efficiency components to minimize losses throughout the entire power architecture. This high level of efficiency results in average electrical savings of 33%, or $21k annually per 10 teraflops of compute power (based on an electricity cost of $0.092/kWh), compared to more typical cluster implementations. If data center facility infrastructure efficiency is also considered, the annual electrical savings doubles to $42k per 10 teraflops of compute power. Data center infrastructure efficiency, commonly described by Power Usage Effectiveness (PUE) = {Total Facilities Power / IT Equipment Power}, is typically 2.0 but ranges from 1.6 to 3.0 or higher (The Green Grid, 2007).

SGI Altix ICE employs a combination of high efficiency redun-dant blowers and optional water-cooled rear doors to deliver impressive cooling efficiency results. With the water-cooled option, SGI Altix ICE has minimal effect on ambient data center temperature, since up to 95% of the rack heat is dissipated to chilled water. While actual performance depends on many site and geographic variables, the SGI Altix ICE water-cooled option significantly reduces cooling equipment power consumption. Electrical operating cost can be reduced by 17% or more, amounting to $11k annually per 10 teraflops of compute power. Use of the water-cooled option also increases overall system reliability by mitigating the common problems of hot-aisle/cold-aisle recirculation and hot spots within the data center. SGI Altix ICE addresses the third major constraint of the data center – space – with its breakthrough performance density.

Each 42U (30" W x 40" D) rack holds four blade enclosures with 16 2-socket nodes each, achieving a density of up to 512 Intel Xeon cores, or 6 teraflops per rack. This results in up to 70% higher compute power density per floor tile (based on gigaflops per square foot) compared to other blade systems. Despite this performance density, the fully loaded SGI Altix ICE rack stays within data center flooring constraints, with a footprint of 246 lb/ft^2.

SGI Altix ICE performance density, combined with its innovative and highly efficient approaches to power and cooling, ensures maximum utilization of scarce data center resources.

3.3 Reliability

Reliability is another area for concern when venturing into larger scale clusters. Compute nodes and other components inevitably malfunction over time, and cluster installations often lack sufficient redundancy to deal robustly with component failures. The networks tying the cluster nodes together can also suffer from reliability issues that grow exponentially as clusters scale. The complexity of scale-out cluster environments, with their multiplying points of failure, leads to further reliability problems.

SGI Altix ICE, with its focus on component integration along with other innovative features, achieves a new standard of reliability for scale-out configurations. Key reliability features include:

- Diskless blades with automatic hot-swap capabilities
- Cable-free blade enclosures (individual rack units), for reduced potential points of failure
- Redundant, hot-swappable system components
- High-efficiency power architecture, for reduced heat dissipation
- Optimal thermal design
- Dual InfiniBand planes for maximum performance and network failover capability
- InfiniBand backplane, for high signal reliability

The SGI Altix ICE blade has no onboard disk, instead using diskless booting from the chassis management controller (CMC) at the blade enclosure level. This innovation offers a number of benefits. It increases blade reliability by eliminating a key failure point. It increases performance density by reducing the blade footprint. It lowers the cost of powering the blade, as well as reducing the cost of the blade component itself. The SGI Altix ICE off-blade storage design also provides an enhanced level of serviceability and administrative control, especially beneficial for managing secure or classified data. This reduces the problems often encountered with the management of classified information in a distributed environment.

The SGI Altix ICE cable-free blade enclosure and integrated switching provides another significant boost in overall system reliability. The SGI Altix ICE approach to interconnect stands in sharp contrast to the cable chaos endemic to cluster systems. This clean and lean design markedly increases both the reliability and the ease of serviceability for the overall system.

Traditional Cluster Cabling

Multiple SGI Altix ICE Racks, Reduced Cabling

To further enhance reliability and maximize uptime, SGI Altix ICE includes redundant, hot-swappable power supplies and blowers. Each blade enclosure includes (7+1) 1625W 12VDC output front-end power supplies and (7+1) 175mm blowers.

The high efficiency power architecture described earlier reduces heat dissipation and associated temperature rises within the system. By so doing, it decreases the likelihood of a common cause of system failure – unsustainable temperatures in crowded data centers.

The reliability of SGI Altix ICE is further enhanced by its optimal thermal design. A common issue with blade-based architectures is the crowding of I/O options across the backplane or mid-plane of the enclosure. The result is often a complex series of baffles and airflow management structures that require small fans or blowers. SGI has greatly improved over other blade-based designs by placing the I/O switches to the sides of the enclosure, enabling large perforations in the backplane. This enables air to flow straight through the compute blade, increasing volume and enabled the utilization of larger, more efficient, and less noisy blowers. The resulting thermal design permits more efficient and consistent cooling across all components, increasing reliability and overall efficiency.

The SGI Altix ICE compute nodes employ Fully Buffered DDR2 DIMM (FB-DIMM) memory, which enhances reliability and performance. Memory data reliability is ensured with enhanced error-correcting codes to allow every node to independently detect and correct errors.

The SGI Altix ICE platform uses a dual-plane InfiniBand backplane that increases overall MPI bandwidth by aggregating bandwidth over two IB networks, and leverages InfiniBand features such as data re-transmission to overcome packet drops and further enhance overall reliability.

3.4 Simplified Management and Scalability

SGI Altix ICE, with its emphasis on component and software integration, sets a new standard for simplicity in the world of scale-out environments. On the hardware side, the clean design of the blade enclosure, with its integrated blades, switches, and interconnect, stands in welcome contrast to the ad hoc maze of many cluster-based systems. This integrated approach continues on the software side. SGI Altix ICE ships with an integrated, complete software solution stack. The standards-based stack, supplemented by SGI tools to simplify system management and enhance performance, ensures fast ramp-up and deployment.

For enhanced scalability and simplified management, SGI Altix ICE is based on a hierarchical management design. This hierarchical management architecture provides a high degree of modularity, enabling SGI Altix ICE installations to scale to very large sizes while maintaining management granularity at every level of the system: node/blade enclosure, rack, and system. This means that administrators can scale the management resources that they need, when they need it – and can also provision, monitor, and service resources at any of these levels.

The entire SGI Altix ICE system undergoes factory integration and testing. SGI Altix ICE arrives fully integrated and ready for out-of-the-box deployment. SGI's more than 20 years of experience in delivering "power up and go" systems ensures immediate productivity.

The simplicity of the platform also simplifies the process of scaling the platform. SGI Altix ICE component integration, minimal cabling, high performance InfiniBand interconnect, and hierarchical management nodes, supported by its suite of software management tools, translates into easy scalability for even the most compute-intensive applications.

4.0 SGI Altix ICE Software Solution Stack

SGI Altix ICE ships with a complete, integrated software stack. The SGI Altix ICE cluster solution includes the full suite of tools that users need to build, deploy, and make the best use of their SGI Altix ICE system. The result is a powerful and cost-effective solution, designed to ensure that SGI Altix ICE users become productive immediately:

Operating System	SUSE® Linux® Enterprise Server 10, Red Hat® Enterprise Linux Server
Performance Optimization	SGI® ProPack™
Platform Management	SGI® Tempo Management Tool or Platform™ Manager (formerly Scali Manage)
Workload Manager	Altair® PBS Professional™
IB Fabric and Subnet Management	SGI InfiniBand Fabric Subnet Management (based on OFED and OpenSM)
Development Tools	Intel C++ and Fortran compilers, VTune, Math Kernel Library

Table 2. SGI Altix ICE Software Solution Stack

SGI Altix ICE runs on the SUSE Linux Enterprise Server or Red Hat Enterprise Linux Server operating systems. The platform builds on SGI's leadership position in the Linux community, unmatched in the industry. SGI has been, and continues to be, a major contributor to the Linux standard, and brings a wealth of experience and expertise to resolve customers' kernel-level issues quickly and efficiently. With the combination of SGI and Linux, SGI Altix ICE offers a scalable, robust, and standards-based software platform.

SGI ProPack 5, SGI's optimization software package, extends the Linux standard with tools to enhance system administration, development, and performance. These tools include linkless FFIO to accelerate I/O calls, resulting in dramatic performance enhancement for I/O intensive applications.

SGI Altix ICE administrators have two options in managing their SGI Altix ICE environment: Platform Manager or SGI Tempo. Platform Manager (formerly known as Scali Manage) is an industry standard cluster management tool for high-performance systems. Sites already running Platform Manager for other systems will appreciate the ICE integration so there's no requirement to switch tools. SGI Tempo, a management tool designed exclusively for SGI Altix ICE simplifies the management of large-scale environments. It enables and manages basic software provisioning, system monitoring, discovery, logging,

and reporting, as well as system and node level command and control. SGI Tempo is used to manage the SGI Altix ICE hierarchical management node infrastructure. It also provides a boot mechanism for rapid parallel booting of the compute nodes.

The feature-rich SGI Tempo management tool draws certain components from the open source community with substantial SGI enhancements for scalability to 50+ racks and thousands of compute nodes. This means that customers familiar tools such as OSCAR should be able to transition easily to SGI Tempo management tool, and still be able to use extensions that they have developed for use with OSCAR.

OSCAR. Support for third party software tools completes the stack. Altair's PBS Professional workload management software allows scheduling and management of jobs and resources, resulting in a higher total utilization of the cluster.

5.0 Integrated Storage

SGI Altix ICE is based on a "diskless node" architecture that removes storage from the compute blades, a design that decreases cost and power/cooling requirements while at the same time increasing overall system reliability. In addition, by moving the storage off the compute blades, SGI Altix ICE allows customers to choose the storage option that best fits their computing environment. Leveraging SGI's comprehensive

SGI Altix ICE Features	Benefit
Integrated Interconnect	Reduced cost and complexity and simplified scalability, with cable-free blade enclosures and integrated switches. Efficiently scales to thousands of nodes without needing to add expensive external switches.
HPC Optimized Compute Blades	Top performance density for optimal data center space utilization. Based on ultra-dense SGI/Intel designed board and Dual or Quad-Core Intel® Xeon® processors – 512 processor cores per rack.
SGI Patented Power Design	Enhanced power efficiency for reduced overall cost of deployment, with +75% power efficiency at rack level.
SGI Water Chilled Doors	Increased reliability, optional features for larger systems maintains optimal operational environmental temperature to reduce overheating and potential system outage.
Hierarchical Management Infrastructure	Simplified scalability and easier management, with ability to manage, monitor and provision at blade, blade enclosure, rack, or system level.
Complete Software Solution Stack	Immediate productivity, with a fully integrated solution that includes the SGI Tempo Management Tool, Platform Manager, SGI ProPack for Linux 5 and Altair® PBS Professional™ Workload Manager.
Industry-standards Based	Fully satisfies IT OS, applications, and security compliancy requirements, and delivers all of the benefits of open industry-standard, with SUSE® Linux® Enterprise Server. Select from an extensive portfolio of 32- and 64-bit applications, with the assurance of industry-leading performance and reliability.
Off Blade Storage	Reduced cost, power/cooling requirements, and increases overall system reliability.
SGI 25+ Years HPC Expertise	Reduced risk, optimal TCO. SGI Professional Services team, rated best by SatMetrix, brings years of industry and technical expertise to help customers with development and deployment, to ensure an optimal solution to precisely meet customer needs, budget, and timeline.
Single Source Support	Simplified administration. All system hardware and software components backed by SGI World-class Customer Service organization.

InfiniteStorage product line, the storage for SGI Altix ICE can be tailored to meet specific application requirements. Another advantage of moving the storage "off-blade" is the centralization of storage resources, promoting better capacity balancing, easier management, and on-line scaling of capacity.

SGI Altix ICE features a low cost/high performance ratio storage transport protocol architecture and supporting NAS product line to help optimize price/performance value. The SGI Altix ICE architecture extends to a purpose built storage platform featuring NFS/RDMA for InfiniBand (IB) connectivity. The SGI InfiniteStorage NEXIS 2000 and NEXIS 9000 NAS appliances connect over an InfiniBand network to Altix ICE:

– **InfiniteStorage NEXIS 2000** offers a balance of performance and affordability for entry level HPC (SAS) or nearline (SATA) storage applications. Exceptional sequential performance of up to 475 MB/sec 100%write and 870 MB/sec 100% read measured with standard benchmark tools (XDD).
– **InfiniteStorage NEXIS 9000** is the high end NAS performance model with optional NFS/RDMA over IB providing up to 2,275 MB/sec 100% write and 3,200 MB/sec 100% read performance measured with standard benchmark tools (XDD). It enables flexible HPC client support consolidating with GbE, 10GbE and InfiniBand connectivity on a single node and can be expanded to support up to 448 SAS drives.

NFS/RDMA, (Remote Direct Memory Access) as an enabler for NAS in the high end provides significant performance optimization. NFS/RDMA over IB allows server and client communication using an RDMA capable transport. This results in a significant reduction in CPU utilization and an increase in throughput. Key benefits of this type of approach include:

• Increased data bandwidth
• Reduced CPU overhead
• Reduced latency
• Improved scalability
• Reduced client overhead

In addition SGI offers a wider choice of storage solutions for block and file data that can be implemented by SGI's industry-leading professional services group.

6.0 Summary:
A New Era in Performance Computing
SGI Altix ICE was designed with High Performance Computing in mind, and delivers a unique blend of "performance" and "productivity". This innovative new platform from SGI raises HPC bar, easily scaling to meet virtually any processing requirements without compromise on ease of use, manageability, or price/performance.

7.0 About SGI
SGI is a leader in high-performance computing. SGI delivers a complete range of high-performance server and storage solutions along with industry-leading professional services and support that enable its customers to overcome the challenges of complex data-intensive workflows and accelerate breakthrough discoveries, innovation, and information transformation.

SGI helps customers solve their computing challenges, whether it's enhancing the quality of life through drug research, designing and manufacturing safer and more efficient cars and airplanes, studying global climate, providing technologies for homeland security and defense, or helping enterprise manage large data. With offices worldwide, the company is headquartered in Sunnyvale, California, and can be found on the Web at www.sgi.com.

8.0 References
AFCOM, 2006. "Five Bold Predictions for the Data Center Industry That Will Change Your Future", AFCOM's Data Center Institute, Data Center World Conference, Atlanta, GA, March 2006.

The Green Grid, 2007. "Green Grid Metrics: Describing Data Center Power Efficiency", The Green Grid, February 17, 2007. See http://www.thegreengrid.org/pages/content.html.

6