White Paper

SGI® Altix® 4700 and 450 Servers Reliability, Availability and Serviceability

Optimizing RAS for the World's Most Versatile Linux Computing Platform

Table of Contents

1.0 Introduction

SGI Altix combines the power of Intel® Itanium® 2 processors and industry-standard Linux with the SGI NUMAflex™ architecture and global shared memory to create a platform that is powerful, uniquely flexible, and highly reliable. SGI has a rich heritage designing, building and deploying high performance systems for the most demanding workloads, resulting in continuous improvement in reliability across the entire spectrum of SGI systems from small systems to large. As of 2007, SGI has deployed thousands of Altix servers serving the needs of highly demanding and mission-critical users in government, research, and the enterprise.

The architecture and fundamental design parameters of Altix systems provide an important and unique value with direct advantages for large-scale, data-intensive technical and enterprise computing. Altix provides the ability to scale global shared memory, compute power, and I/O bandwidth from very small to essentially unlimited sizes. The SGI Altix 4700 system is capable of supporting multiple terabytes of shared memory, 1,024 processor cores under a single instance of Linux, and 10GB/sec or more of sustained bandwidth to a single file system. The SGI Altix 450 scales to 76 processor cores and 608 GB of shared memory, making even this mid-range server a uniquely scalable platform. Both systems use a compact, blade design for space efficiency and ease of maintenance.

With the rapid increase in system memory capacity and CPU count combined with growth in the size of running jobs and associated data sets, reliability, availability, and serviceability—RAS—is becoming increasingly critical for the Altix platform. Jobs that run for days or weeks necessitate a system platform designed to ensure execution success.

Even as SGI pushes system sizes to new heights, the company is bringing new levels of RAS to high-end Linux environments. SGI's RAS efforts leverage its unique experience in building the world's largest and most robust server systems combined with focused investments to enhance memory reliability for our large, shared memory environments.

This white paper describes the RAS capabilities of the SGI Altix platform.

2.0 Altix: Designed and Built to be Dependable

When it comes to the design and manufacture of Altix systems, SGI leverages over two decades of experience creating some of the world's largest high performance computing (HPC) systems. SGI has pioneered technologies to increase the number of processor cores supported in a single system image (single instance of Linux) and refined the large, high-bandwidth memory and I/O systems necessary to support the computational power of thousands of processors working in parallel.

The SGI Altix 4700 platform is the culmination of years of focused effort. It is comprised of reliable, modular blades—interchangeable compute, memory, I/O and special purpose blades that deliver unprecedented configuration flexibility. The blade-to-NUMAlink™ architecture of the Altix 4700 enables users to mix and match eight blade options to create a system with the exact capabilities they require—up to a maximum of 1,024 processor cores and over 100 terabytes of globally addressable memory.

To achieve the necessary level of reliability for large-scale systems, SGI engineers pay special attention to logic and circuit design rules as well as comprehensive design verification and test. Even the smallest details, such as the specification of solder rated for low alpha particle emission, are not overlooked in the quest for the greatest possible hardware reliability.

In addition to careful selection of design parameters, SGI engineers specify high quality components necessary to deliver a highly reliable product. SGI sources from top-tier suppliers to ensure high visibility to quality control for high component reliability. SGI maintains ISO 9001 quality management system certification and uses it as part of the supplier selection process. Critical components undergo additional stress or performance testing per SGI engineering's tough specifications. Further, SGI conducts rigorous PFA or "parts failure analysis" testing for component assemblies to identify critical system issues that could impair reliability.

SGI Altix system blades undergo stringent environmental testing during the manufacturing process to ensure reliability. Samples are thermally stressed in oven chambers with simultaneous voltage margining for up to 8 hours, testing the operational limits of every component. Assembled systems are further tested using a comprehensive suite of confidence tests that have been refined and enhanced by engineers with years of experience developing and maintaining high performance systems.

All of these design, manufacturing, quality control and testing procedures assure highly reliable product assemblies. However, SGI takes this one step further with testing and qualification processes that are specific to the system configurations that our customers request. In this way, we assure not only high quality components, but that the system as a whole will function reliably as deployed in the datacenter.

3.0 Designed-in Reliability

3.1 System Components and Data Paths

While not all hardware faults are preventable, much can be done to reduce the number and impact of errors that occur through careful design, and selection of system components that automatically detect and correct errors.

All SGI Altix models utilize Intel Itanium 2 processors. As a result, the platform benefits from the processor's built in RAS capabilities (Table 1) which detect and correct most on-chip errors. As Intel makes improvements to Itanium RAS technology, these features are increasingly leveraged.

Itanium 2 includes an enhanced Machine Check Architecture (MCA) which defines processor, chipset, firmware and operating system roles in error handling. Altix systems with dual-core Itanium 2 processors leverage the Intel Cache Safe technology which disables L3 cache lines when an uncorrectable error occurs, allowing systems to continue operation in the face of errors that would otherwise cause a disruption.

Features	Functions
Intel Cache Safe Technology: automatic cache recovery	Allows processor and server to continue normal operation in case of cache error: automatically disables cache lines in the event of cache memory error
Enhanced Machine Check Architecture: extensive error detection and correction capabilities	Address and data path error correction; system-wide ECC protection; automatic error detection, logging, and correction

Source: Intel Corp. 2006

Table 1. Overview of Intel Itanium 2 RAS features

Altix systems are designed to reliably detect all errors in system memory, directories, and data paths. SGI utilizes ECC (Error-Correcting Code) on all system buses and memories to detect and correct single bit errors and detect double bit errors, plus CRC (Cyclic Redundancy Check) on all NumaLink 4 channels between Altix blades.

3.2 Environmental Monitoring, Power, and Cooling

Another critical element in system reliability is environmental control. Clean power and proper cooling can dramatically improve observed hardware reliability. The Altix has an extensive environmental monitoring and control system to protect hardware operation. Redundancy in both power and fans protects against failures of these components. Variable speed fans ensure that a system always runs at the optimal temperature. While the power-efficient design of Altix makes it an unlikely occurrence, systems are automatically shut down to prevent damage due to over-temperature conditions.

The very high power efficiency of Altix also serves to facilitate cool operation and lower failure rates. SGI has consistently chosen power-efficient components and incorporated them in a system architecture designed to maximize performance while minimizing physical server footprint, power consumption and cooling requirements. SGI has, in fact, made power efficiency a priority since the company launched its Altix line in 2003.

Altix blade servers, clusters and supercomputers feature a power supply and conversion architecture designed to convert AC power to DC voltages with over 90% efficiency (this compares to efficiencies in the 60 to 70% range for other vendors). Further, SGI Altix blades are designed to minimize power loss, with 12V DC blade input voltage requiring only one additional conversion to useable logic-level voltage. Finally, Dual-Core Intel® Itanium® 2 processors are among the most energy efficient on the market, in some cases running at a fraction of the power of competitive RISC CPUs.

The superior energy efficiency of Altix servers means that the vast majority of Altix users do not require water cooling. However, as a result of a long history of deploying very large systems in densly populated datacenter facilities, SGI was an early adopter of water cooling technologies for traditional air-cooled servers, beating our rivals to this advancement in system cooling (Figure 1). Radiator-like cooling coils intercept hot air as it exits each rack, efficiently cooling the air and preventing machine room hot-spots and the common problem of hot-aisle to cold aisle recirculation. This stabilizes the ambient inlet air temperature and results in increased reilability.

Figure 1. Details of SGI water cooling available for Altix 4700 server deployments.

4.0 Availability Features

SGI includes a number of features in Altix systems which enhance overall availability beyond the reliability provided by the base hardware components. As a leader in the deployment of high-end Linux solutions, SGI has done more than almost any other system vendor to ensure the performance and reliability of Linux. Examples of SGI contributions relating to RAS are listed in Table 2. SGI releases Linux enhancements that improve availability to the broader Linux community whenever they are generally applicable.

SGI Contributions to Linux RAS
Memory UCE (Uncorrectable Error) recovery enhancements
Improved hardware error reporting
Reduced panics on double-bit errors
Improved fault containment for cross-partition jobs
Many more in progress for error avoidance and management

Table 2. SGI Contributions to Linux RAS

4.1 System Partitioning

SGI Altix hardware partitioning allows a single system to be subdivided into multiple, logical systems without physical re-cabling. Partitioning capabilities are designed into Altix hardware to ensure ease-of-use and highly reliable operation and a number of unique hardware features increase the robustness of hardware partitioning on Altix versus other implementations. And while partitioning does change the OS configuration, memory used by applications can still be selectively shared across partitions using Altix global shared memory capabilities. Through this technique, an Altix system can be partitioned and run as a cluster for improved availability without losing all of the large-memory advantages of Altix.

Memory Protection. Altix has memory protection built into the SGI-designed chipset that resides in each CPU blade. This feature provides fault containment by protecting each partition from unexpected writes from other partitions. Other systems that lack this hardware feature may be subject to memory corruption if a mis-configured kernel or poorly behaved application attempts to access the wrong memory.

Altix memory protection can be modified to allow cross-partition access to specific memory pages as necessary. For example, XPMEM support in SGI MPI libraries allow the hardware to change memory protection on pages being shared with other partitions so CPUs in other partitions can directly load and store to shared memory without opening up all memory to remote access.

Reset Fences. This capability is also built into Altix hub and NUMAlink 4 router chips to protect a partition from hardware resets occurring in another partition. Reset fences ensure that each partition operates independently and reliably in the face of restarts or hardware and software failures occurring in other partitions, and provides support for concurrent replacement of system modules.

Block Transfer Engine (BTE). The block transfer engine built into Altix memory hub chips provides a reliable way to transfer data between partitions without changing CPU memory protection. This allows partitions to share data via high speed copying if desired. To ensure fault containment, BTE is designed so that a disruption in a remote partition will not crash or hang a partition that is actively performing a block transfer.

Altix hardware partitions can be rebooted independently without affecting operations in other partitions, providing a number of availability benefits:

- Necessary hardware repairs in one partition can be undertaken without disrupting other partitions.
- When upgrades are necessary, rolling kernel updates can be used to update each partition in turn without bringing the entire computing infrastructure to a halt.
- Organizations doing software development and testing can use partitioning to create a development and test environment that closely approximates the production environment. This environment can be re-started as necessary or brought down by ill-behaved software without affecting other partitions.

4.2 Memory Enhancements

Memory errors remain the most common errors experienced by servers. In typical Linux environments, memory configurations are relatively modest compared to Altix. For this reason, SGI is committed to improve the robustness of Linux for large memory systems.

When a memory location is determined to be bad, either by exceeding a threshold number of single-bit errors or by encountering an uncorrectable error, the Altix memory flawing feature allows the OS to mark the page containing that memory as flawed. The operating system avoids using the flawed page, and it is reported to the System Abstraction Layer (SAL) so that the page is permanently retired.

SGI has further enhanced Linux to recover from main memory failures while a memory page is zeroed out. This typically occurs when a process is allocating memory. Other recent enhancements improve error logging capabilities at both the SAL provided by the Itanium 2 Machine Check Architecture and at the Linux level for both SUSE® and Red Hat® Linux. This basic capability improves data capture for error thresholding and predictive failure analysis.

SGI has also enhanced Linux to ensure that if an MCA event due to a hardware failure disrupts the system, the complete hardware state is captured. This improves root cause failure analysis to ensure that the right components are quickly replaced to restore the system to full operation.

4.3 Reliable I/O

When it comes to I/O, SGI has been a leader for many years, particularly when it comes to fibre channel and high performance storage. SGI has pioneered the deployment of high-performance, highly redundant fibre channel storage infrastructures along with the software necessary to utilize them efficiently.

Systems with multiple fibre channel host bus adapters spread across multiple blades and connected either directly or through a fabric to SGI InfiniteStorage RAID arrays result in I/O infrastructures with no single points of failure. Multi-path I/O balances I/O load across channels and shifts the load from a failed port or HBA over to survivors.

SGI created the InfiniteStorage File System XFS, to accommodate the I/O requirements of HPC while providing the reliability of journaling for error recovery and rapid restarts. XFS has been released as Open Source to the Linux community and is available in standard Linux distributions. For shared data access in clusters, the SGI InfiniteStorage Shared File System CXFS builds on XFS to create a highly-reliable, high-performance storage infrastructure that lets cluster members read and write data directly to disk at full SAN speeds. Altix systems that have been partitioned can use CXFS to allow partitions to share access to the same data sets without compromising performance.

4.4 High Availability Cluster Configurations

To further enhance Altix availability, clustered configurations can be designed using SGI InfiniteStorage Cluster Manager for Linux or any of a number of 3rd party High Availability software solutions for Linux. With Cluster Manager, highly available application services can be created that span separate Altix systems, or partitions in a single Altix system. Applications fail over from one cluster member to another should anything affect the running service.

Oracle 10g Real Application Clusters (Oracle 10g RAC) is a specific option available on Altix to provide high availability solutions. Among the features of Oracle 10g RAC are Automatic Storage Management which handles disk failure, and Transparent Application Failover which provides failover for certain applications and functions.

5.0 Maximum Serviceability

The Altix design includes important serviceability features that provide advanced system control capabilities, system health monitoring, online system management and maintenance, and failure analysis. The advanced, modular blade design of Altix enhances serviceability by its very nature, with individual system components easily accessed for service, maintenance or upgrade. The SGI goal is to make most components of the Altix platform serviceable by an administrator with minimal or no system disruption.

5.1 SGI Altix Features to Maximize Uptime During Service

Altix blades are housed in a chassis referred to as an Individual Rack Unit (IRU) – shown in Figure 2. The Altix 4700 IRU enclosure contains up to ten single-wide blades or two double-wide blades and eight single-wide blades. A single Altix 4700 rack houses up to 4 IRUs (40 blades) as shown in Figure 3. Similarly, the Altix 450 features a smaller-scale IRU that houses up to 5 blades, including one slot that can also accommodate double-wide blades. It is available in a half-rack form factor housing up to 4 IRUs (20 blades) as well as a 40 blade tall rack.

Components in each IRU are electronically isolated so that they can be replaced without powering down the IRU. Power supplies (illustrated in Figure 2), fans (located on the opposite side of the IRU), and individual PCI cards can be hot swapped without interrupting the operation of the system or partition containing the component in most cases. If a system has been partitioned, a failed blade in one partition can be replaced without affecting the operation of other partitions. However, the partition containing the failed blade must be shut down while the operation is in progress.

Other related features of Altix help minimize downtime from component failure. First, compute or memory blades can be disabled and a system can run without them until scheduled maintenance becomes possible. Further, the advanced RAS capabilities of Itanium 2 such as Cache Safe minimize the likelihood of a CPU failure. Likewise, individual memory pages can be marked as flawed and retired while operations continue. Processors and memory are always subjected to self-test at boot time and automatically de-allocated if failures occur. The system then boots without the affected resource so that operations can continue.

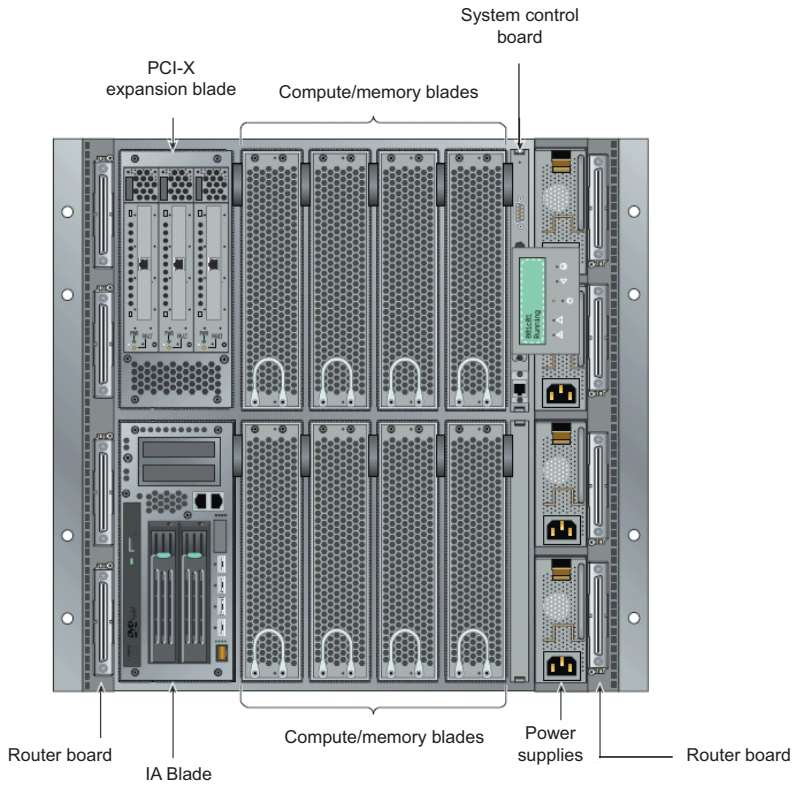

Figure 2. IRU and rack design (Altix 4700 IRU and rack shown)

Figure 3. Altix 4700 tall rack containing 4 IRUs

5.2 System Controller Network

All Altix 4700 and 450 Individual Rack Unit (IRU) enclosures contain an embedded system controller that runs off standby power and is operational whenever the enclosure is connected to an active power source.

The L1 or Level 1 system controller provides control and monitoring functionality for each IRU in an enclosure and communication to other L1s in adjacent enclosures connected via NUMALink 4 cables. The L1 is active even when the system is not booted or powered off.

The console mode L2, or Level 2 system controller provides control over multiple L1s in different system enclosures and peer communication to other L2s. The L2 is resident when an enclosure is connected by an Ethernet connection to a Local Area Network (LAN).

The Altix system controller network manages the hardware partitions within each system, providing pinpoint power control, system booting, and run time system monitoring, along with support for configuration control, hot swapping, and diagnostics execution. The system controller can transparently extract all internal register states and actions from compute, memory, IO, and router blades while the system is running, providing a wealth of input data that allows a fault analyzer to produce failure data reports down to the field replaceable unit (FRU) level.

The system controller is able to read the complete hardware configuration down to the level of individual FRU serial numbers in real-time. This capability supports the rapid and accurate notification and transmittal of essential information for system service actions.

Overall, the system controller network provides the following functionality:

- Power control to the entire system.
- Power control to individual Blades and Dense routers.
- Environmental monitoring.
- Monitoring status and error message information.
- Specific commands to monitor or change system functions.
- Executable diagnostics and system boot control.

6.0 SGI Service Options to Assure Uptime

6.1 SGI Customer Support

As a long-standing global organization, the SGI customer service organization offers a broad range of support, up to and including mission-critical 7x24 system support. The SGI customer service organization consistently ranks among the top in the industry according to the SatMetrix™ third party evaluation.

6.2 SGI MAS (Managed Services) Console

The SGI Solution for Console Server Management is a valuable tool to help system administrators monitor and manage SGI servers during a system-down situation by providing an interface to the system even when network access is not available. It is a combined hardware, software, support and on-site installation package for management of a single or multiple heterogeneous servers.

6.3 SGI UpSafe™ Uninterruptible Power Supply (UPS)

UPS systems are critical to protect electronic equipment from power problems such as blackouts, brownouts and electrical surges/sags caused by the weather or by events such as the switching off of heavy industrial equipment (elevators, factory machines, etc.). A UPS is especially critical to those located in a multi-tenant building where there is competing demand for power and momentary blackouts are common. Through UpSafe, SGI delivers a full line of UPS solutions configured to meet the specific needs of the datacenter environment and Altix server configuration, One solution covers the uninterruptible power system, power monitoring software, and support.

6.4 Electronic Support – Embedded Support Partner

SGI® Electronic Support is a fully integrated suite of services that work together seamlessly to monitor and manage systems and proactively protect against problems.

For example, when SGI Embedded Support Partner (ESP) is activated, if an anomaly is detected, SGI Electronic Support initiates the following actions:

- Notifies the system administrator and an SGI professional
- Searches for possible fixes
- Sends field-proven solutions to the identified system problems

SGI Electronic Support provides a smooth, seamless, multipoint support experience that saves downtime, administrative resources, and money. In many cases, problems may be detected and corrected before users are aware of them. It is available, at no additional cost to system owners who have a valid SGI Warranty, SGI® FullCare™, SGI® FullExpress™, SGI FullExpress 7x24™, or Mission-Critical support contract.

7.0 Conclusion

The Altix product line has seen steady improvement in performance, serviceability, and overall capability since the original product launch of the Altix 3000 series in January 2003. Altix RAS capabilities also demonstrate this trend of constant improvement. SGI manufacturing employs strong feedback and process control methods for continued enhancement of reliability. The RAS infrastructure of Altix including the system controllers, operating system software, and internal firmware undergo steady RAS capability improvements as techniques for monitoring and managing systems are refined and optimized in customer environments. Future development will include continuous advancement in RAS capabilities based on experience and customer feedback. SGI considers the RAS capabilities of our flagship systems a vital and integral component of a reliable and dependable high performance computing solution.

Corporate Office
1140 E. Arques Avenue
Sunnyvale, CA 94085
(650) 960- 1980
www.sgi.com

North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.912.7500
Japan +81 3.5488.1811
Asia Pacific +1 650.933.3000

3994 [03.13.2007]

J15262