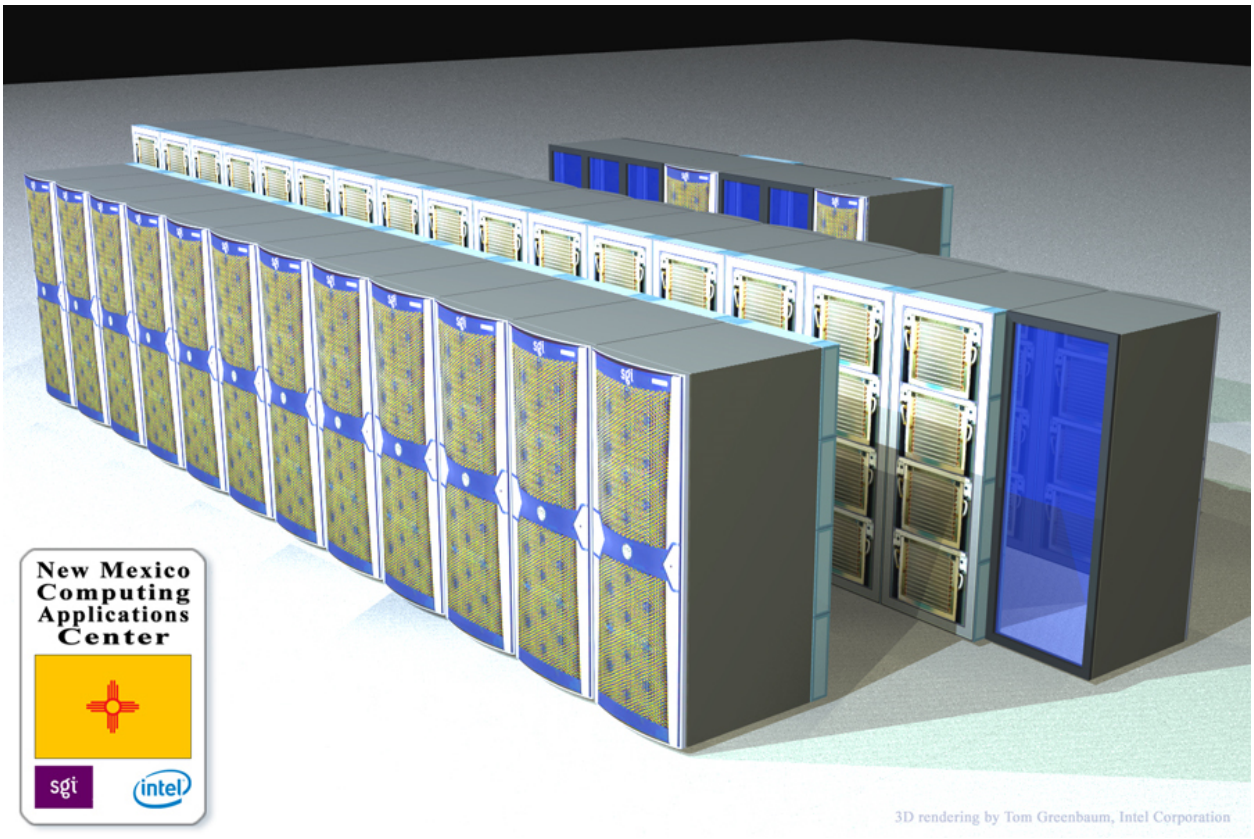
New Mexico Computing Application Center

Cluster Design and Architecture



Nicolas L Behrmann, PMP

NMCAC Project Manager

Department of Information Technology

715 Alta Vista

Santa Fe, NM 87501

Office -505-827-0656

Cell 505-660-3265

TABLE OF CONTENTS

Tables

Figures

Overview

Purpose

Supercomputing is rapidly becoming an essential element of innovation and competitiveness. Computer-based methods now play a central role in all areas of economic development, education, and research. Economically, computational analysis, modeling, and simulation have become a distinguishing factor in business competitiveness. Educationally, developing computing skills in students is an essential part in preparing for a future in a wide range of careers.

Establishing the New Mexico Computing Applications Center will greatly accelerate these trends to the benefit of New Mexico. The Center will provide a nation-wide fabric of scientists and institutions — including leading-edge industrial partners — that are working together (in New Mexico!) on issues important to the State and the Nation.

The NMCAC will provide an ideal environment for researchers from industry, national laboratories, and universities to collaborate both on-site and in "virtual interaction labs" in efforts to accelerate innovation. A major focus will be to bring together computational and simulation tools with researchers and businesses in the State that to establish New Mexico as a global leader in targeted areas of scientific discovery and technological innovation. In fact, one of the requirements of any partnering institution is that their researchers and students must spend time in New Mexico working at the NMCAC. This will help develop the personal relationships that are key to encouraging businesses to permanently site some of their R&D efforts in New Mexico. Once the R&D centers develop new technologies that result in new products, the State will have a basis of attracting the manufacturing of those new products.

In sum, the goals are to:

1) Promote economic development in New Mexico by establishing and maintaining the State as a world leader in innovation
2) Incubate new, innovative businesses and give a competitive advantage to New Mexico business, thus creating high-paying sustainable jobs,
3) Support science and technology education in our State and build a pipeline of highly qualified computational scientists and engineers, with job opportunities in New Mexico
4) Become an example for the way in which disparate organizations with common interests can interact with significant impact.

Funding and Acceptance Project Background

A. Capital Funding from the SENATE FINANCE COMMITTEE, Senate Bill 827 Section 61 - DEPARTMENT OF FINANCE AND ADMINISTRATION PROJECTS-- GENERAL FUND.; 7. fourteen million dollars ($14,000,000) to plan, design, construct, renovate, improve, purchase and equip a state center for advanced computing;
B. State of New Mexico Purchasing RFP 70-361-00-00003 issued for the NMCAC. This RFP covers the acquisition of high performance computing equipment, services, infrastructure, and operating software, as well as for the provision of a facility to house the computing equipment. The request for proposal is expected to result in the procurement and deployment of a high performance computer capable of at least 100 trillion operations per second (100Teraflops) to be housed at a facility in the greater Albuquerque, New Mexico area.
C. State of New Mexico Contract established with Silicon Graphics #08-361-1005
The Contractor shall be responsible for the following:

- Manufacturing, installing, and testing a computing cluster with peak theoretical performance of 172 Tflops.
- Sitting this cluster in a facility that meets the requirements of this Agreement.
- Providing electrical power, cooling, hardware, software, and systems administration support for the cluster.
- Providing connectivity to Lambda Rail or successor agency
- Providing 3 exemplars with each having approximately 2.1 Tflops of peak theoretical performance.
- Providing all associated equipment for the main cluster and exemplars.
- Achieving the successful performance for the cluster, the associated equipment, exemplars, and connectivity to the Lambda Rail, hosting facility, power, cooling, and support.

Contract Deliverables

1. 1A Inventory of Cluster on Site - January 2008
2. 1B Performance Requirements - February through June 2008
3. 2A and 2B Exemplars at UNM, NMSU, NM Tech – March through June 2008
4. Network Connectivity between Rio Rancho Site and UNM Switch at 505 Marquette, ABQ -June 2008
5. System Documentation – June 2008
6. Final Acceptance - June 27th 2008

Initial Collaborative Partners

Overview of NMCAC HPC Systems

Phase 1 - Encanto and Exemplars

The overall collection of NMCAC High-Performance Computing systems includes the large 14,336-core "Encanto" supercomputer installed at the Intel Facility in Rio Rancho and three small "Exemplar" systems situated at three different universities/colleges in New Mexico. The following figure depicts the overall NMCAC network of systems.

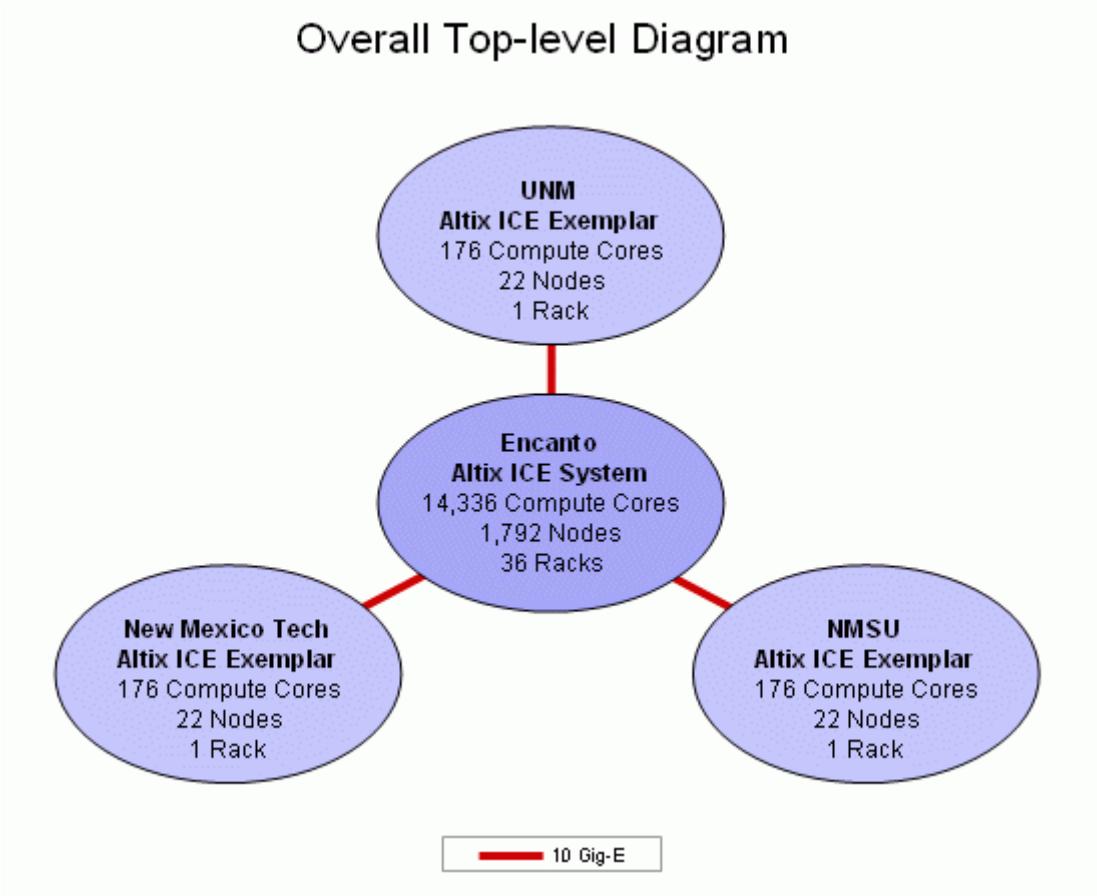


Figure 1. NMCAC Network of Supercomputers

The initial $14 Million dollar appropriation included $11 Million dollars for the Encanto Cluster and smaller compute development and test environments for the UNM, NMSU and NM Tech campuses. University students and researchers would develop and test application codes on the campus exemplars and then run the applications on the Encanto cluster. For more information on the design and architecture of the exemplars see the section at the end of this document.

Phase 2 – Visualization Gateways

Following the testing and acceptance by Department of Information Technology of the Encanto Cluster, an RFP or series of RFPs will be issued for the expansion of the three campus High Performance Computing capability connected to NMCAC and Encanto.

This expansion will consist of Visualization capability and collaboration features that will leverage the computing capacity of Encanto and the local exemplars on campus. For other educational institutions and centers, the visualization gateways will consist of smaller Microcluster compute boxes and collaboration and visualization facilities. Through the Microcluster boxes and telecommunication devices, all these sites could be linked for conferencing and collaborative research.

The following diagram depicts how these gateways connect through the State of New Mexico high-speed network into Encanto and the National Lambda Rail network.

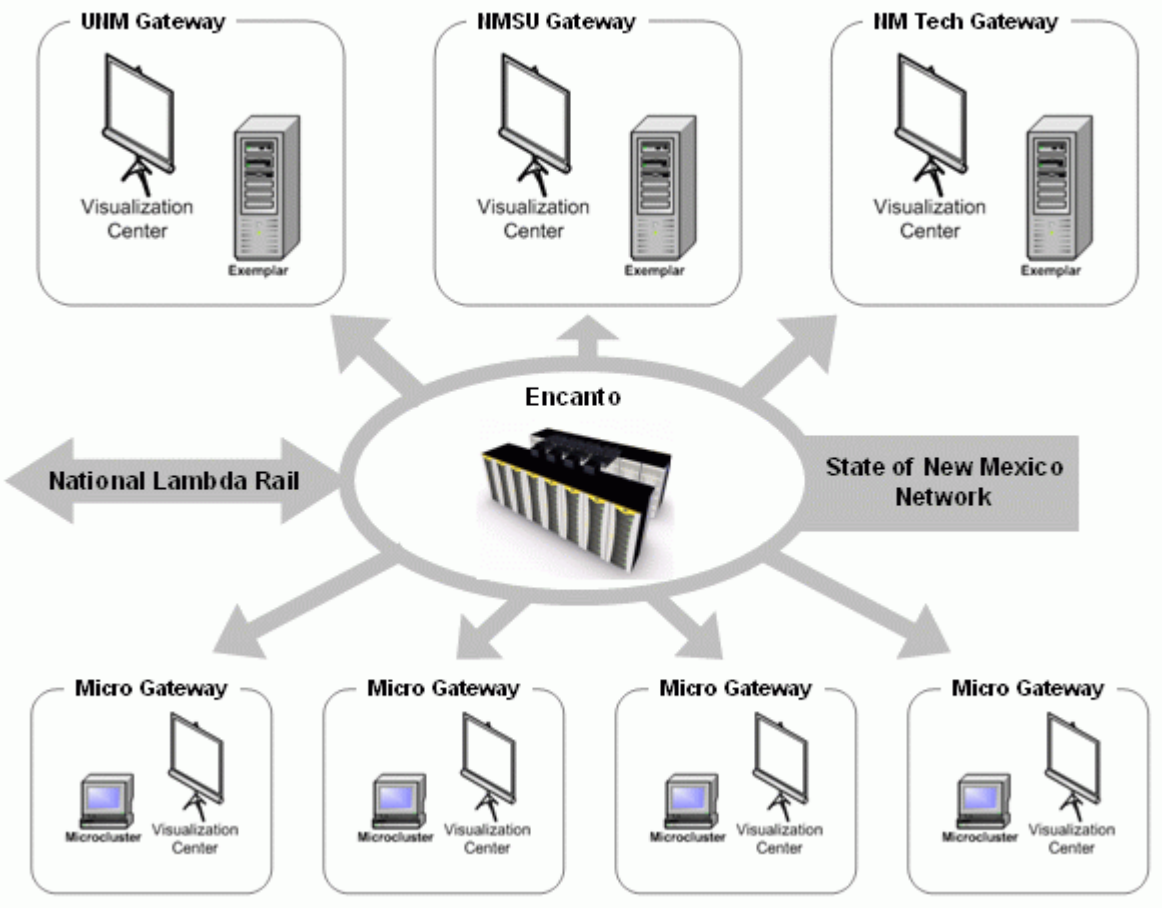


Figure 2. NMCAC Visualization Gateways

The SGI Altix ICE 8200 Cluster - ENCANTO

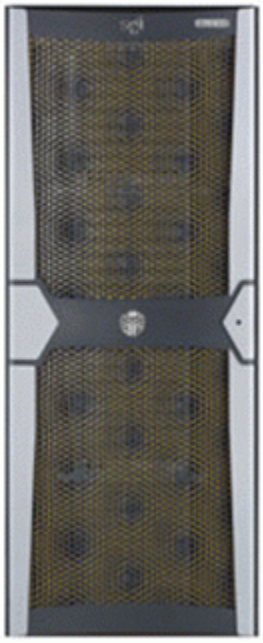
The center of the NMCAC research efforts will be the 1792 Node cluster housed at Intel's Rio Rancho facilities. This cluster has been named "Encanto".

Encanto Overview

Main System "Encanto" Overview

Major Components

- 172 TFLOPS Altix ICE 8200 Cluster
- 1792 Nodes, 3,584 Sockets, 14,336 Cores
- 28.7 TB memory (2GB/core)
- 172 TB Parallel File Subsystem (Lustre)
- 20 OSS Service Nodes
- 20 TB NAS & Archive Storage Subsystem
- 1 System Admin Controller
- 28 Rack Leader Controllers
- 1 Batch Service Node
- 5 Login Service Nodes
- Lambda Rail Connectivity
- 36 Racks – 28 Compute – 5 Storage – 3 Support



The Encanto Cluster system is an SGI Altix® ICE 8200 ultra-dense, Xeon®-based, water-cooled, blade server consisting of 2048 compute nodes (blades) with a total of 14,336 processor compute cores. The peak performance of the system is 172 TFLOPS. Tightly integrated into this high-performance compute cluster is an ultra-high-performance Lustre®-based, parallel storage subsystem of approximately 172 Terabytes (TB) of usable capacity. Additionally, the Encanto system includes a 20-TB NAS subsystem combined with a complete backup/archiving subsystem.

The Encanto Altix ICE 8200 uses the latest 3.0-GHz, x86-64, quad-core Xeon processors from Intel augmented with 32 TBs of ECC, fully-buffered memory. The 32 TB's of main memory equates to three times the amount of information in the Library of Congress. The entire system uses a completely integrated InfiniBand® infrastructure to interconnect the 2048 compute blades into a high-bandwidth, low-latency fabric. The InfiniBand switches are built directly into each of the system's

112 Blade Enclosures. This design greatly reduces the complexity of the InfiniBand hardware connections while allowing for a more sophisticated fabric topology and greatly improved airflow through the compute blades.

The following table highlights the many Altix ICE 8200 value-priced, cost-of-ownership features.

Table 1. SGI Altix ICE 8200 Features

SGI Altix ICE 8200 Features
▪ Integrated blade platform – enhanced serviceability and simplified scalability. ▪ Ultra-high package density – 512 Intel Xeon Processor cores per rack. ▪ InfiniBand Interconnect aggregate bandwidth is 256 GB/second/plane. ▪ Top reliability: – Hot-swappable compute blades – Cable-free system chassis (Blade Enclosures) – Diskless compute blades – N+1 power supply and fan redundancy ▪ Leading energy efficiency – chilled water cooling. ▪ Highly efficient power supplies. ▪ SGI Tempo Platform Manager for flexible management/monitoring at blade, chassis, rack, and system levels.

Table 2. Proposal Deliverables

Proposal Deliverables
▪ 4096-socket, 16384-core, Altix ICE 8200 cluster system – 172 TFLOPS peak performance ▪ 172-TeraByte (TB) Lustre-based parallel storage sub-system (10 GB/second aggregate bandwidth) ▪ 20-TB NAS storage sub-system ▪ Complete Backup/Archive sub-system ▪ Lambda Rail 10Gig-E connection ▪ Physical / cyber security ▪ System software (Novell® SLES10 Linux, SGI ProPack™, PBS Pro, Time Navigator, NAS Manager, Intel Compilers and debuggers, and much more) ▪ "FullCare™" Hardware, software, and systems support through final acceptance

Proposal Deliverables
▪ Complete installation, integration, and testing of entire system ▪ State-of-the-art computing facility

Table 3. Reliability Features

Reliability Features	
Bladed architecture	Fewer cables and connectors
Hot-swap compute blades	Reduced downtime
N+1 redundant power supplies	Minimize system failure due to power reliability
Hot-swappable redundant cooling	Easily replaceable fans
Continuous hardware monitoring	Monitors failed nodes, disks, and switches
Continuous environmental monitoring	Monitors power, temperature, and environment
Redundant RAID storage/controllers	Highly reliable RAID storage with hot spares
Reliable and robust SUSE Linux OS	Mature, well tested Linux operating system

The SGI Altix ICE 8200 ultra high-density blade-based system completely supports compute, service, and I/O functions. The various major components of this system are detailed in the following sections.

NMCAC "Encanto" Top-level Diagram

The functionality of the Encanto cluster includes the 14,336 compute cores or processors which do the computations; the Lustre-based Parallel Storage subsystem that provides temporary high-speed storage for "live" computations so that applications can then access the data for the next computation step; Login service nodes through which the researchers can access the cluster; The NAS storage and archive subsystem which provides both persistent storage for user's home directories as well as backup and archive functionality for user's applications and resultant calculations.

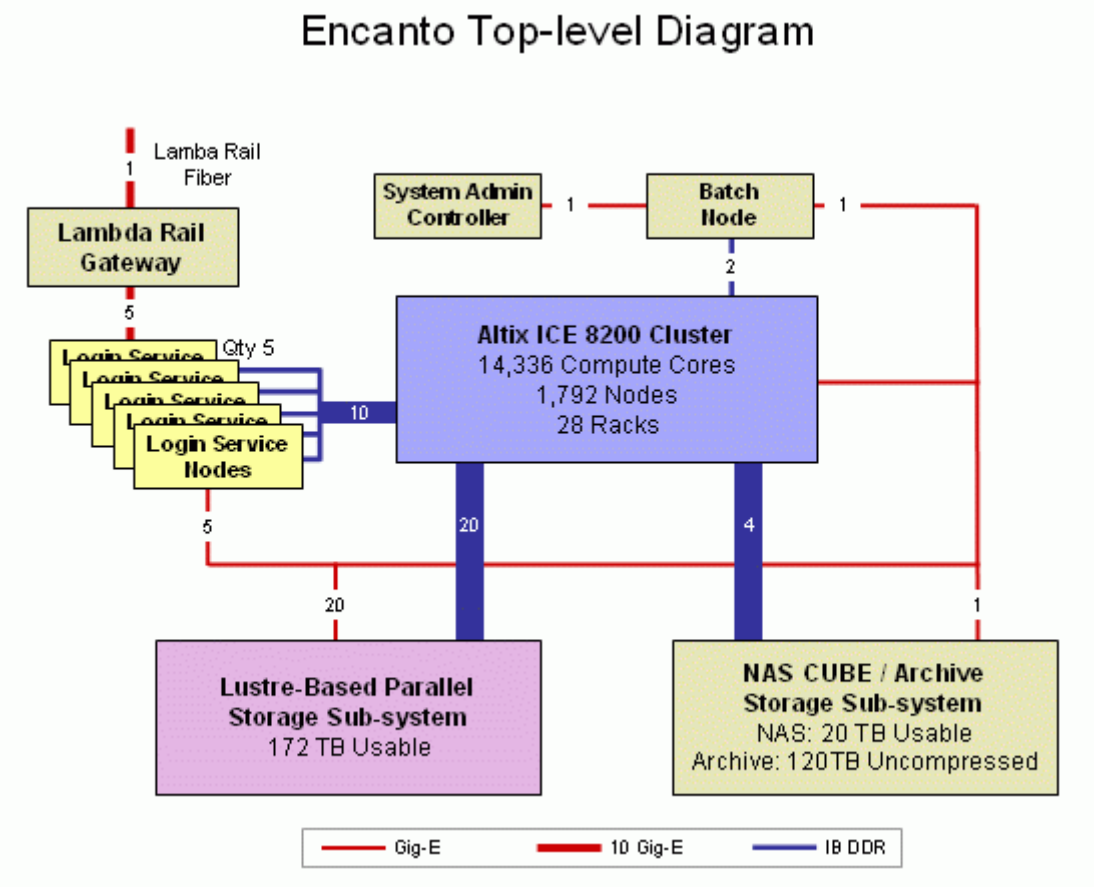


Figure 3. Encanto Top-level Block Diagram

Encanto Floor Layout

The Encanto supercomputer is installed in the old "Fab 7" building of the Intel Facility in Rio Rancho. This system occupies only 800 square feet of floor within this building. The scope of the Encanto Cluster is shown in the floor plan map in the following figure.

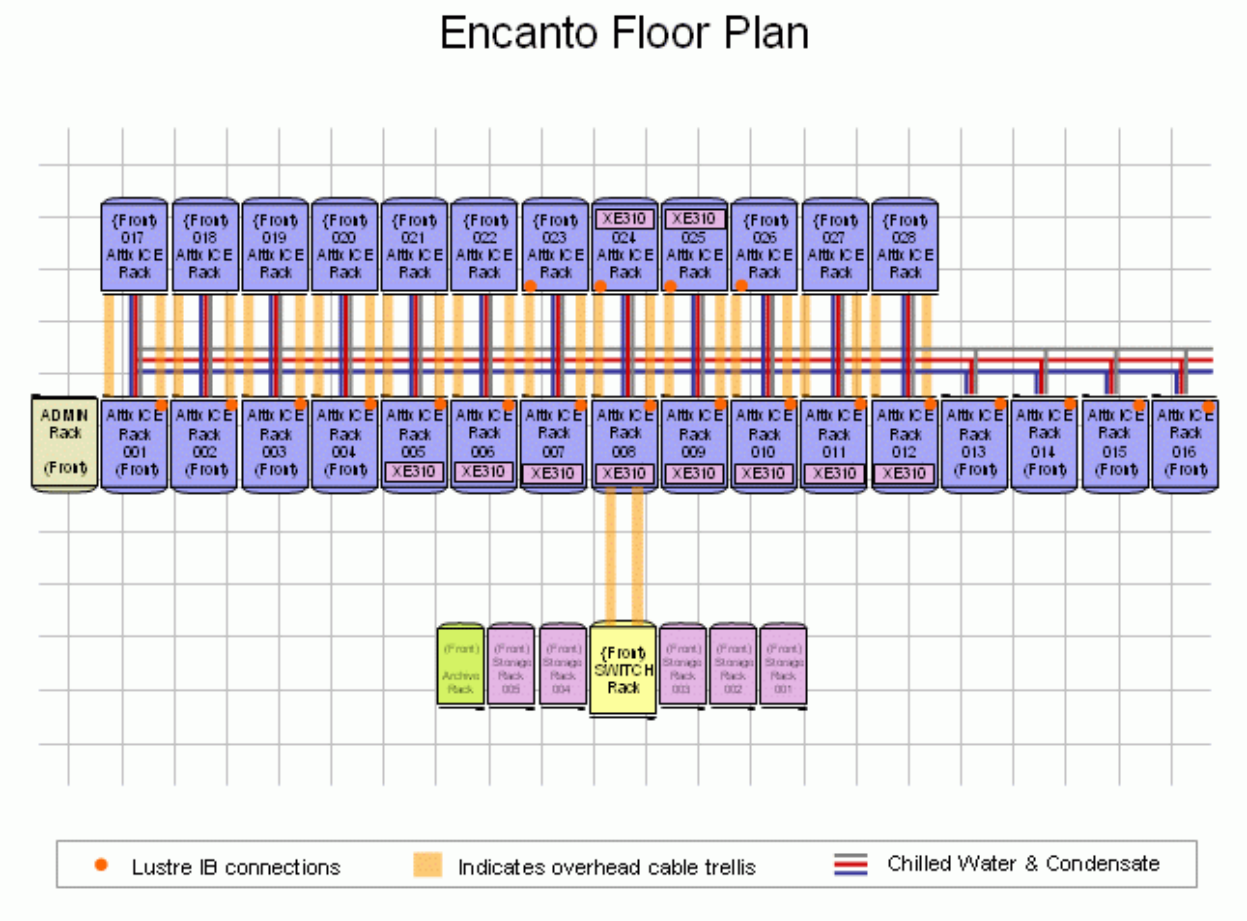


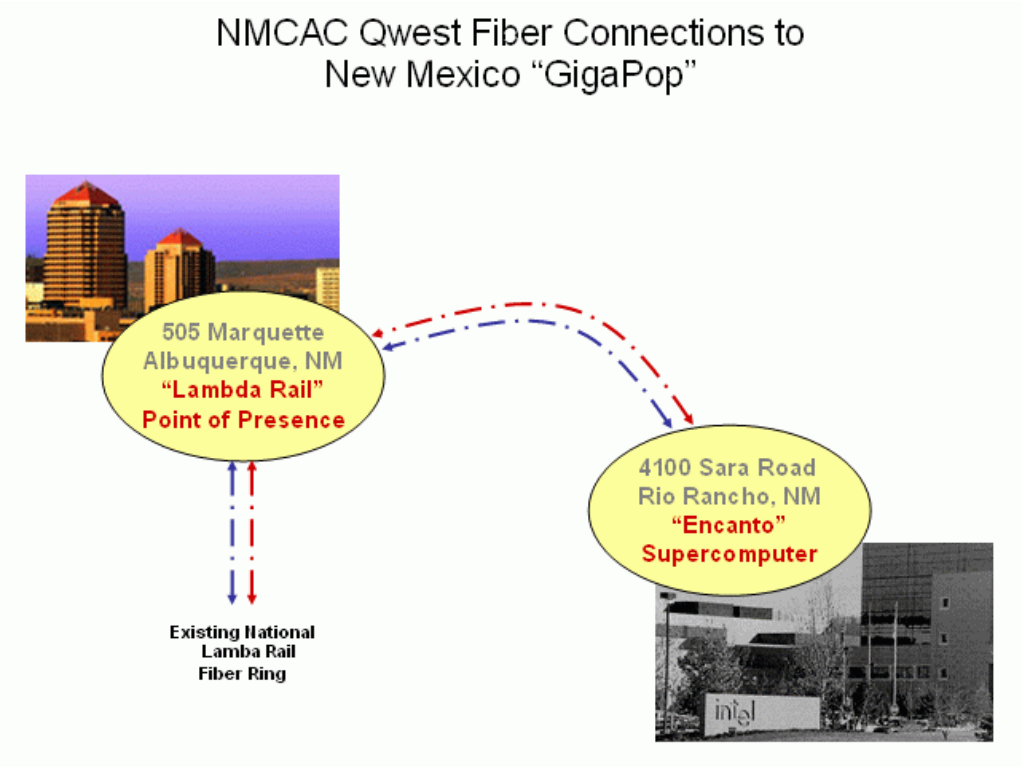
Figure 4. Encanto Floor Plan

This integrated blade platform reduces complexity, simplifies management, and lowers total cost of ownership because of the following design points:

- Up to 70% more compute power per floor tile (based on GFLOPS per sq. ft.).
- Fewer components reduce potential points of failure.
- Leading energy efficiency: average $50K in annual savings per 10 TFLOPS compute power.
- SGI Tempo Platform Manager for flexible management/monitoring at blade, chassis, rack, and system levels.
- Redundant system components and hot-swappable blades.

Encanto Network Connectivity for User Access

The SGI-Department of Information Technology contract for the capital acquisition of Encanto contained the requirement for providing connectivity between the Intel facility and the New Mexico Lambda Rail "Point of Presence" (PoP) in Albuquerque.



Included in this connectivity were Cisco Switches at both the Encanto site and the 505 Marquette Point of presence for NLR. These switches are shown on the following diagram. The UNM switch connection for Encanto will facilitate connection with NLR, Internet2 both of which UNM manages in New Mexico, as well as for the Encanto to UNM Exemplar.

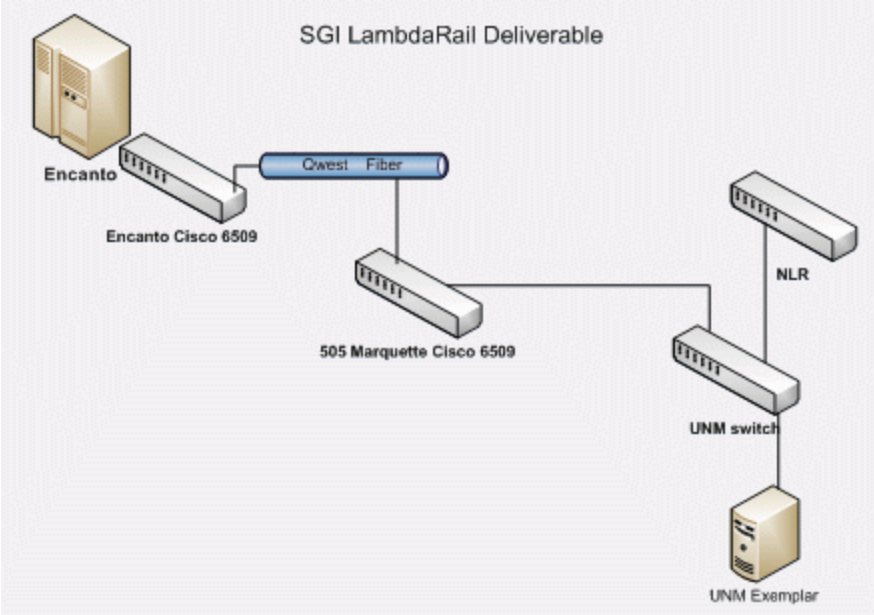


Figure 5. Lambda Rail Connectivity

The following illustration presents a more detailed map of the connectivity between Encanto, NLR, UNM, NMSU and NM Tech, showing the Qwest circuit as well.

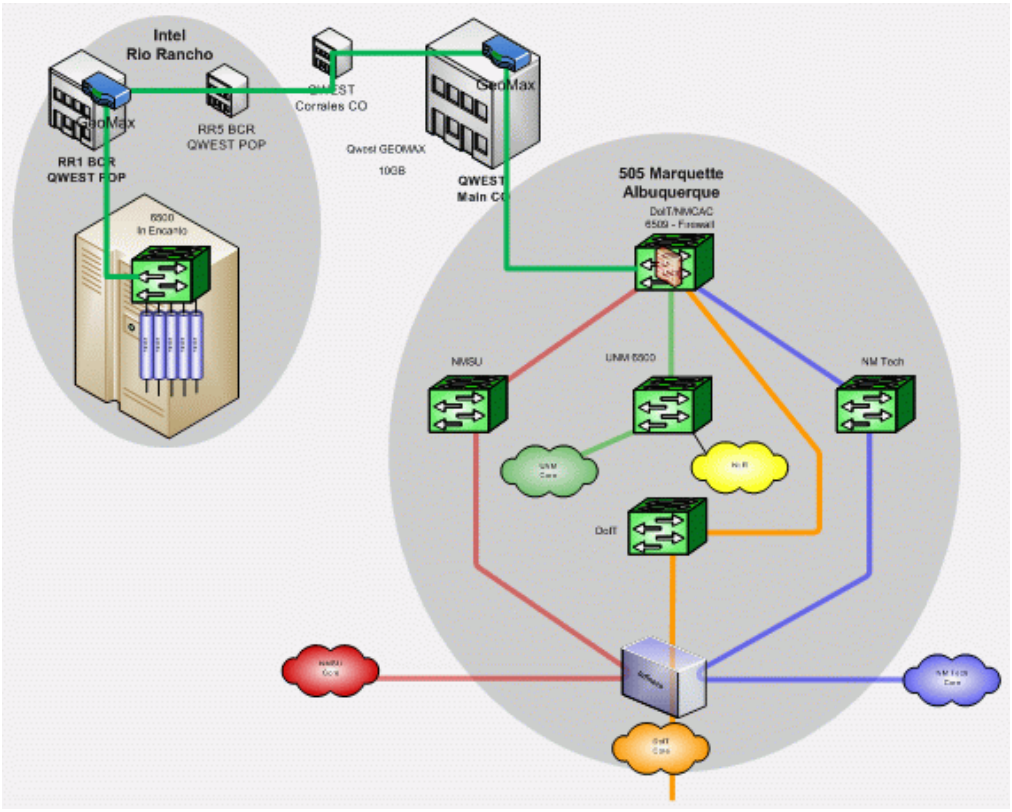


Figure 6. Qwest Network Circuit Detailed Diagram

Network Connectivity at Encanto

The following schematic depicts the connections between the switches and the login nodes at Encanto.

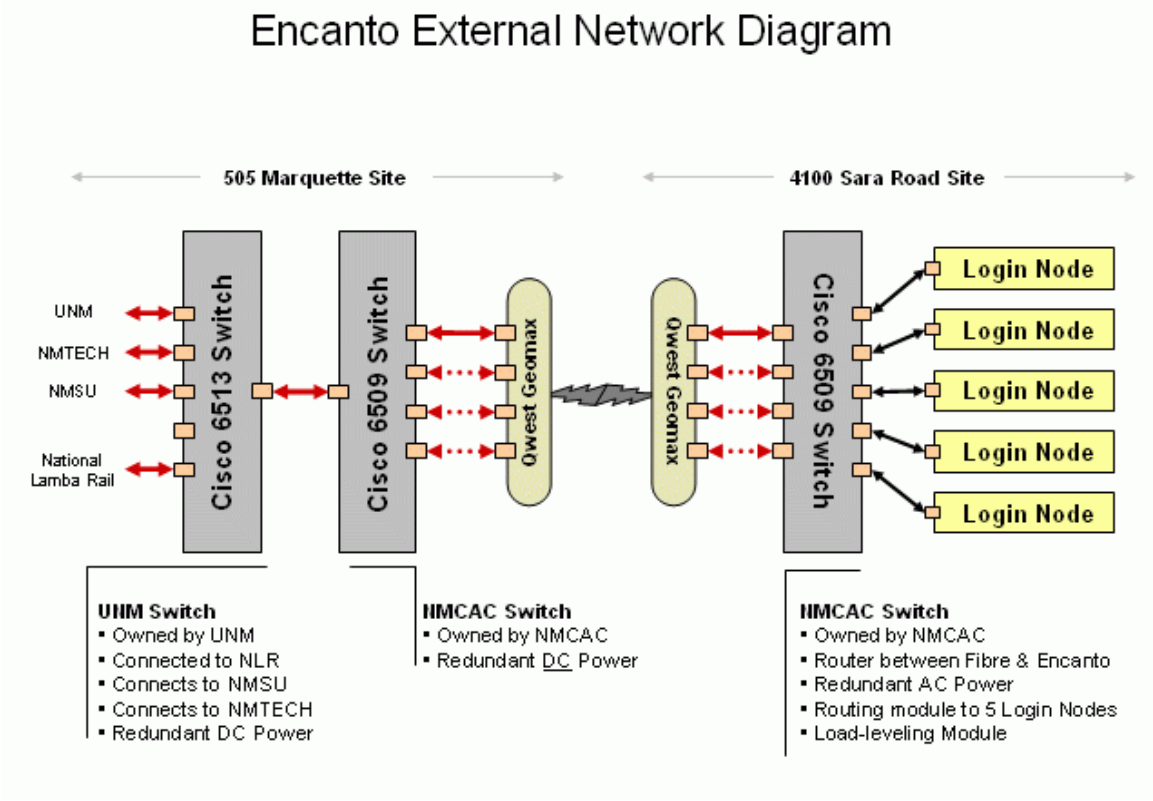


Figure 7. Encanto External Network Diagram

The diagram below shows the connections between the Encanto located Switch router, the login nodes and the compute cluster nodes. What is not shown below is the PBS job scheduler which handles the job requests of the users who have logged onto the cluster.

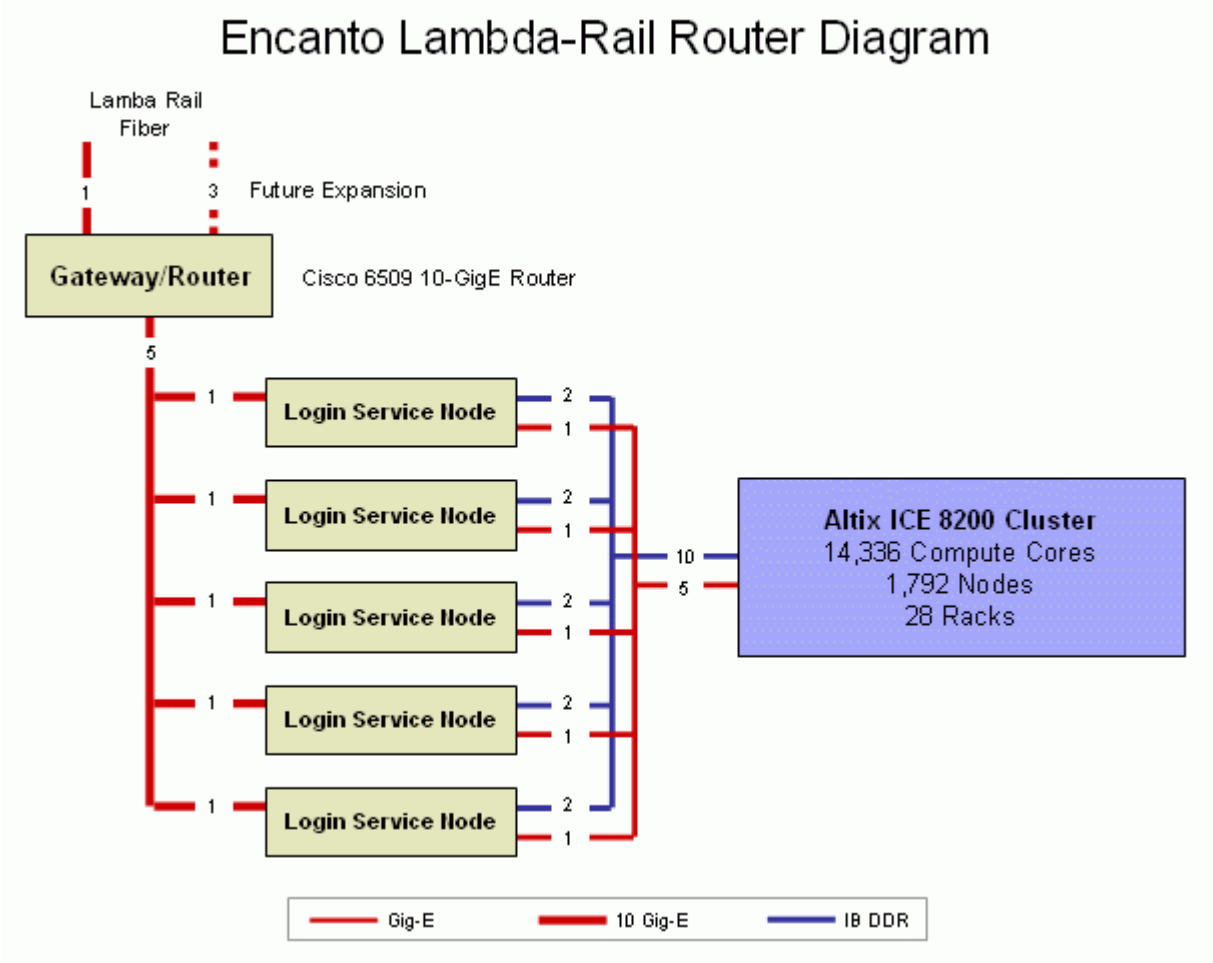


Figure 8. Encanto Login Service Node Router Diagram

Scalable Hypercube Topology

The SGI Altix ICE 8200 system uses a fully, integrated set of 4x DDR InfiniBand switches that are internal to each of its Blade Enclosures. The following figure shows how the compute blades connect to the integrated InfiniBand switches.

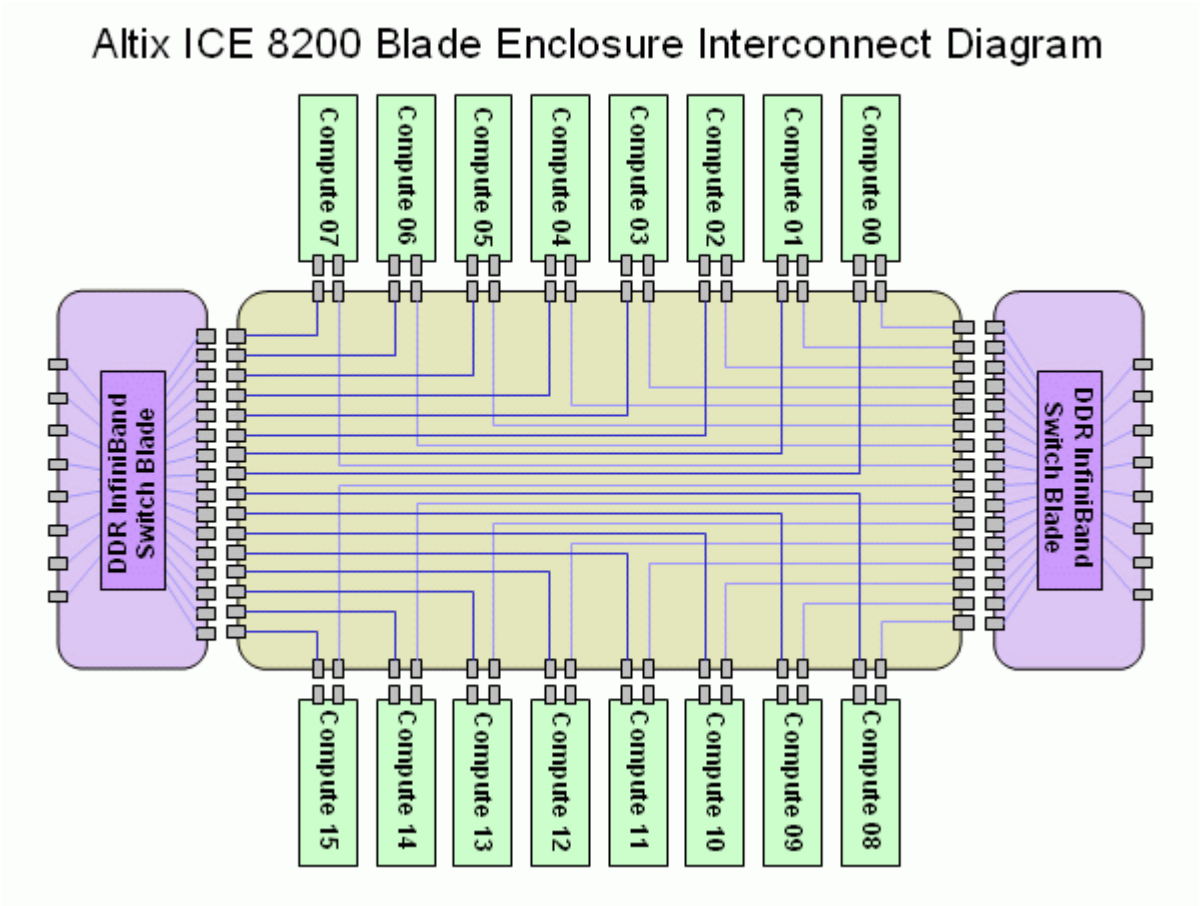


Figure 9. Blade Enclosure Logical Block Diagram

These switches are interconnected to form a dual-plane, bristled 7-D hypercube topology. The hypercube interconnect topology is the basis for all inter-node message passing (using MPI) within the Altix ICE cluster.

The Altix ICE 8200 Cluster architecture links up multitudes of IB switches and compute nodes by emulating a multidimensional, geometrical construct known as a *hypercube*. A hypercube possesses the following properties:

- A hypercube has *n* spatial dimensions where *n* can be any positive integer.
- A hypercube has 2*n* vertices.
- There are *n* connections (lines) that meet at each vertex of a hypercube.
- All connections into a hypercube vertex are mathematically orthogonal to one another.

Conceptually speaking, each IB switch is equivalent to the vertex of a hypercube and each cable or backplane connection is equivalent to a line that connects two hypercube vertices. Thus, the hypercube is the underlying interconnection architecture into which the nodes are connected.

A common method for evaluating the effectiveness of a system's interconnectivity is to look at the system's "bisectional bandwidth". The bisection bandwidth is determined by dividing a system in half (bisecting) and counting the number of "cut" connections that normally would have allowed one half of the system to communicate with the other. The following figure shows several multidimensional hypercubes that are bisected by a plane represented by a red dashed line.

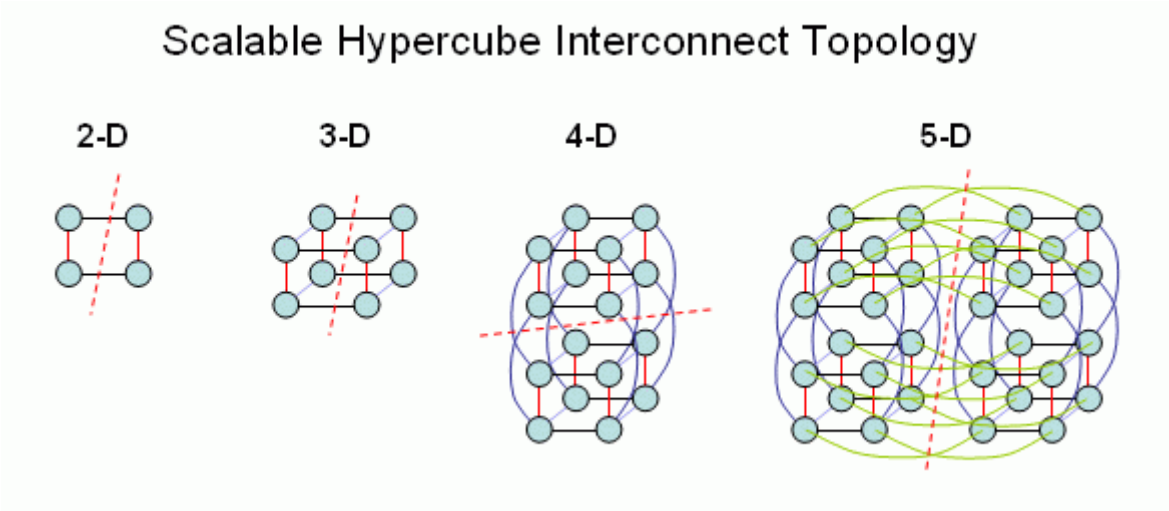


Figure 10. Scalable Hypercube Interconnect Topology

As the number of nodes in an Altix ICE Cluster grows, it is desirable to have the bandwidth grow proportionally to prevent bottlenecks. As the preceding figure demonstrates, the number of connections cut by each bisecting plane scales as the system scales. This architecture linearly scales the bisection bandwidth as the dimensionality of the hypercube increases.

It can be seen that the hypercube architecture enjoys a constant "cuts to node" ratio, regardless of size of the hypercube, maintaining a constant "bisection bandwidth to node" ratio. This clearly demonstrates how the Altix ICE Cluster has the ability to scale to extremely large sizes.

The following figure shows the logical interconnect topology for the Encanto system. Each small light blue circle is one Blade Enclosure.

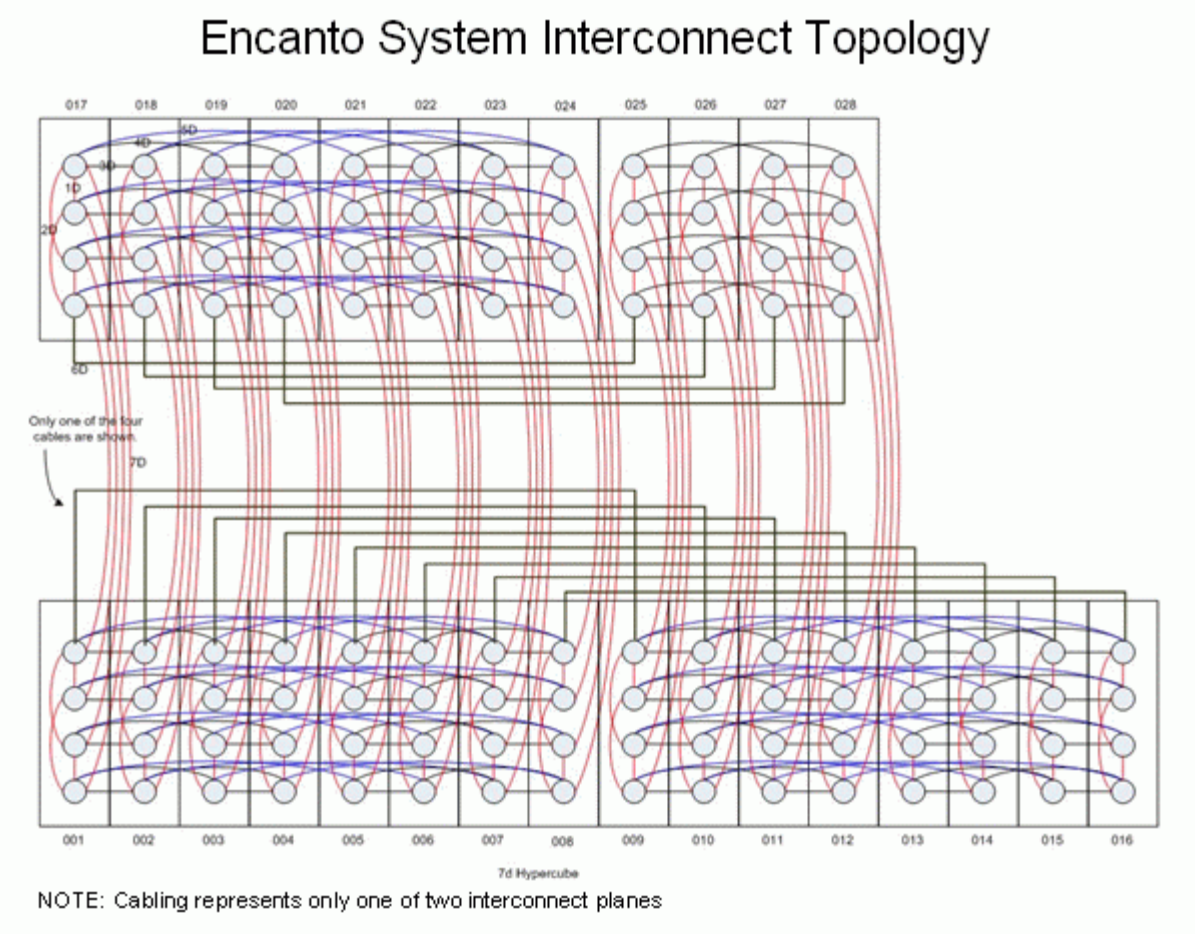


Figure 11. Encanto System Interconnect Topology – 7-D Hypercube

There are actually two InfiniBand hypercube fabrics (or planes) in an Altix ICE 8200. One plane is dedicated for inter-node communication through the use of the Message Passing Interface (MPI). The other IB plane is dedicated for balanced, high-performance I/O. This plane is used primarily to access the Lustre-based, high-performance, parallel storage subsystem.

Physical System Overview

Compute Cluster

The SGI Altix ICE 8200 system is a blade-based, scalable, high-density compute server. The building block of this cluster is the "Blade Enclosure". The Blade Enclosure provides power, cooling, system control, and network fabric for 16 compute blades. Each compute blade supports two quad-core Xeon processor sockets and eight fully-buffered, DDR2 memory DIMMs. Four Blade Enclosures reside in a custom-designed 42U-high rack. One rack supports a maximum of 512 processor cores and 2 TB of memory.

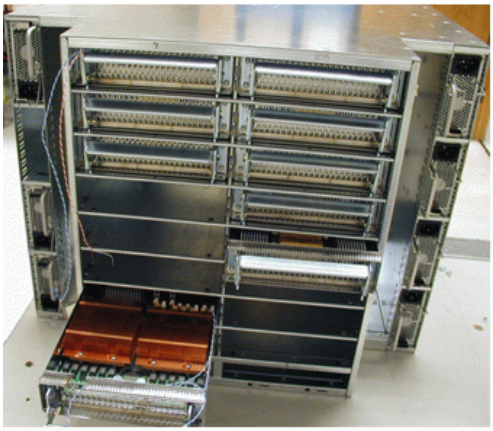
The SGI Altix ICE 8200 system is designed around a "hypercube" topology. This interconnect fabric is implemented through a series of 4x DDR InfiniBand switches.

Blade Enclosure

The Blade Enclosure is a 10U-high chassis that provides power, cooling, system control, and network fabric for up to 16 compute blades. Up to four Blade Enclosures can be installed into a single 24" ICE rack.

Compute Blade

The compute blade has two processor sockets and eight memory DIMM slots. Up to 16 compute blades can be installed in a Blade Enclosure.



InfiniBand Switch Blade

The switch blade contains one InfiniBand ASIC. The switch blade provides the interface between compute blades within the same chassis and also between compute blades in separate Blade Enclosures. The switch blade has one DDR 4x InfiniBand ASIC. Two switch blades are used with each Blade Enclosure to support a dual-plane topology for enhanced bandwidth.

Chassis Management Control Blade

Each Blade Enclosure has one chassis manager, which is located directly below compute blade slot 0. The chassis manager supports powering up and down of the compute blades and environmental monitoring of all units within the Blade Enclosure.

Compute Rack

The custom-designed 42U-high rack holds up to four Blade Enclosures and one Rack Leader Controller. The rack is designed to support both air and water-assisted cooling. The Encanto system is water-cooled. The following depicts the physical layout of the custom compute rack.

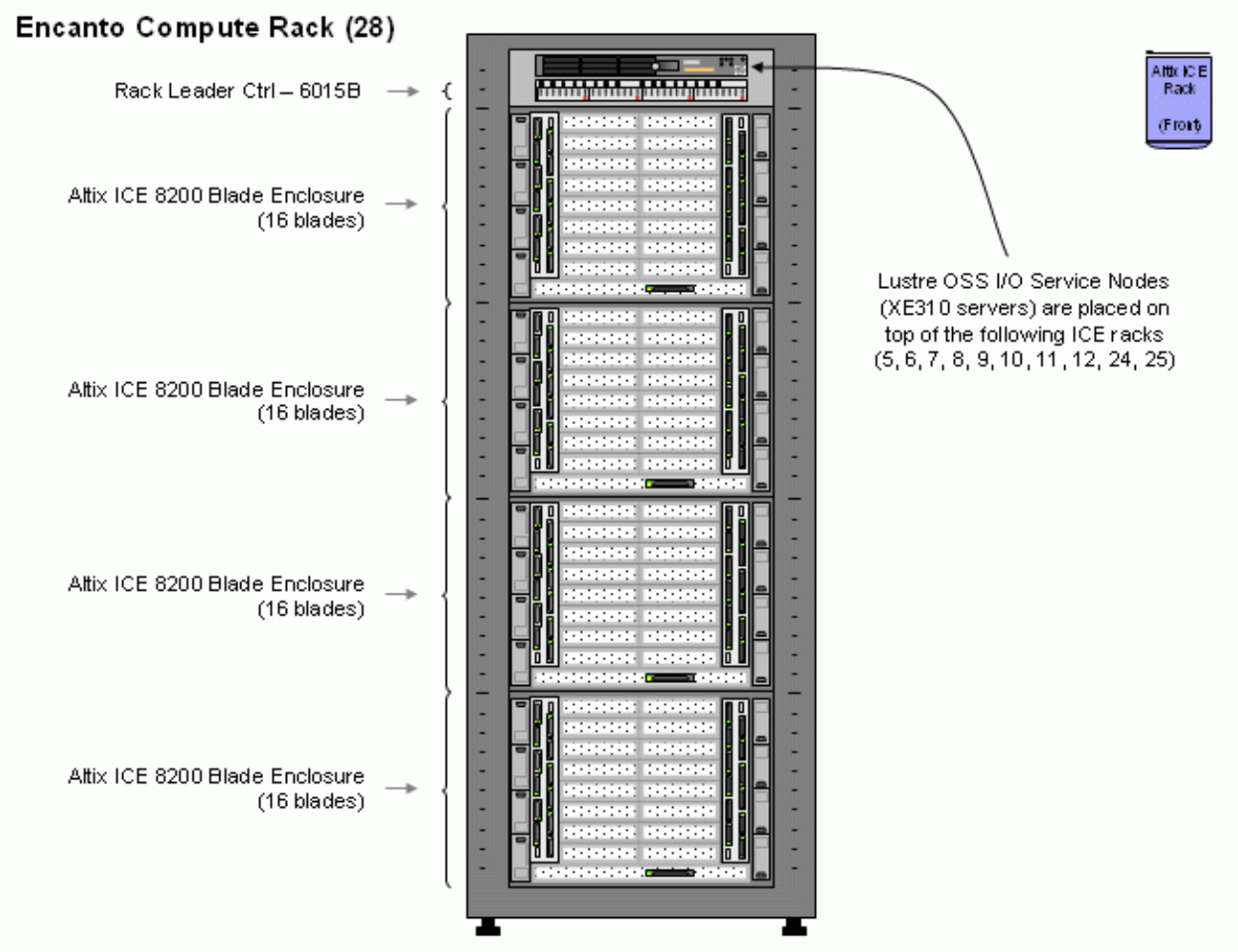


Figure 12. Encanto Compute Rack Physical Layout

System Admin Controller

The Encanto Altix ICE 8200 includes one System Admin Controller. The System Admin Controller is an SGI Xeon-based server. Only one System Admin Controller is needed per system and is used to manage the entire cluster. The System Admin Controller is configured with one 2.66-GHz dual-core Xeon processors, 8-GB memory, and four system HDDs (250-GB SATA drives configured as hardware-enabled RAID1). This node is also configured with a set of "n+1" redundant power supplies.

Rack Leader Controller

The Encanto Altix ICE 8200 includes one Rack Leader Controller per ICE rack. The Rack Leader Controller is an SGI Xeon-based server. These servers are used for system startup and as a "boot server" for the diskless compute nodes. In this configuration it also runs the fabric management software to monitor and control the IB fabric. There is one Rack Leader Controller for each rack. The Rack Leader Controller is configured with two 2.33-GHz dual-core Xeon processors, 4-GB memory, and two system

HDDs (250-GB SATA drives configured as hardware-enabled RAID1). The node is also configured with a set of "n+1" redundant power supplies.

Login Node

The Encanto Altix ICE 8200 includes five Login nodes. The Login node is an SGI Xeon-based server. The Login node provides access to the user's home file system, and allows that user to develop, debug, and run one or more parallel jobs. There can be one or more Login nodes per system and they may be combined with the Batch node or Gateway service nodes in many configurations. Each Login node is configured with two 2.66-GHz dual-core Xeon processors, 8-GB memory, and four system HDDs (250-GB SATA drives configured as hardware-enabled RAID1). This node is also configured with a set of "n+1" redundant power supplies. Additional Login nodes can be added as demand requires.

Batch Node

The Encanto Altix ICE 8200 includes one Batch node. The Batch node is an SGI Xeon-based server. There can be one or more Batch nodes per system and the Batch node can also be combined with a Login node or a Gateway node. For the Encanto system, a single Batch node is configured with two 2.66-GHz dual-core Xeon processors, 8-GB memory, and four system HDDs (250-GB SATA drives configured as hardware-enabled RAID1). This node is also configured with a set of "n+1" redundant power supplies. Additional Batch nodes can be added as the total number of user logins increase. The following depicts the physical layout of the IO and service node rack.

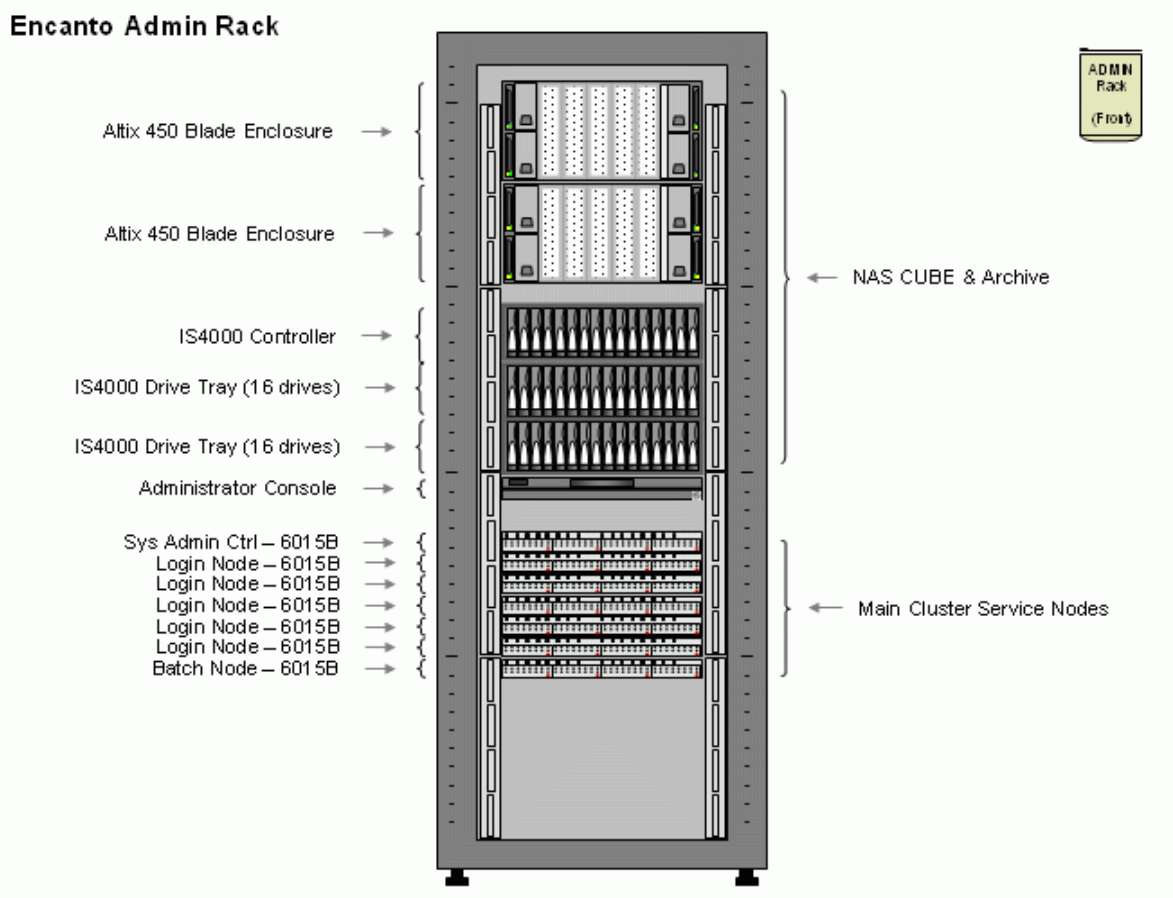


Figure 13. Encanto Admin Rack Physical Layout

Lustre Storage Node

The Encanto Altix ICE 8200 includes twenty Lustre Storage nodes. Each InfiniBand-connected Lustre storage node consists of two Xeon-based OSS (Object Storage Server) servers. An OST (Object Storage Target) RAID storage sub-system is attached to the two OSSs through multiple Fibre Channel 4 (FC4) connections. Data objects are dynamically distributed horizontally across multiple Lustre Storage nodes to provide an object-oriented, high-performance, parallel file system.

Atoka-P Compute Blades

Each “Atoka-P” compute blade contains the processors, memory, and two imbedded 4X DDR InfiniBand HCAs. The Atoka-P node board is a collaborative design by SGI, Intel, and SuperMicro. The Atoka-P node board is the basic building block for the SGI Altix ICE 8200 compute blade and was designed with a high-performance InfiniBand interconnect from its very inception. This compute blade contains no hard drives for either system or data. The blade is designed to boot “disklessly” from each Rack Leader Controller.

The Encanto SGI Altix ICE 8200 system uses Atoka-P node blades configured with two Intel Quad-Core, 64-bit x86, Xeon, 5300 series processors, with a 1333-MHz system bus speed and a core frequency of 3.0 GHz. Each blade is configured with 16 GB of Fully Buffered DDR2 DIMMs providing 2 GB memory per core. All memory is fully qualified ECC memory, burned in at the factory and from the same manufacturer.

The SGI Altix ICE 8200 system uses the Atoka-P node board as its basic building block for the compute blades. Sixteen Atoka-P blades are housed in a single 10-U Blade Enclosure. See the following figure for the Atoka-P block diagram.

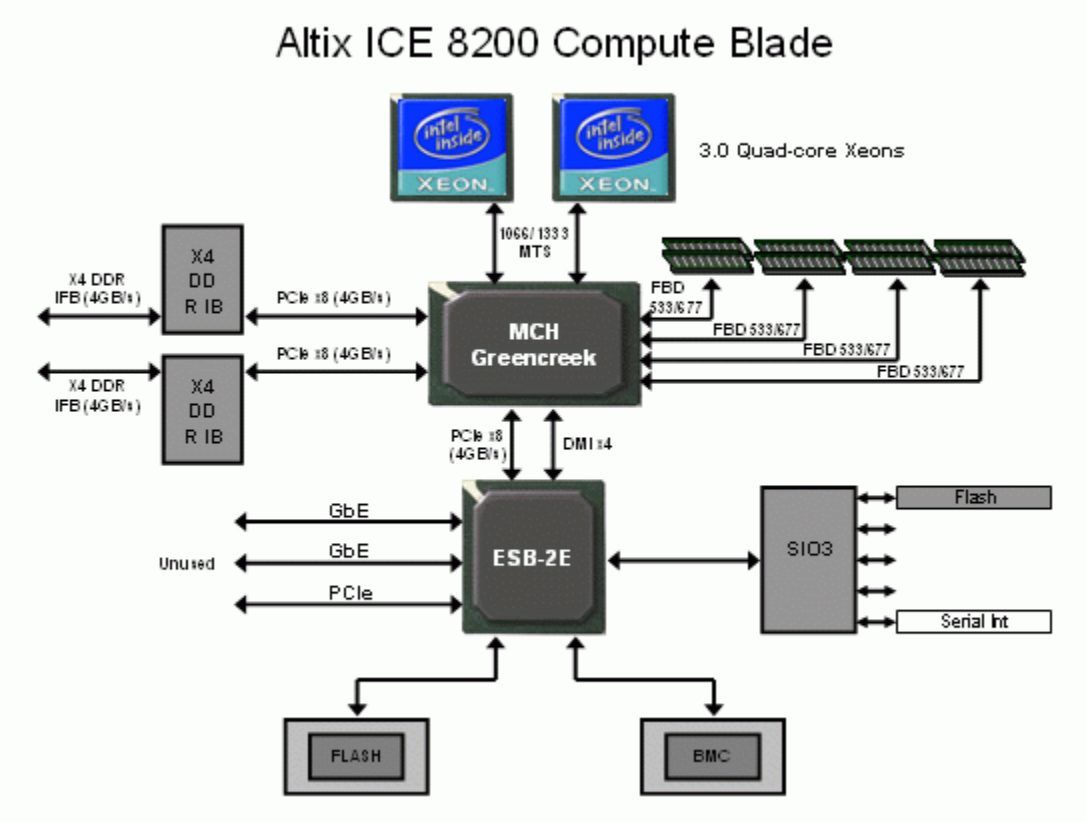


Figure 14. Altix ICE 8200 Compute Blade Block Diagram

Processor Support

The Encanto system uses Atoka-P node boards configured with two Quad-Core Intel Xeon processor 5300 series processors, with 1333-MHz system bus speed running at a core frequency of 3.0 GHz. The node board can be configured with one or two Xeon processors.

Memory Controller Hub

The Memory Controller HUB (MCH) is located on the Atoka-P node board. It is a single 1432-pin FCBGA package that includes the following core platform functions:

- System bus interface for the processor subsystem
- Memory controller
- PCI Express® ports including the Enterprise South Bridge Interface (ESI)
- FBD thermal management
- SMBUS interface

System Bus Interface

The MCH is configured for symmetric multi-processing across two independent front-side-bus interfaces that connect the quad-core Intel Xeon processors 5300 series. Each front side bus on the MCH uses a 64-bit wide 1066- or 1333-MHz data bus. The 1333-MHz data bus is capable of addressing up to 64 GB of memory. The MCH is the priority agent for both front-side-bus interfaces, and is optimized for one socket on each bus.

Memory Subsystem

The MCH provides four channels of Fully Buffered DIMM (FB-DIMM) memory. Each channel can support up to two Dual Ranked Fully Buffered DDR2 DIMMs. FB-DIMM memory channels are organized into two branches for support of RAID 1 (mirroring). The MCH can support up to eight DIMMs or a maximum memory size of 32-GB physical memory in non-mirrored mode and 16-GB physical memory in a mirrored configuration.

The read bandwidth for each FB-DIMM channel is 4.25 GB/second for DDR2 533 FB-DIMM memory, which gives a total read bandwidth of 17 GB/second for four FB-DIMM channels. This provides 8.5 GB/second of write memory bandwidth for four FB-DIMM channels.

For DDR2 667 FB-DIMM memory, the read bandwidth is 5.3 GB/second, which gives a total read bandwidth of 21 GB/second for four FB-DIMM channels. This provides 10.7 GB/second of write memory bandwidth for four FB-DIMM channels.

The total bandwidth is based on read bandwidth; thus the total bandwidth is 17 GB/second for 533 and 21.0 GB/second for 667.

The NMCAC SGI Altix ICE 8200 system is configured with 2 GB of memory per core.

ESB-2 IO Controller

The ESB-2 is a multi-function device that provides the following four distinct functions:

- I/O controller
- PCI-X® bridge
- Gb Ethernet controller
- Baseboard Management Controller (BMC)

Lustre-based Parallel File System

A Lustre-based parallel file system is used for the Encanto Altix ICE 8200 configuration. The Lustre parallel file system redefines I/O performance and scalability standards for the world's largest and most complex computing environments. Ideally suited for data-intensive applications requiring the highest possible I/O performance, Lustre is an object-based cluster file system that scales to tens of thousands of nodes and Petabytes of storage with ground-breaking I/O and metadata throughput. The following diagram outlines the Encanto Lustre parallel storage subsystem.

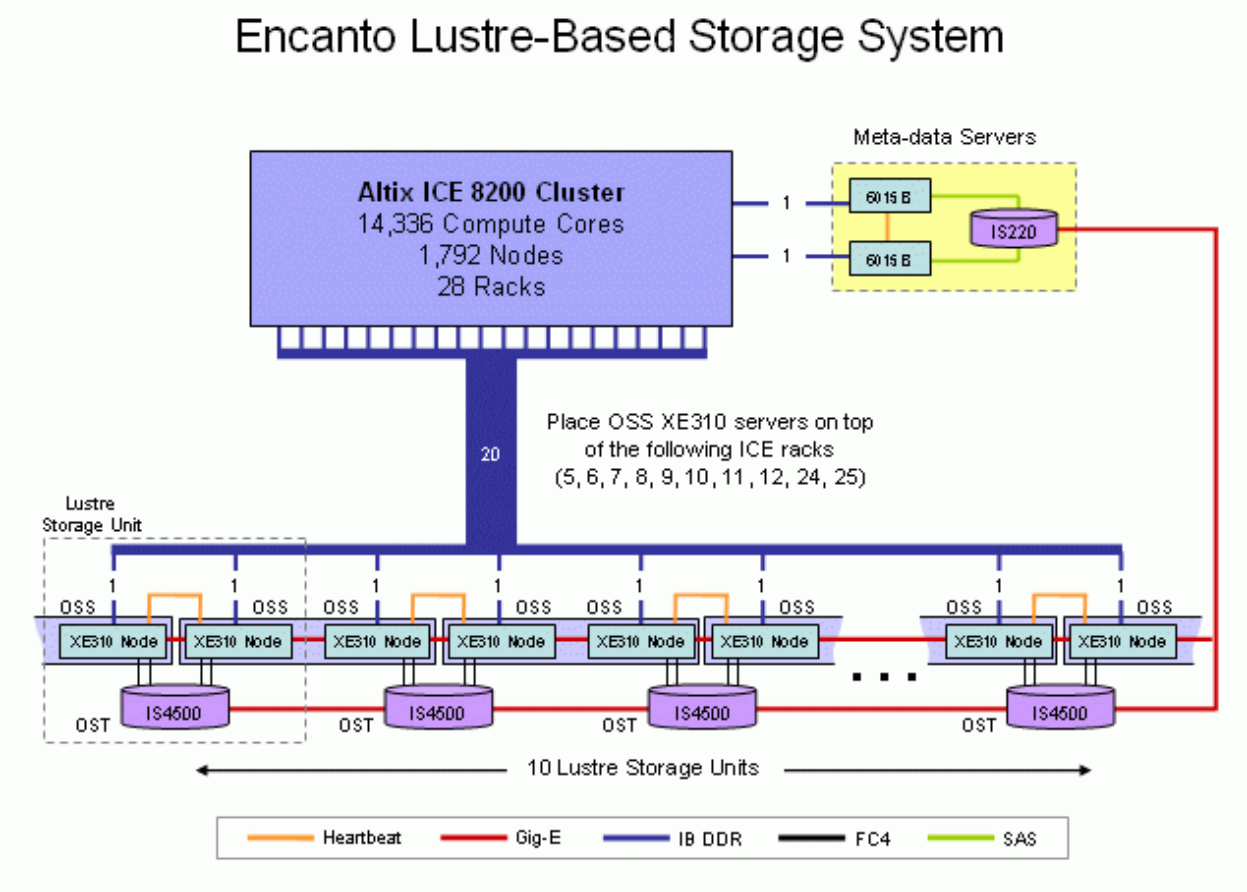


Figure 15. 172-TB Lustre-based Parallel Storage Subsystem Diagram

Lustre is a highly scalable distributed file system that combines open standards, the Linux operating system, an open networking API, and innovative protocols. Together, these elements create a very large data storage and retrieval system. Building a Lustre clustered file system requires a Lustre Meta-Data Server (MDS) and Lustre Object Storage Servers (OSSs), each with an Object Storage Target (OST). Data objects are dynamically distributed horizontally across the servers and a pool of client systems accesses these servers through one of many supported networks. The following diagram shows cabling details for each Lustre Storage Unit.

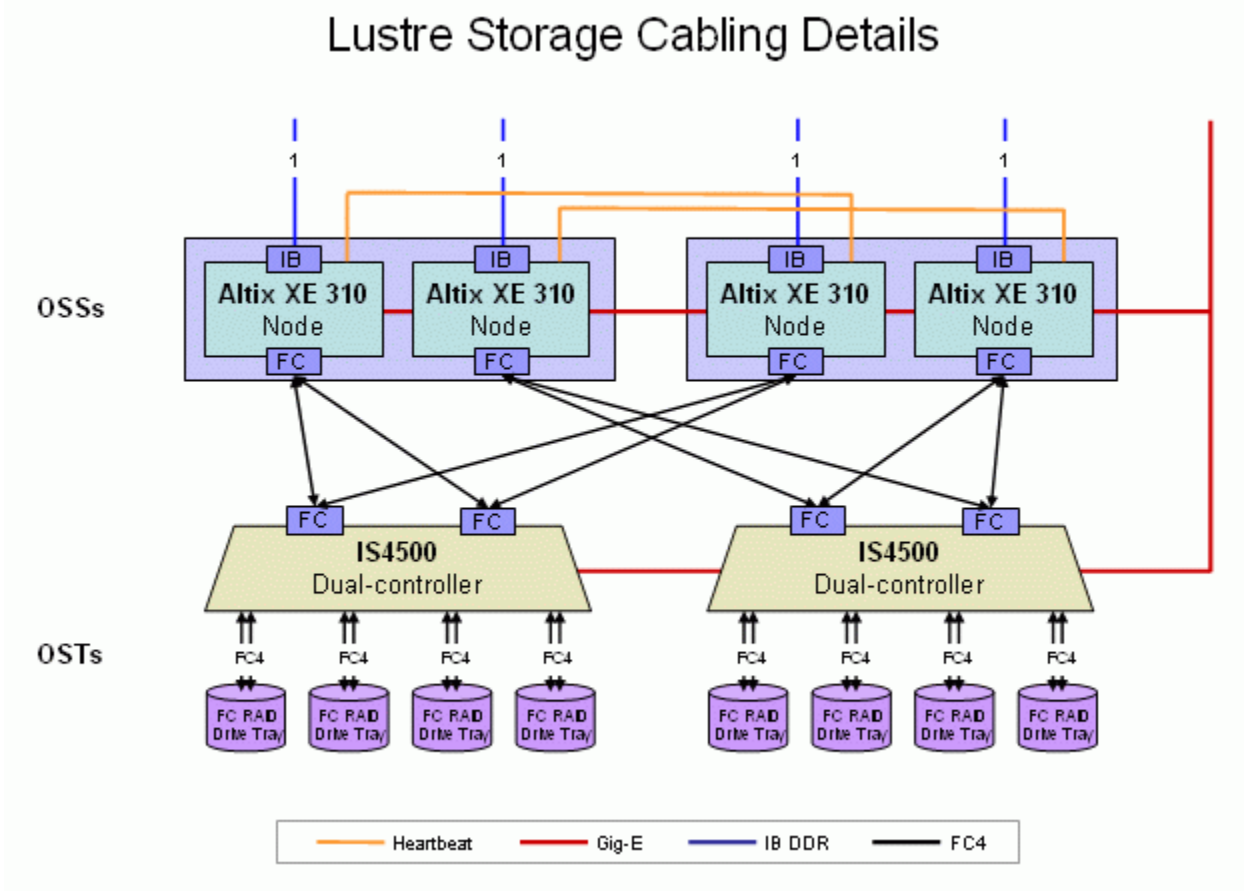


Figure 16. Lustre Storage Unit Cabling Details

Lustre file operations bypass the Meta-Data Server completely and fully utilize the parallel InfiniBand 4x DDR data paths to all OSS's within the storage subsystem. This unique approach – separating metadata operations from data operations – results in significantly enhanced performance.

Lustre has been developed and maintained as open-source software under the GNU General Public License (GPL), enabling broad support for industry-standard platforms.

Lustre network architecture provides flexible support for a wide variety of networks and high-performance features. Lustre interoperates with network vendor-supplied libraries through Lustre Network Drivers (LND), which utilize advanced features such as Remote Direct Memory Access (RMDA), OS-bypass for parallel I/O, and vector I/O for efficient bulk data movement.

Lustre revolutionizes configuration simplicity. Routine formatting and mounting of server devices aggregates them into a global high availability cluster file system.

Lustre employs a distributed lock manager to handle access to files and directories and to synchronize updates, improving on the metadata journaling approach used by modern file systems. To dramatically reduce bottlenecks and to increase overall data throughput, Lustre uses an intent-based locking mechanism, where file and directory lock requests also provide information about the reason the lock is being requested. For example, if a directory lock is being requested to create a new, unique file, Lustre handles this as a single request. In other file systems, this action requires multiple network requests for lookup, creation, opening, and locking.

The Lustre lock manager automatically adapts its policies to minimize overhead for the current application. A single lock, eliminating additional lock overhead, covers files being used by a single node. Nodes sharing files get the largest possible locks, which still allow all nodes to write at full speed.

The following figure depicts the physical layout of the Lustre Storage Rack. Five of these racks are used to house the entire Lustre Storage Subsystem.

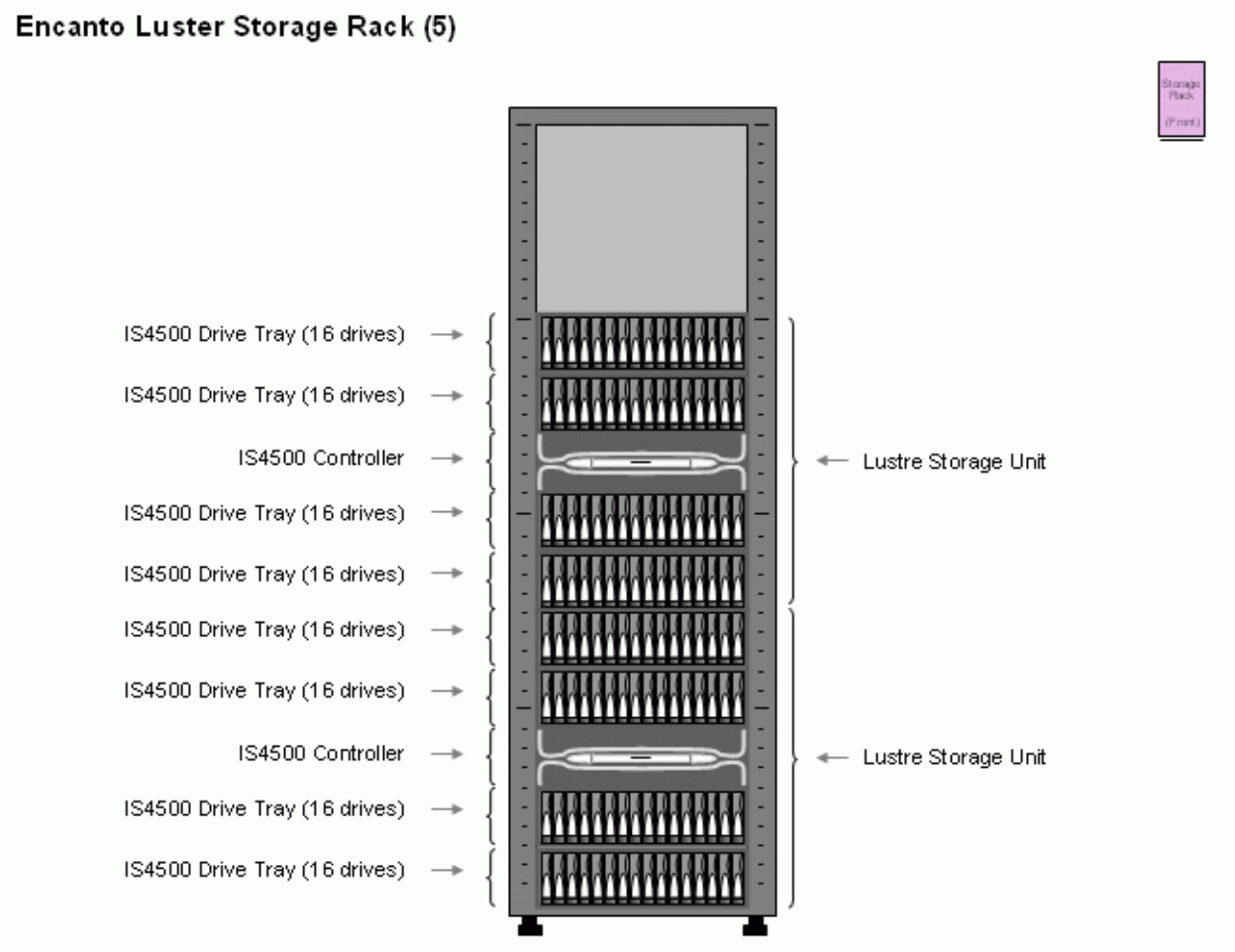


Figure 17. Encanto Lustre Storage Rack Physical Layout

The following diagrams show details of the host-side and back-end Fibre Channel connections.

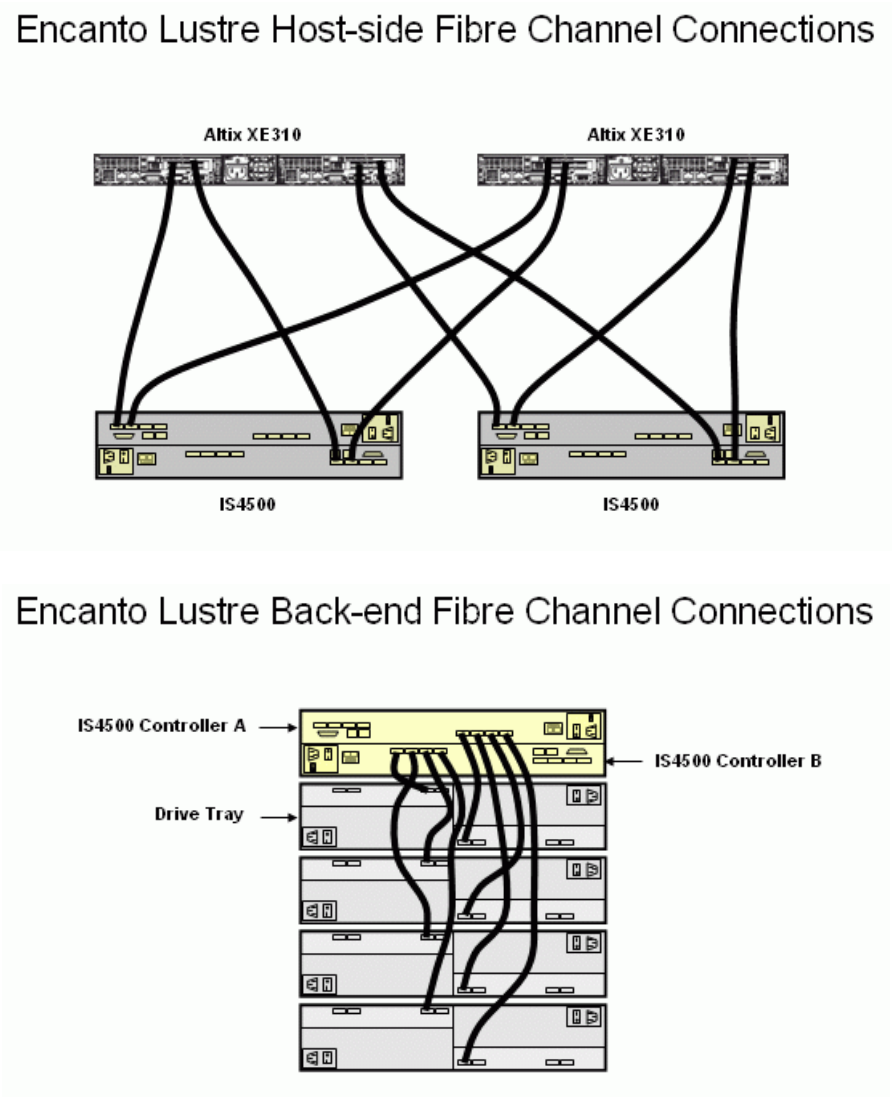


Figure 18. Lustre Storage Fibre Channel Connections

NAS Storage Subsystem

The Encanto SGI Altix ICE 8200 system includes an SGI InfiniteStorage "NEXIS" NAS storage subsystem. The usable capacity of this NAS storage is configured to be 20 TB of usable, RAID-based, scalable storage. This storage subsystem is accessible via a common, global name space by all nodes in the Encanto Cluster system through four InfiniBand 4x DDR connections. On the drive side, this SGI NAS storage is provisioned as 400 GB SATA disks, distributed across three drive trays. This NAS storage array is configured as RAID5 (parity protected) with additional hot spares.

The SGI InfiniteStorage NEXIS NAS family benefits from the industry-leading SGI file system technology. Architected to scale up to exabytes of addressable space, this exceptional, journaled SGI file system can be applied to a wide range of hardware platforms and configurations.

The following diagram shows the NAS storage subsystem along with the Archival tape library that is discussed in the next section.

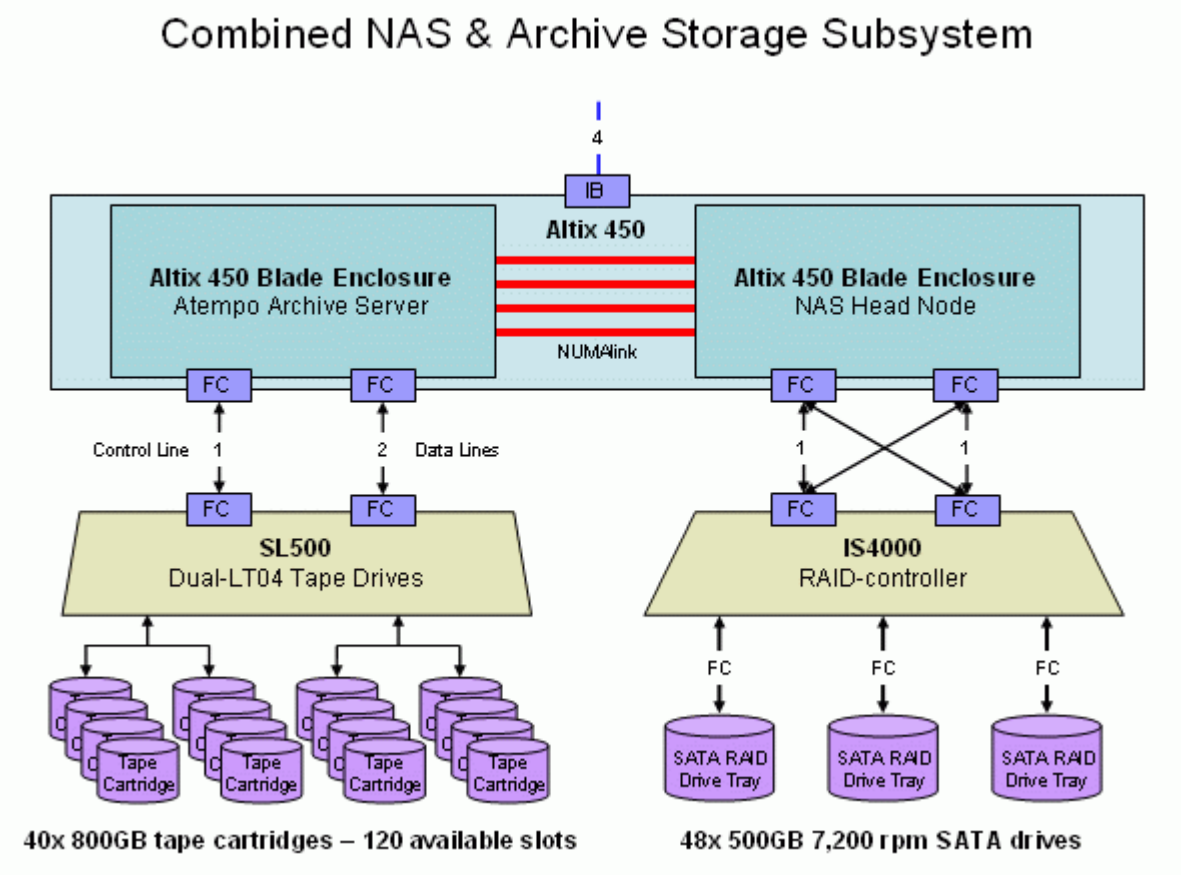


Figure 19. Combined NAS & Archive Storage Subsystem

Archive and Backup

The Encanto system provides a complete backup, archive, and restore subsystem. This system will provide up to 120 TB of uncompressed archive-level storage. The hardware component of this sub-system is an SL500 modular tape library system from StorageTek™. The tape cartridge library currently contains two LT04 tape drives configured with 150 cartridge slots. The library is populated with (40) 800GB tape cartridges for a total of 32 TB of uncompressed data. The software component utilizes Atempo's Time Navigator (TINA) archival solution. Time Navigator combines the best of data protection and storage security to enable corporations to meet critical business requirements for digital privacy, regulatory compliance, and non-repudiated long-term archiving. The product's graphical user interface makes it easy to manage snapshots, replication, disk and tape backups, and archival and recovery operations in addition to simplifying the management of multiple servers and remote sites.

ATEMPO TIMEnavigator™

Atempo's award winning Time Navigator is the fastest restore data protection solution on the market. Built upon a single, unified architecture Time Navigator delivers restore times equal to backup times plus a rich feature set that includes many of today's hottest features. Time Navigator is targeted to the open, heterogeneous environment and is fully flexible where end-users need it. With its unique concept of "time navigation", Time Navigator is the only software product on the market that allows end-users single file access at any point in time. A unique and easy to use directory tree makes locating restored files a simple and visual process. In addition, Time Navigator addresses firewall security by requiring only a single port to be opened for backup through the firewall. Time Navigator also offers in-demand features such as multi-streaming and synthetic full backup, giving end-users the highest degree of control over time and resources in the backup and restore process.

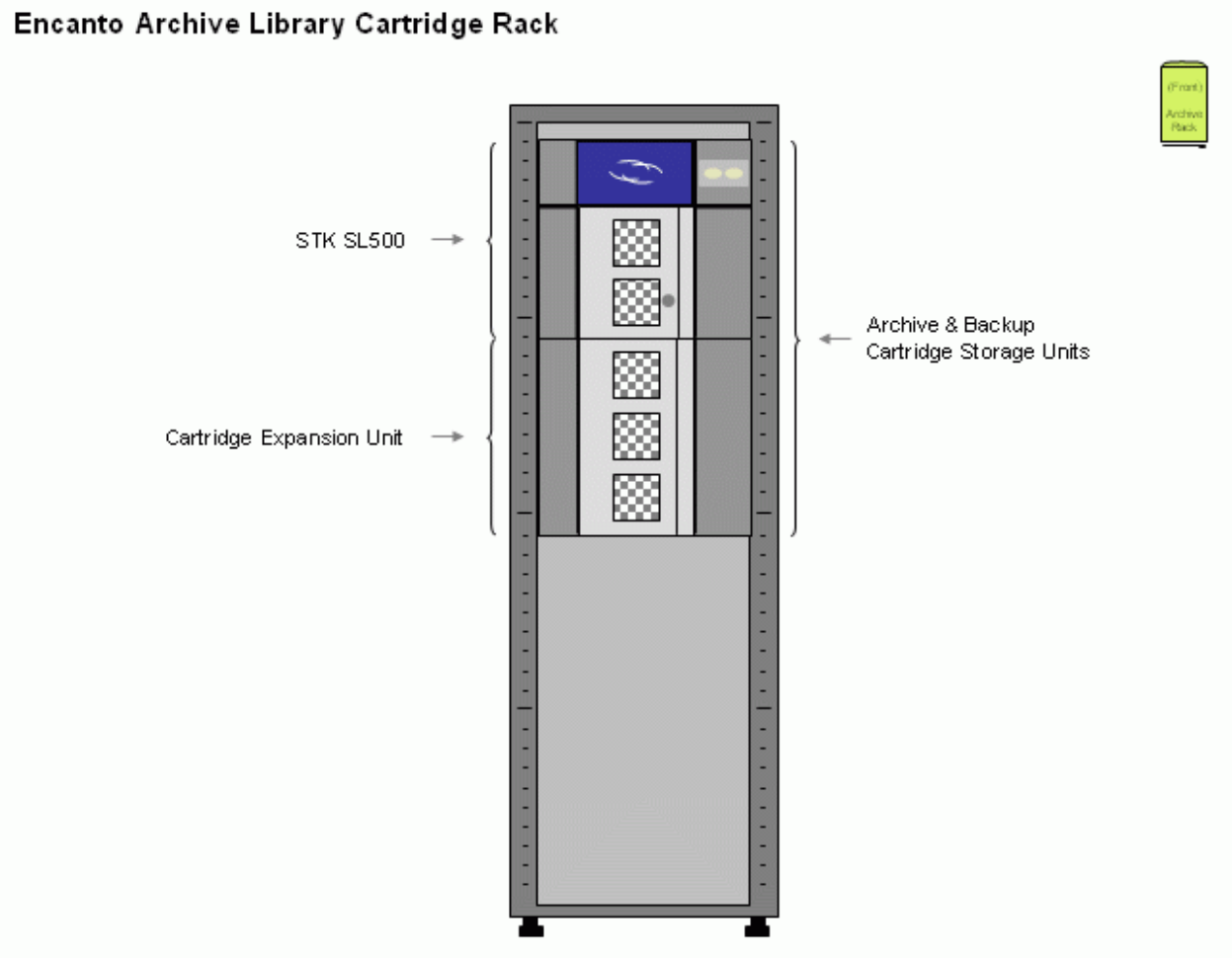


Figure 20. Encanto Archive Rack Physical Layout

The SL500 modular library system is one of the most scalable, rack-mounted tape-automation solutions in the industry. Slot Expansion Modules that hold 120 cartridges can be easily added within the same storage rack. Additional tape read heads can be configured to improve overall backup and restore performance.

Cooling Technology

The SGI Altix ICE 8200 systems are designed to be air-cooled and maintain adequate thermal control within the hardware specifications.

This chilled-water design is based on the third-generation water-cooling technology originally created for the Altix 3000/4000 line of Itanium® systems. This solution provides exceptional cooling efficiency, with a 95% reduction of the facility air-cooling requirement. The following diagram depicts how the Altix ICE 8200 distributes the chiller technology to each compute rack.

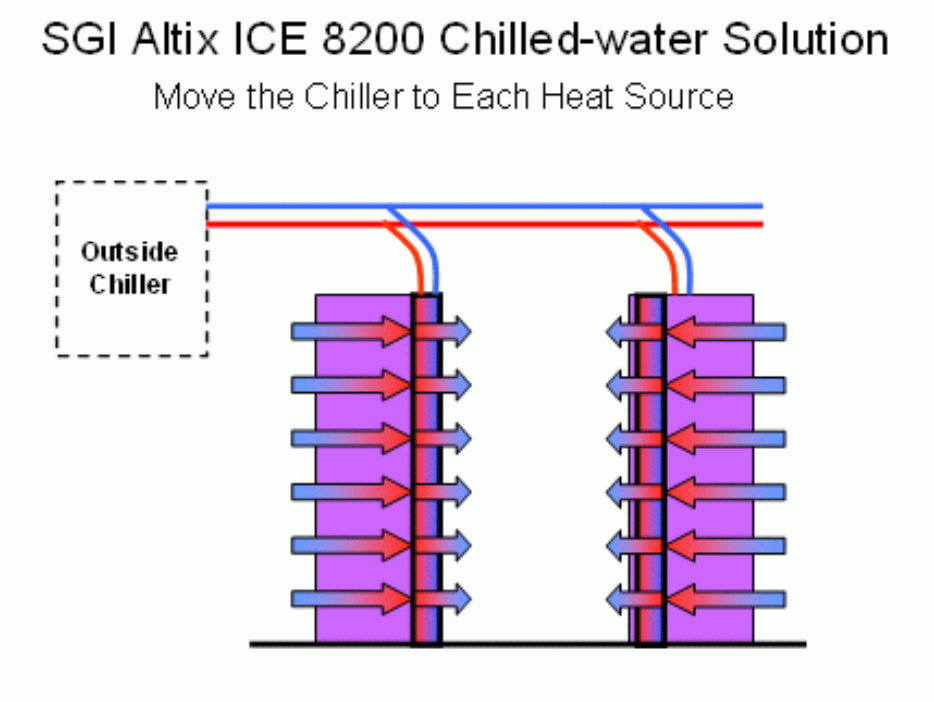


Figure 21. Encanto Chilled-water Solution

The following photo shows a single rack configured with four individual water-chilled doors.

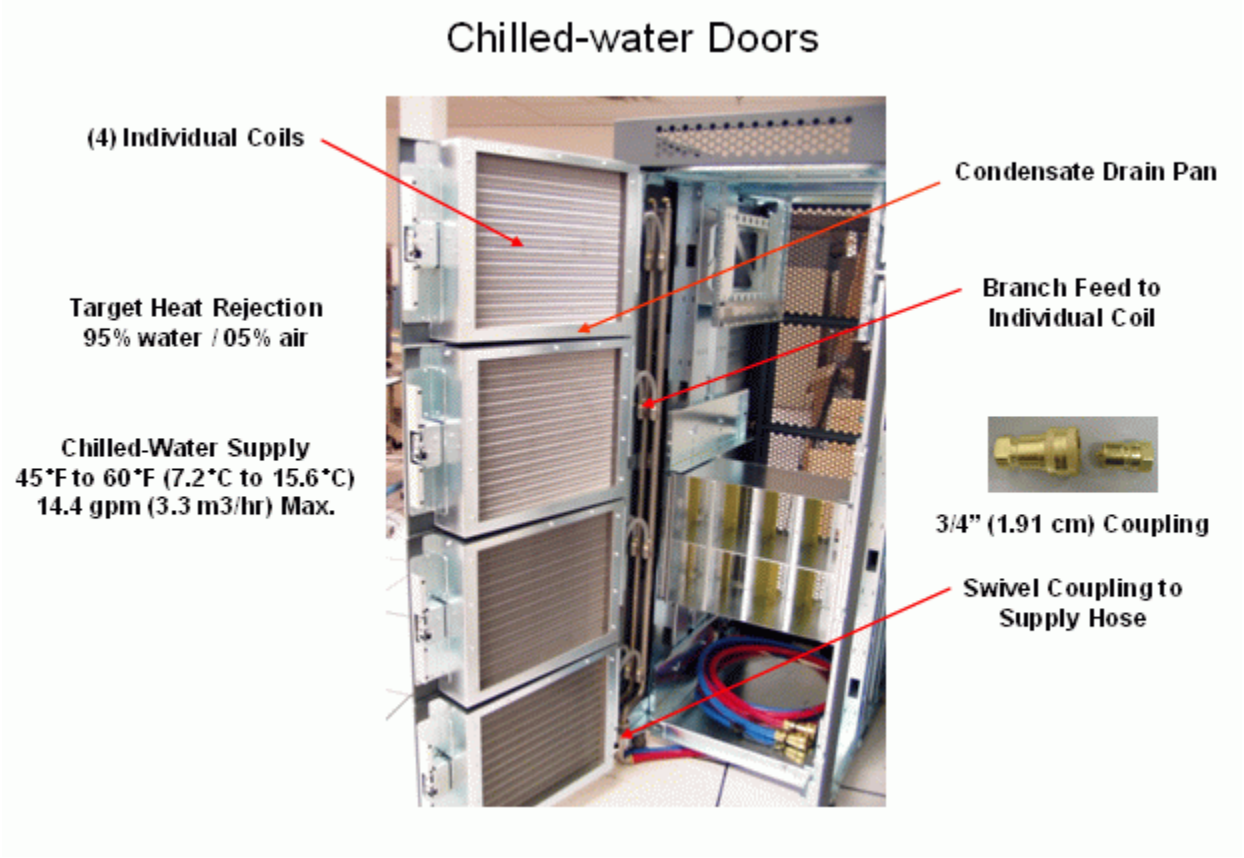


Figure 22. Encanto Single Rack with Four Water-chilled Doors

Software and Operating Environment

SUSE Linux Enterprise Server

The Encanto Cluster system uses the Linux operating system. The specific distribution that is supported is the SUSE™ Linux Enterprise Server 10 (SLES10). SLES10 is a highly scalable foundation for secure HPC computing. Backed by Novell®, it offers comprehensive functionality to SGI's high-performance supercomputers.

The system also includes SGI ProPack™, which offers numerous optional performance and management enhancements, is included as an overlay to the SLES10 operating system. ProPack provides features such as: PCP (Performance Co-Pilot™) for use in debugging multi-processor jobs, CPUSET, memory placement tools, and much more.

Compiler Support

The Encanto Cluster includes the complete collection of GNU compilers for C, C++, and FORTRAN. Part of the popular GNU Compiler Collection (GCC), these compilers are designed to optimize code for platforms based on Intel processors. A GNU compiler with FORTRAN extensions to support FORTRAN 95 is also available from the Free Software Foundation.

The SGI Cluster can be optionally licensed to run the Intel, ANSI-compliant compilers (Fortran 90, C, and C++ languages) on all nodes of the system. SGI has worked closely with Intel on the specifications of its compilers and tools to develop and support a suite of compilers and libraries that provide optimal performance for applications on the SGI Altix ICE 8200 system. The compilers are designed and tuned to take advantage of SGI Altix system architecture and are optimized for high-performance computing.

Intel licenses can be purchased directly from SGI, with single-user, floating, small cluster licenses, with one year of support included. Up to four additional years of support are also available for those customers wishing to secure support for multiple years.

Intel FORTRAN Compiler

The Intel FORTRAN Compiler for Linux makes it easy to get outstanding performance from all 64-bit Intel processors. The compiler provides optimization technology, threaded application support, features to take advantage of hyper-threading technology, and compatibility with leading tools and standards to produce optimal application performance. The compiler delivers leading-edge performance and excellent compatibility, and comes with first-class customer support.

FORTRAN compatibility features:

- Ability to handle big endian files: This is a mechanism to read and write data files in big endian mode. In the technical computing world, data files are generated on various platforms, including systems that use the big endian model of computing. This option allows users to use files that have been generated on different platforms with ease.
- Compatible with widely used Linux software utilities: The Intel FORTRAN Compiler is substantially compatible with common tools used in developing Linux applications such as make, Emacs, and gdb.

For a detailed description of the complete optimization features and techniques available with the Intel compilers, please refer to the following URL:

http://www.intel.com/software/products/compilers/

FORTRAN standards support:

- ISO FORTRAN 95 support: The Intel FORTRAN Compiler is fully compliant with the ISO FORTRAN 95 standard.
- Mixed language support (C, FORTRAN): The Intel FORTRAN Compiler supports mixed language programming and debugging between C and FORTRAN.

Intel C++ Compiler

The Intel C++ Compiler delivers optimal performance on all Intel processors. It offers improvements to its advanced optimization features and interprocedural optimization and profile-guided optimization. In addition, the Intel C++ Compiler has increased levels of Linux and industry standards support that provide improved compatibility with GNU C, a wider support of Linux distributions, and support for the C++ ABI object model, which enables stronger binary compatibility with gcc version 3.2. It supports glibc 2.2.4, 2.2.5, or 2.3.2. It also provides development and optimization capability for threaded applications through its enhanced support of the OpenMP™ 2.0 standard for C/C++, and the Auto-Parallelization preview feature.

C++ compatibility features:

- Substantial GNU C compatibility features: The Intel C++ Compiler for Linux is substantially source- and object-code compatible with GNU C. This allows recompilation of existing software with the Intel C++ Compiler as a simple way to add performance to applications. Alternatively, applications can be built by compiling specific modules with the Intel C++ Compiler and linking them with modules compiled with GNU C. This is especially useful if one wants to start using the Intel compiler on a few modules first. The Intel C++ Compiler for Linux includes language features that provide the capability to build the Linux kernel with minor modifications.
- Compatible with widely used Linux software utilities: The Intel C++ Compiler is substantially compatible with common tools used in developing Linux applications at such as make, Emacs, and gdb.

For a detailed description of the complete optimization features available with the Intel compilers, please refer to the following URL:

http://www.intel.com/software/products/compilers/

Other Development Software

Intel Math Kernel Library (MKL): The optional Intel MKL Cluster Edition is composed of highly optimized mathematical functions for engineering and scientific applications requiring high performance on Intel platforms. The functional areas of the library include linear algebra consisting of LAPACK and BLAS, Fast Fourier Transform (FFT), vector transcendental functions, ScaLAPACK, and distributed memory FFTs.

Python: Python, an interpreted, object-oriented, high-level programming language is included as part of the SLES10 operating system.

MPI Python: MPI Python (MYMPI) is fully supported under the SLES10 operating system as part of the Altix ICE system.

Java: Java2 1.4.2 is included in SLES10. BEA JRockit is the first high-performance Java Virtual Machine (JVM) uniquely optimized for Intel platforms, enabling Java applications to run with increased reliability and performance on lower cost, standards-based platforms. Unlike other JVMs, BEA JRockit is designed to power demanding server-side Java applications – delivering superior performance, manageability, and reliability for enterprise applications. Java2 is available from BEA JRockit (a third-party vendor). Java 2.5 is available from BEA JRockit.

Batch Queuing

The Altix ICE 8200 cluster system can be optionally configured with an advanced batch queuing system. The preferred workload management software for Altix ICE systems is PBS Professional (PBS Pro) from Altair Engineering.

The Portable Batch System™ (PBS) is a workload management solution for HPC systems and Linux clusters. PBS was originally designed for NASA because existing resource management systems were inadequate for modern parallel/distributed computers and clusters. From the initial design forward, PBS has included innovative new approaches to resource management and job scheduling, such as the extraction of scheduling policy into a single separable, completely customizable module. The professional edition, PBS Pro, operates in networked multi-platform UNIX environments and supports heterogeneous clusters of workstations, supercomputers, and massively parallel systems.

Sites using PBS Pro to manage their computing resources can expect many tangible benefits including:

- Increased utilization of costly resources.
- Unified interface to all computing resources.
- Reduced burden on system administrators freeing them to focus on other activities.
- Enhanced understanding of computational requirements and user needs.
- Expandability: PBS Pro supports dynamic distribution of production workloads across wide-area networks, and the logical organization of physically separate computing systems.

Key features of PBS Pro include:

- User interfaces: Designed for production environments, xPBS provides a graphical interface for submitting both batch and interactive jobs; querying job, queue, and system status; and tracking the progress of jobs. Also available is the PBS command line interface (CLI) providing the same functionality as xPBS.
- Job priority: Users can specify the priority of their jobs and defaults can be provided at both the queue and system level.
- Job interdependency: PBS enables the user to define a wide range of interdependencies between batch jobs. Such dependencies include: execution order, synchronization, and execution conditioned on the success or failure of a specified other job.
- Cross-system scheduling: PBS provides transparent job scheduling on any system by any authorized user. Jobs can be submitted from any client system or any compute server.

- Security and access control lists: Configuration options in PBS permit the administrator to allow or deny access on a per-system, per-group, and/or per-user basis.
- Job accounting: For chargeback or usage analysis, PBS maintains detailed logs of system activity. Custom charges can be tracked per-user, per-group, and/or per-system.
- Desktop cycle harvesting: Take advantage of otherwise idle workstations by configuring them to be used for computation when the user is away.
- Comprehensive API: Included with PBS is a complete application programming interface (API) for sites that want to integrate PBS with their applications or have unique job scheduling requirements.
- Automatic load leveling: The cluster jobs scheduler provides numerous ways to distribute the workload across a cluster of machines: based on hardware configuration, resource availability, and keyboard activity (all this information is available via PBS).
- Enterprise-wide resource sharing: PBS does not require that jobs be targeted to a specific computer system. This allows users to submit their jobs and have them run on the first available system that meets their resource requirements. This also prevents waiting on a busy system when other computers are idle.
- Username mapping: PBS provides support for mapping user account names on one system to the appropriate name on remote server systems. This allows PBS to fully function in environments where users do not have a consistent username across all the resources they have access to.
- Parallel job support: PBS supports parallel programming libraries such as MPI, MPL, PVM, and HPF. Such applications can be scheduled to run within a single multiprocessor system or across multiple systems.
- Fully configurable: PBS can be easily tailored to meet the needs of different sites. The Job Scheduler module was designed to be highly configurable.
- Automatic file staging: PBS provides users with the ability to specify any files that need to be copied onto the execution host before the job runs, and any that need to be copied off after the job completes. The job is scheduled to run only after the required files have been successfully transferred.

PBS Professional 8.0 is the first version that will support SGI ProPack 5. PBS Pro supports features unique to the SGI environment and related ProPack environment. It makes use of the SGI "cpuset" facility to gain control over system partitioning, thus increasing the consistency of job runtimes within large (> 32 CPUs) systems. PBS Pro support for cpusets permits job scheduling to be restricted to a dynamically defined set of processors. PBS Pro also supports SGI Process Aggregation Groups (PAGG) and provides integration with SGI Comprehensive System Accounting (CSA) for Altix systems.

Cluster Management

The Encanto Cluster uses "SGI Tempo Cluster Management Software". This cluster management software provides a single point of administration for the entire system and enables the system to be administered as a single system. Every SGI Altix ICE 8200 is configured with an administrative node which, in conjunction with SGI Tempo, automatically provisions and functions all other nodes in the hierarchical management scheme, an implementation designed to improve automation and enable the management infrastructure to scale as the customer's computational needs scale. From the administrative node an administrator can manage the system as either a single computing system or at a granular level.

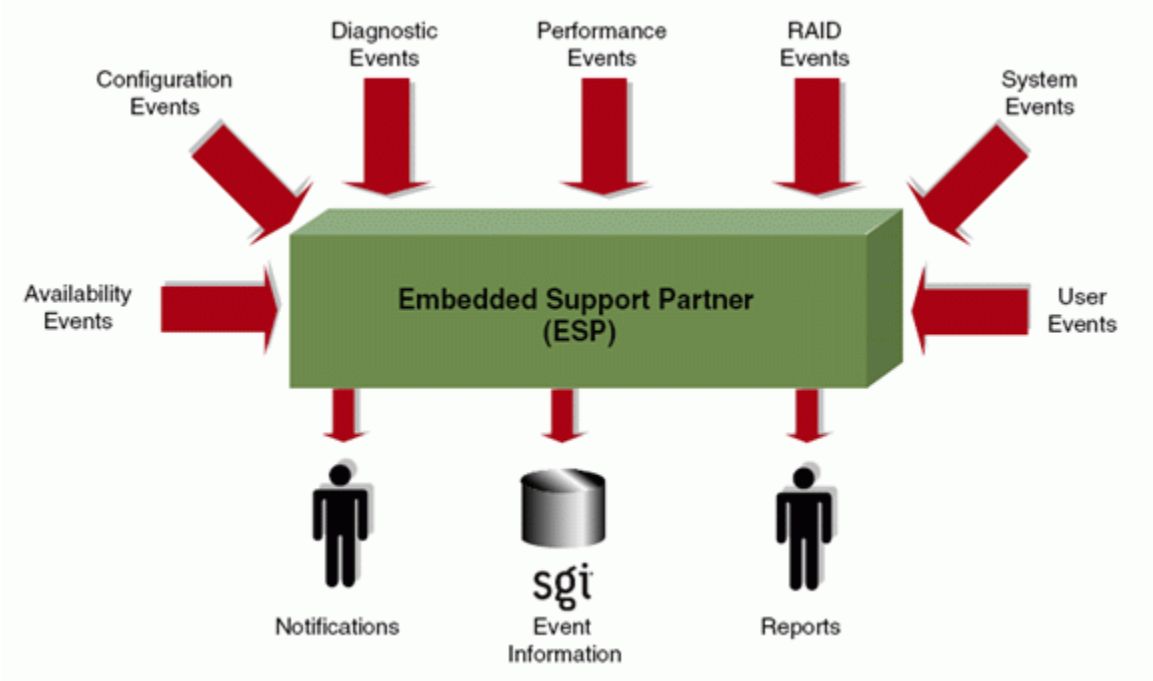
Some features are:

- Issue system-wide as well as node-specific commands such as power on and off.
- Collect and output information on all or some components within the cluster.

- Provide an interface to monitor all or some of the components within the cluster.
- Provide an interface to discover all or some of the components within the cluster.
- Manage and provision via a single interface all the compute node boot images created and stored in the cluster, as well as assignments of specific boot images to a specific range of compute nodes.

Status Information Tools

The Encanto Cluster includes management tools that will provide status information for all the major components of the cluster. These management tools will provide status information for major components within the cluster including: compute blades (nodes), blade enclosures (also called the Individual Rack Unit or IRU), power supplies, blowers, Gigabit networks, InfiniBand networks, and all service nodes (i.e. System Admin Controllers, Rack Leader Controllers, Login nodes, Storage nodes, etc.). Administrative tools will be scalable so that global administrative commands can be completed in a short time.



The Altix ICE 8200 cluster employs administrative tools that are hierarchical by design and are highly scalable. Global administrative commands occur in parallel, using the hierarchical design of the Altix ICE system. By processing these commands in parallel, they can be completed in a much shorter timeframe.

The hierarchical design starts with the System Admin Controller. This administration node can communicate with all components within the management infrastructure and utilizes the database caching capabilities of the Rack Leader Controllers to further improve management scalability. Each Rack Leader Controller communicates to all the chassis management controllers (CMC) within each blade enclosure (IRU) located within the rack. And each CMC communicates with the IRU's power supplies, fans, and compute blades. As a result, an administrative command is issued to each Rack Leader Controller and propagated to each CMC.

Therefore, commands are delivered in parallel in a highly efficient and scalable fashion.
The SGI Altix ICE 8200 system uses a hierarchical management design that provides for easy and automated upgrades. Additional SGI racks (and Rack Leader Controllers), blade enclosures (IRU), and compute blades may be added to the system and discovered as an available resource. Because each rack and blade enclosure (IRU) is a distinct part of the structure, it has also been designed as a "modular" unit that can be added incrementally to the management infrastructure. The diskless compute blades also facilitate rapid deployment of additional compute blades. As each compute blade is discovered, it becomes available for any existing system image (boot image) to start running on that blade, so administrators do not have to build an image for every new node as they would on a conventional server node.

In addition, as every SGI Rack Leader Controller, Login node, Batch Scheduler node, or Gateway node is introduced into the system, it is provisioned by the System Admin Controller. This allows for easy installation as well as ensures that all Rack Leader Controllers are predictably running the same software.

PBS Pro provides comprehensive accounting and utilization logging that tracks the requested and actual utilization of each job in an easily parsed, human-readable ASCII format. PBS also supplies tracking and reporting utilities to extract and analyze this data. Since the log format is well-documented and easy to parse, custom reporting and analysis is easily implemented with standard UNIX/Linux utilities (e.g. awk, perl, sed, etc.).

PBS Pro is robust, field-proven grid cluster infrastructure software that increases productivity even in the most complex of computing environments. Resource managers and administrators at more than 1400 sites worldwide use PBS Pro to simplify job submission, enabling scientists, designers, financial analysts, and engineers to remain focused on their fields of expertise. It efficiently distributes workloads across cluster, SMP, and hybrid configurations, scaling easily to hundreds or even thousands of processors.

System partitioning, user controls, and network access controls can be imposed to ensure against unauthorized data access to the proposed SGI Cluster. Users can only access the cluster via the service nodes. User access controls will be completely implemented on these service nodes. The system security and access controls can be defined and implemented by SGI Professional Services engineers. Various service nodes for the Altix ICE 8200 provide specific functions and capabilities. It is the responsibility of the Batch Service node to configure and control system partitioning. User access to the main cluster and storage sub-systems is established and enforced by one of the several Login nodes.

As far as system network access, the internal networks on the Altix ICE 8200 are not directly connected to the customer network. They are only connected to the service nodes. By default, there are no network gateways defined to route traffic automatically between the customer network and the internal system networks. If network gateways are created, they would only be defined with security considerations appropriate to the site.

The proposed SGI Cluster provides complete job queue monitoring and information about the status of jobs both locally and remotely. Customers may, at their discretion, prevent users from obtaining information about other users' jobs with a configuration option. PBS also supplies an X-based GUI, xpbs, which provides an auto-refreshed display of this same information as well as a convenient job submission GUI.

Ganglia a software graphic interface shows the administrator pictures of the loads and other useful data. The following figure shows historical Ganglia charts for system cluster loads.

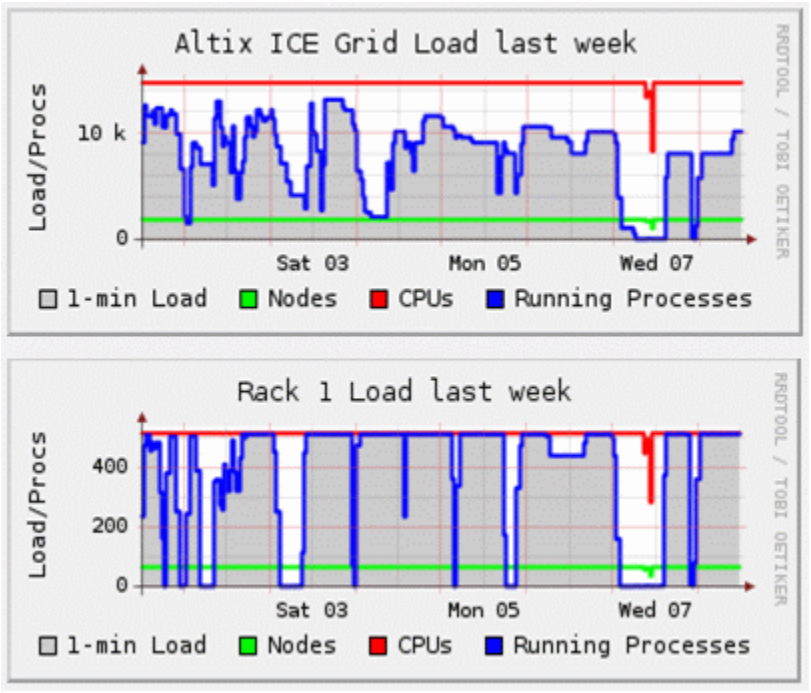


Figure 23. Typical Ganglia Web-based Load Chart

Test Environment

Within the Encanto "Switch Rack" resides a very small "test or development system". This system consists of a 16-node, 128-core, Altix ICE 8200 and a single Login node. This test system will allow the Encanto system administrator to install and test new system software, utilities, and applications to ensure that they function properly before deploying them to the production cluster. The following diagram shows the physical layout of this test system. Note that this same rack houses the Cisco 6509 network switch that provides Lambda Rail connectivity.

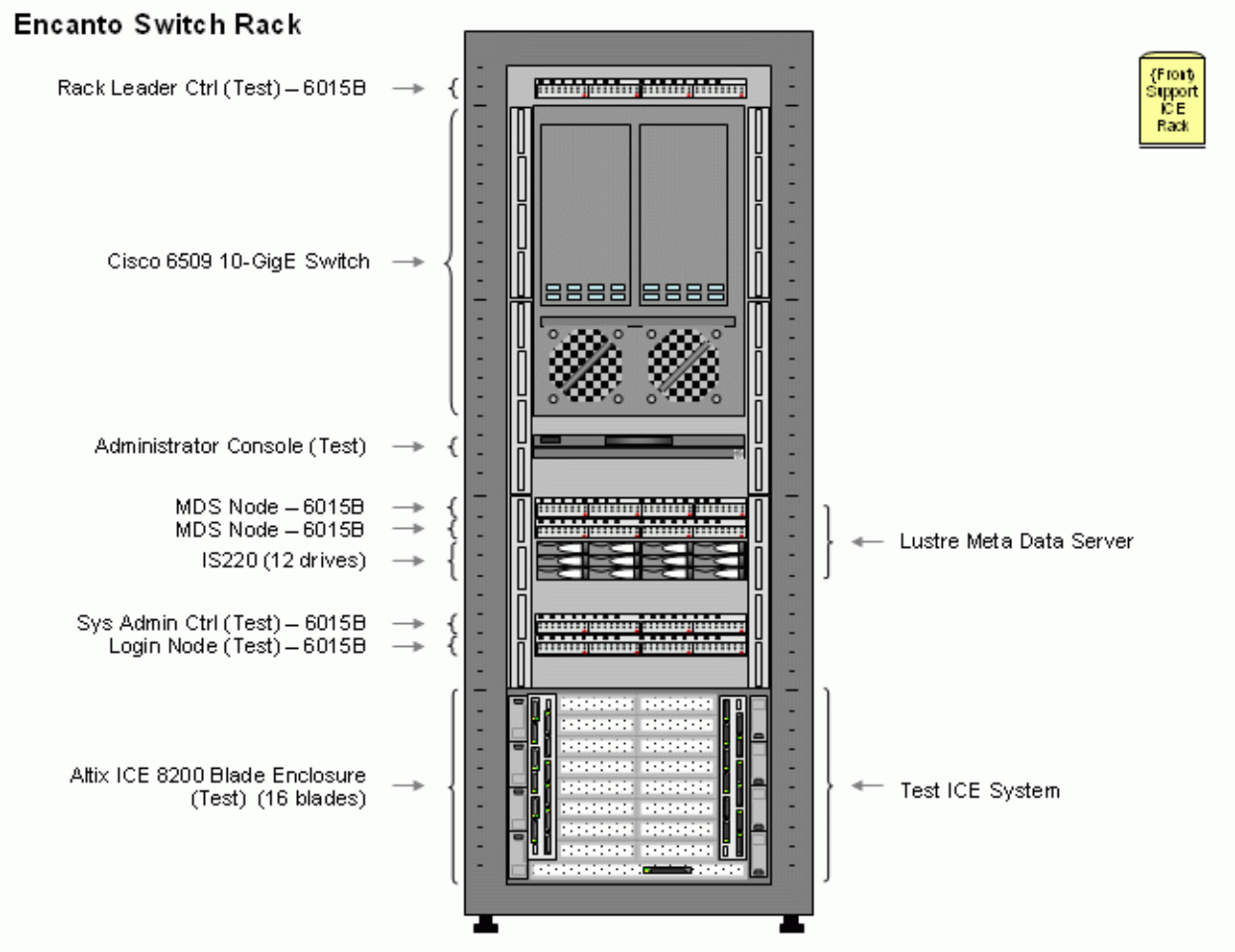
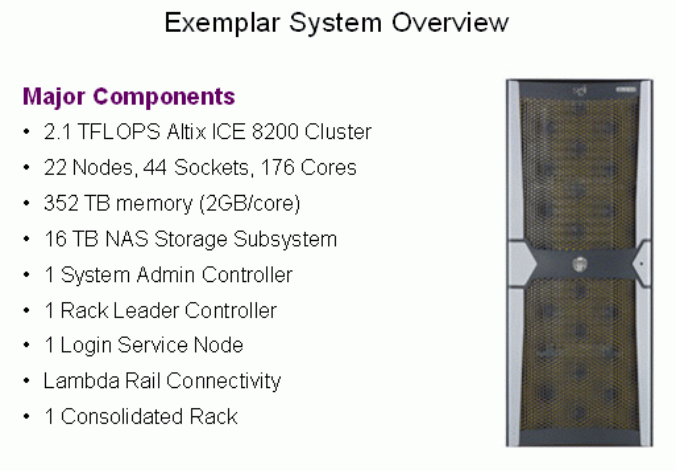


Figure 24. Encanto "Switch" Rack Physical Layout (with Test System)

Exemplar Systems

In addition to the main Encanto Altix ICE 8200 supercomputer system, three small Altix ICE "Exemplar" systems have been deployed to three different Universities/Colleges within the State of New Mexico.



These small Altix cluster systems are each approximately 1.2% the compute power of Encanto. Each of these systems are configured with 22 compute blades (nodes) each provisioned with two, quad-core Xeon 3.00 GHz processors and 16 GB of memory. These Altix ICE 8200 systems each have a single Login Node to support a number of local users and developers. Each system also includes a small, 10TB, NAS storage subsystem.

The following is a block diagram of an Exemplar system.

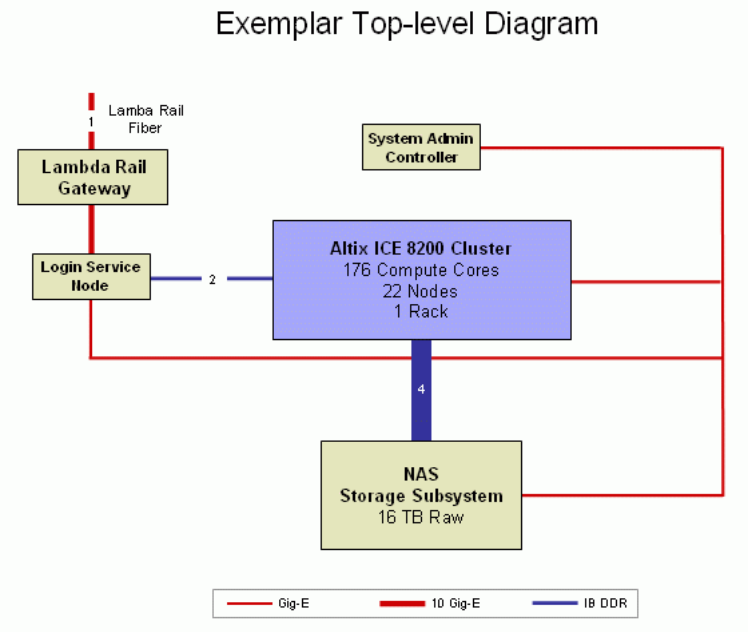


Figure 25. Exemplar Top-level Block Diagram

The following diagram shows the physical layout of an Exemplar system.

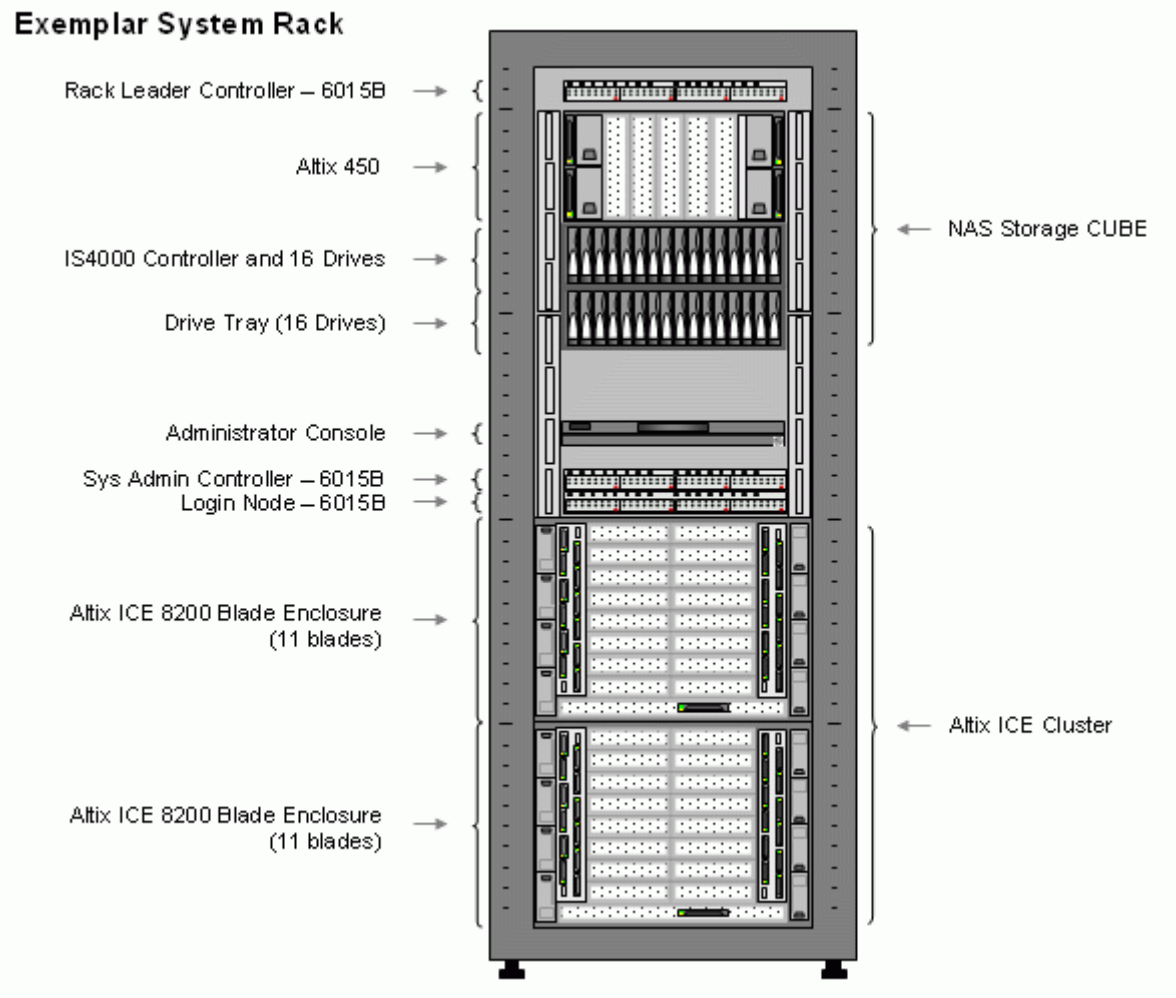


Figure 26. Exemplar Rack Physical Layout

Software Inventory

The following table enumerates the various software packages as part of this proposal. Note that many of these packages are part of the SGI ProPack.

Table 4. NMCAC Software Inventory

Open Source Software	Closed Source Software
Novell SLES10	SGI ProPack Cluster Tools
SGI external interrupt (extint) driver header file	CPUSET processor and memory placement utilities
SGI IOC4 driver for x86_64	ESP for Linux Package
A Debug Version of the Kernel	Pre-loadable FFIO Library
The Standard Kernel	REACT Real-time development library
Kernel for kdump	LK - License Keys
Kernel with Multiprocessor Support	The Performance Co-Pilot™ visualization tools
The Linux Kernel Sources	SGI REACT documentation
Kernel Symbol Versions (modversions)	SGI ProPack documentation
Multi-word bitmask abstract data type	SGI ProPack support tools
CPUSET processor and memory placement library	Rhino System Administration Client files
LSI Fusion-MPT host adapter management utility	Rhino System Administration libraries
Utilities to Load Modules into the Kernel	Rhino System Administration Server files
NUMA utilities	Cluster Manager GUI client software
Library for Configuring Boot Loaders	XVM GUI client software
SGI REACT™ simple configuration	XVM GUI server software
The SGI public gpg key for rpm signature verification	XVM GUI web server software
SGI external interrupt (extint) driver header file	XVM Command line tools
	Displays or modifies model-specific registers (MSRs)

Open Source Software	Closed Source Software
SGI IOC4 driver for x86_64 NUMA utilities SGI ProPack configuration recommendations SGI Linux release file A set of tools for SCSI Disks A udev helper for doing XSCSI device names	Intel Compiler Runtime Libraries

Encanto - Software List

Type	Description
Open	SLES10 Linux Operating System
Closed	HPC cluster/compute
Closed	SGI ProPack Server
Closed	SGI ProPack for HPC Compute Nodes
Closed	SGI Tempo management suite
Closed	Ganglia
Closed	SystemImager
Closed	Conserver
Closed	IPMI-based Tools
Closed	Embedded Support Partner (ESP)
Closed	InfiniBand Fabric Subnet Management Software (based on OFED v1.2rc2 and OpenSM 1.2.0)
Closed	PBSPro Scheduling System
Closed	Time Navigator Server for Linux
Closed	Time Navigator Agents for Windows
Closed	Time Navigator Tape Drive Connection for Linux
Closed	Time Navigator Multiprocessor option
Closed	Time navigator Agent for Unix
Closed	Device drivers for HP LTO family of tape drives
Closed	ISSM Workgroup Edition for SGI InfiniteStorage 220 - for Linux
Open	Lustre Open Source Software (with install & support)
Closed	INFINITESTORAGE NAS Manager
Closed	TPSSM license and S/W for Linux hosts
Closed	Intel EM64T Fortran, C, and C++ compilers
Closed	Intel Math Kernel Library (MKL) Cluster Edition
Closed	Intel MPI
Open	MVAPICH2
Closed	SGI MPT
Closed	Intel Debugger (idb)
Closed	Intel Trace Analyzer and Collector

Open	GNU gdb
Open	Python
Open	MPI Python
Open	Java