WHITE PAPER

SGI® Altix® ICE

Selecting the Most Effective InfiniBand Topology

TABLE OF CONTENTS

1.0 Introduction

Across a wide range of disciplines, InfiniBand technology now enables clusters that range from a few systems to the largest technical computing clusters in the world. In only a few years, clustering with InfiniBand has come to easily dominate the top 100 of the Top500 list of supercomputing sites (*www.top500.org*). As new grand-challenge problems and other computational challenges emerge, larger and larger clusters will be required. Even with now-routine advances in processor speed and memory capacity, scaling with cluster size will likely remain the simplest way to grow computational capacity for the world's most tenacious computational problems. While InfiniBand can be deployed in multiple topologies, choosing the optimum InfiniBand topology can be difficult, with trade-offs in terms of scalability, performance, and cost.

SGI has considerable experience in the design and deployment of some of the largest InfiniBand clusters in existence. While some vendor's limitations drive them to push one topology choice above others, SGI understands that the best topology is one that matches the needs of the application. Based on high-performance Intel® Xeon® processors, the SGI® Altix® ICE rackmount system is designed for flexible and optimized InfiniBand topology configuration.

Beyond offering compelling InfiniBand-based products, SGI has done extensive performance testing of InfiniBand cluster topologies using the SGI Altix ICE system. This paper describes the clustering capabilities of SGI Altix ICE in the interest of helping organizations make informed choices as they compare and select InfiniBand topology for their applications.

2.0 InfiniBand Topology Considerations and Trade-offs

SGI Altix ICE supports multiple InfiniBand topology choices, including All-to-All, Fat Tree (CLOS), as well as Hypercube and Enhanced Hypercube topologies. Choosing the right topology involves understanding the needs of the application as well as comparing key metrics and cost implications.

2.1 SGI Altix ICE Topology Choices

InfiniBand fabrics present different advantages and limitations. The SGI Altix ICE system is designed to flexibly support multiple InfiniBand topologies, including:

- *All-to-All*. All-to-All topologies are ideal for applications that are highly sensitive to Message Passing Interface (MPI) latency since they provide minimal latency in terms of hop-count. Though All-to-All topologies can provide non-blocking fabrics, and high bisection bandwidth, they are restricted to relatively small cluster deployments due to limited switch port counts.
- *Fat Tree*. Fat Tree or CLOS topologies are well suited for smaller node-count MPI jobs. Fat Tree topologies can provide non-blocking fabrics and consistent hop counts resulting in predictable latency for MPI jobs. At the same time, Fat Tree topologies do not scale linearly with cluster size. Cabling and switching become increasingly difficult and expensive as cluster size grows, with very large core switches required for larger clusters.
- *Standard Hypercube*. Standard Hypercube topologies are ideal for large node-count MPI jobs, provide rich bandwidth capabilities, and scale easily from small to extremely large clusters. Hypercubes add orthogonal dimensions of interconnect as they grow, and are easily optimized for both local and global communication within the cluster. Standard Hypercube topology provides the lightest weight fabric at the lowest cost with a single cable typically used for each dimensional link.
- *SGI Enhanced Hypercube*. Adding to the benefits of Standard Hypercube topologies, SGI Enhanced Hypercube topologies make use of additional available switch ports by adding redundant links at the lower dimensions of the hypercube to improve the overall bandwidth of the interconnect.

2.2 Hypercube vs. 3D Torus Topologies

Torus topologies are not currently supported by SGI as they can present scalability challenges and introduce variable hop count and latency with increasing cluster size. Inconsistent hop count can result in unpredictable application behavior due to uneven latency between nodes. It is worth noting that Hypercube topologies in particular can be viewed as "supersets" of Torus topologies. Hypercubes also provide addititional connectivity and offer numerous advantages, including:

- Hypercube topologies provide linear latency (hop count) scalability as compared with Torus fabric (Figure 1), and offer higher connection capabilities and resources at the node level.
- Orthogonal dimensions of the interconnect are added with every doubling in hypercube cluster size, and each dimension of the hypercube interconnect scales linearly as cluster size increases.
- Hypercube topologies also allow switch requirements to scale linearly with system size, with cabling that is distributed and more reliably managed.
- The dimensional aspects of Hypercube topologies allows independent paths for local and global traffic, facilitating greater available bandwidth.

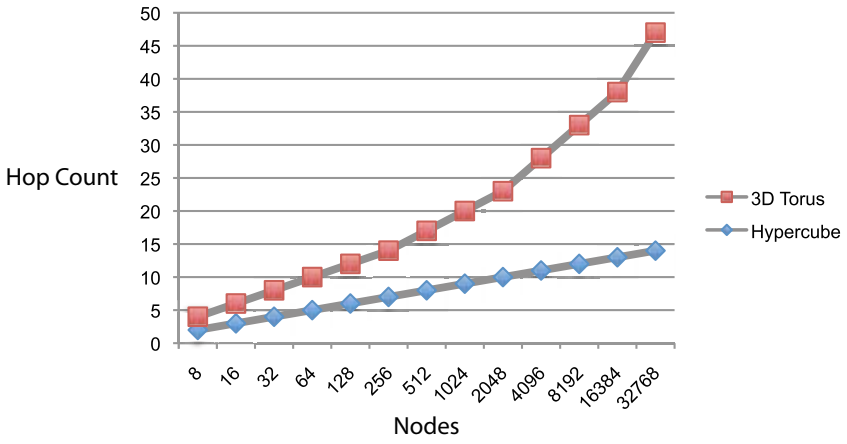

Figure 1. *Hypercube topologies provide consistent and linear hop-count scalability vs. 3D Torus topologies.*

2.3 Interpreting Key Metrics

In evaluating different InfiniBand topologies, it is essential to understand key metrics related to their performance and cost. Unfortunately, some metrics designed around one topology can be misleading when applied to another topology.

Latency

Depending on the application, latency can be an important consideration. A side effect of topology choice, latency is typically expressed in terms of hop count between nodes. Different topologies imply different levels of latency. Table 1 illustrates varying levels of latency by topology across a two-rack 128-node SGI Altix ICE configuration.

		Latency — Maximum Number of Hops		
Number of Switches/rail	Number of Nodes	Hypercube	All-to-All	Fat Tree
1	8	2	2	2
2	16	3	3	4
4	32	4	3	4
8	64	5	3	4
16	128	6	3	4

Table 1. *Two-rack SGI Altix ICE latency across multiple topologies*

Blocking Factor and Bisection Bandwidth

Blocking factor and minimum bisection bandwidth metrics were initially designed to compare between different Fat Tree implementations. As such, these metrics often focus across the system at the top of the fabric only. Blocking factor and minimum bisection bandwidth can easily be misinterpreted if taken literally for topologies other than Fat Tree. For example, mapping Fat Tree metrics such as these onto hypercube-based topologies inherently understates the capabilities of these designs, where job traffic patterns are evenly distributed throughout the system.

Though hypercube based topology job traffic can be orthogonal and include multiple paths, blocking factor and bisection bandwidth only allow one path to be counted by definition. Instead, hypercube-based topologies can be described by the aggregate bandwidth of all dimensions of the hypercube interconnect. For optimum job performance, the focus should be at the node level, or within the fabric just above the node — where most of the job traffic occurs. As shown in Table 2, depending on chosen topology and the number of rails (or planes), SGI Altix ICE system design provides substantial bisection bandwidth. For more detailed information, Appendix A includes tables that detail blocking ratio and bisection bandwidth across each set of dimensional links for SGI supported topologies.

		Theoretical Bisection Bandwidth — GB/s/node						
		Standard Hypercube		Enhanced Hypercube		All-to-All		Fat Tree
Switches/Rail	Nodes	1 Rail	2 Rails	1 Rail	2 Rails	1 Rail	2 Rails	1 Rail
1	8	4	8	4	8	4	8	4
2	16	2	4	4	8	2	4	4
4	32	0.5	1	4	8	1	2	4
8	64	0.5	1	2	4	2	4	4
16	128	0.5	1	2	4	4	8	4

Table 2. *Two-rack Altix ICE topology global bandwidth comparison.*

Cost

As InfiniBand clusters move into petascale deployments and beyond, predictable infrastructure costs are as key to managing growth as increasing bandwidth requirements. In this context, cost is often a reflection of infrastructure complexity in terms of external switching and cabling. Because Fat Tree topologies require more cables and external switching, infrastructure costs do not scale linearly as shown in Figure 2. In contrast, Standard Hypercube and Enhanced Hypercube topologies provide lower complexity and consistent and linear switch infrastructure costs.

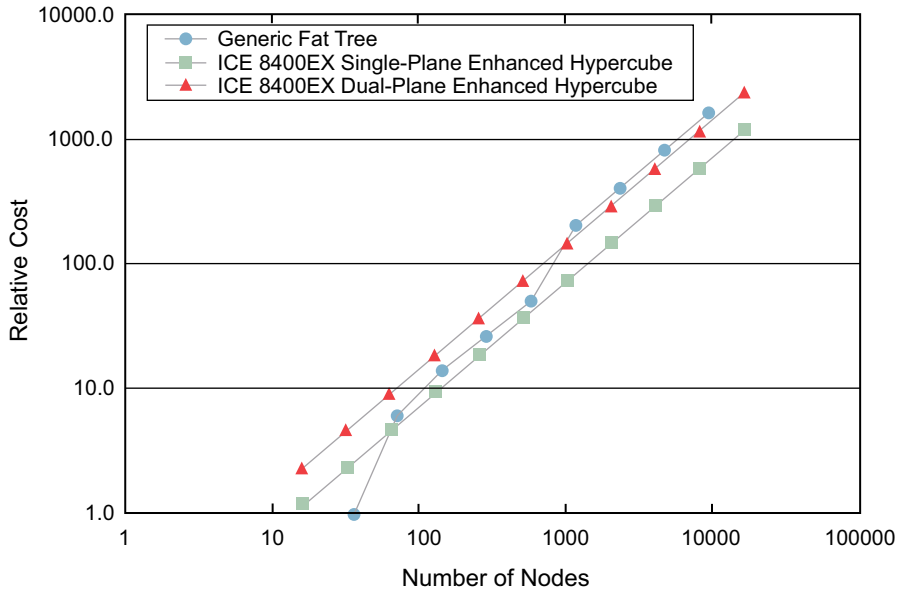

Figure 2. *Hypercube topologies provide linear and predictable infrastructure costs.*

3.0 SGI Altix ICE: Designed for InfiniBand

The SGI Altix ICE platform is fundamentally architected to provide cost-effective high-performance InfiniBand infrastructure. With Intel Xeon processor 5600 Series as its multi-core engine, the SGI Altix ICE 8400 platform in particular is capable of achieving industry-leading scalability without sacrificing application performance efficiency. The platform offers a variety of interconnect options that let organizations scale their applications across hundreds or thousands of processor cores.

The SGI Altix ICE 8400 system can accommodate up to 16 compute blades within each Individual Rack Unit (IRU). The IRU is a 10 rack unit (10U) chassis that provides power, cooling, system control, and network fabric for up to 16 blades via a backplane. Up to four IRUs are supported in each custom-designed 42U rack, with a choice of either air cooling or water cooling for all configurations. Each rack supports:

- A maximum of four IRUs
- Up to 64 compute blades (up to 128 Intel Xeon Processor sockets, and 768 processor cores)
- A maximum of 12.2TB of memory (64 x 192GB)

3.1 Single–or Dual-Plane Design

Fast interconnect speeds are essential for technical computing, and multi-rail networks can help optimize communication traffic. Multi-rail networks can improve MPI communication performance by dividing large messages into chunks, and distributing those chunks across multiple independent InfiniBand networks, or rails. Messages can also be circulated in round-robin fashion across multiple raile. Multi-rail systems can also separate MPI communication from I/O traffic, effectively dedicating separate networks for MPI and I/O respectively. With the dual-plane design provided by SGI Altix ICE 8400, MPI applications can make use of both InfiniBand rails. Four topologies are currently supported on the SGI Altix ICE 8400 system:

- All-to-All (from 1 to 8 IRUs)
- Standard Hypercube and Enhanced Hypercube
- Fat Tree

All topologies are based on the 4x Quad Data Rate (QDR) InfiniBand interconnect. The All-to-All, Hypercube, and Enhanced Hypercube topologies do not require external InfiniBand switches, but are instead built by interconnecting the internal InfiniBand ASICs on switch blades that are inserted into the IRU. The switch blades provide the interface between compute blades within the same IRU and also between compute blades in separate IRUs. Each switch blade contains one 4x QDR (ConnectX-2) 36-port InfiniBand ASIC. SGI provides a choice of backplanes, switch blades, and compute blades to support a range of topologies as described in the sections that follow.

3.2 Eight-Node Bristled Backplane

The Eight-Node Bristled Backplane and associated QDR switch blade are optimized for Standard Hypercube, Enhanced Hypercube, and All-to-All topologies. Illustrated in Figure 3, each group of eight compute blades connects to one 36-port QDR InfiniBand switch ASIC on each plane, providing non-blocking connectivity within the group of eight blades. The two groups of eight compute blades are then connected together through the backplane by four links on each plane. This arrangement provides 2:1 blocking on each plane, but results in a non-blocking (1:1) configuration when the two planes are combined together (eight nodes on each side connected together with eight links). In hypercube topologies, the first hypercube dimension is quad-linked on the backplane by default.

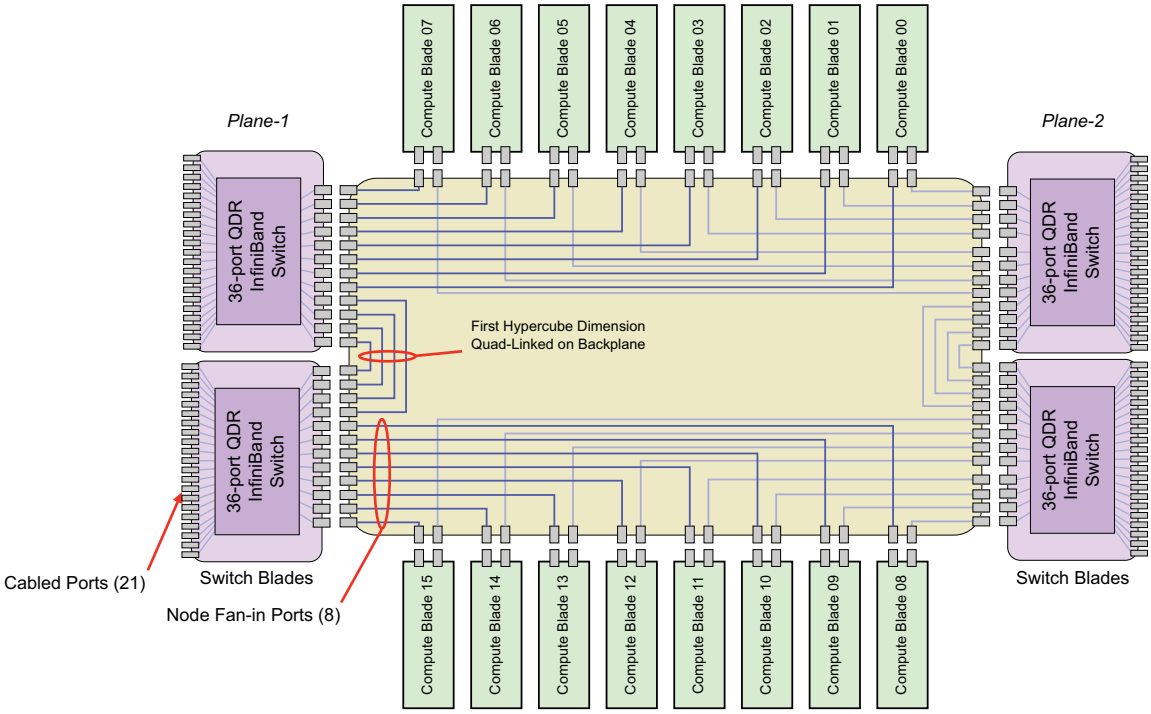

Figure 3. *The Eight-Node Bristled Backplane uses two associated QDR switch blades on each plane.*

The Eight-Node Bristled Backplane uses two switch blades configured on each plane of each IRU. A total of 21 InfiniBand ports are brought out to external connectors on the switch blade. Four of the 21 external ports use mini-SAS connectors. The remaining seventeen ports use Quad Small Form Factor Pluggable (QSFP) connectors. Single plane (single-rail) configuration is achieved with two installed switch blades. Dual-plane (dual-rail) configuration adds two additional switch blades.

3.3 Sixteen-Node Bristled Backplane

The Sixteen-Node Bristled Backplane and associated QDR switch blade are ideal for smaller Fat Tree configurations that can require fewer InfiniBand connections than other topologies. As shown in Figure 4, all sixteen compute blades in each IRU connect to a 36-port QDR InfiniBand switch ASIC on the switch blade. One switch is configured on each plane. This configuration creates a non-blocking topology within the 16 compute blades, and the switch supports non-blocking topology exiting the IRU.

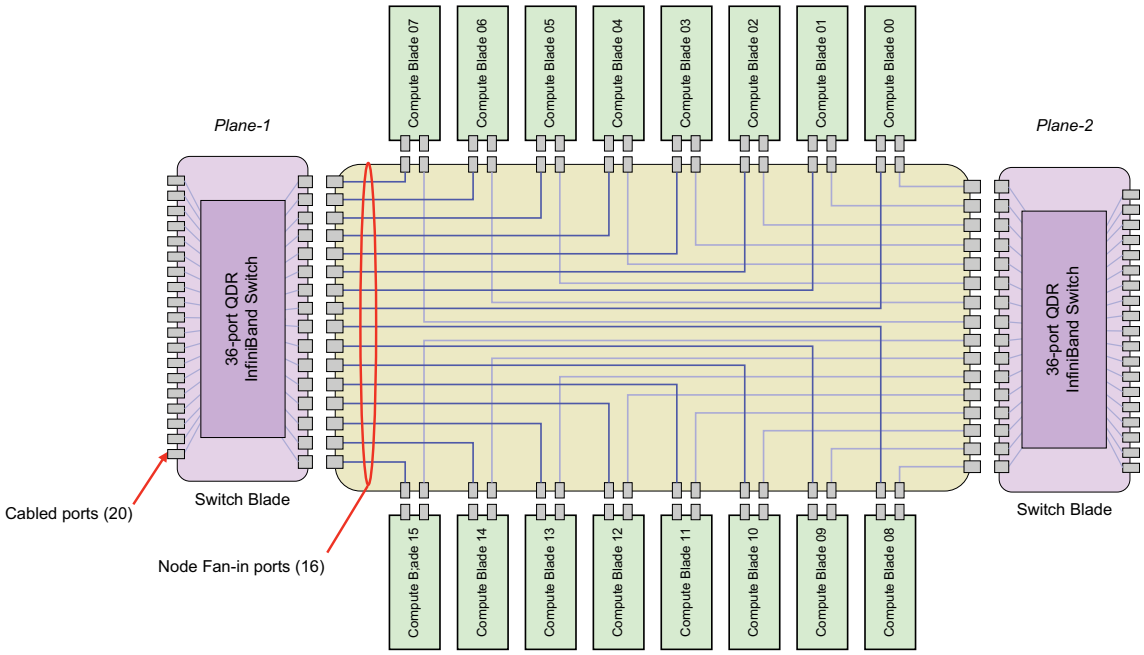

Figure 4. *The Sixteen-Node Bristled Backplane uses a single associated switch blade for each plane.*

The Sixteen-Node Bristled Backplane uses one or two InfiniBand switch blades to connect to a network fabric for either a single- or dual-rail topology. Each switch blade supports 20 external QSFP InfiniBand ports. Dual-rail topologies provide enhanced bandwidth as well as redundancy in case of link failure.

3.4 A Choice of Compute Blades

Depending on the target application and other factors, a multi-rail network topology may not automatically benefit application performance. Application communication patterns ultimately determine whether performance benefit can be derived from the extra bandwidth available in dual-rail topologies. Applications may also not be written to take advantage of the additional network topology at their disposal. For these reasons, SGI offers a choice of compute blades for the SGI Altix ICE 8400 system, including:

- The SGI IP103 compute blade provides a single dual-port ConnectX InfiniBand HCA on a PCIe x8 bus
- The SGI IP105 compute offers two single-port ConnectX InfiniBand HCAs, each on their own dedicated PCIe x8 bus

SGI IP105 blades provide the highest available bandwidth for dual-rail applications since striping a single large message across both ports can almost double the MPI communication bandwidth without being limited by PCIe bus bandwidth.

4.0 Configuring SGI Altix ICE for Various Topologies

When making topology choices, it is helpful to understand how those topologies are implemented on the underlying architecture of the selected platform. The sections that follow describe how various topologies can be configured on the SGI Altix ICE platform.

4.1 Configuring for All-to-All

Figure 5 illustrates an All-to-All topology configured using SGI Altix ICE. Using the Eight-Node Bristled Backplane in each IRU, a two-rack (128-node) All-to-All topology can be built with the SGI Altix ICE 8400. Eight IRUs are provided in two racks. Each bolded (blue) line in the illustration represents the quad links between the two 36-port switch ASICs in each IRU. Each regular (black) line represents a single cabled link.

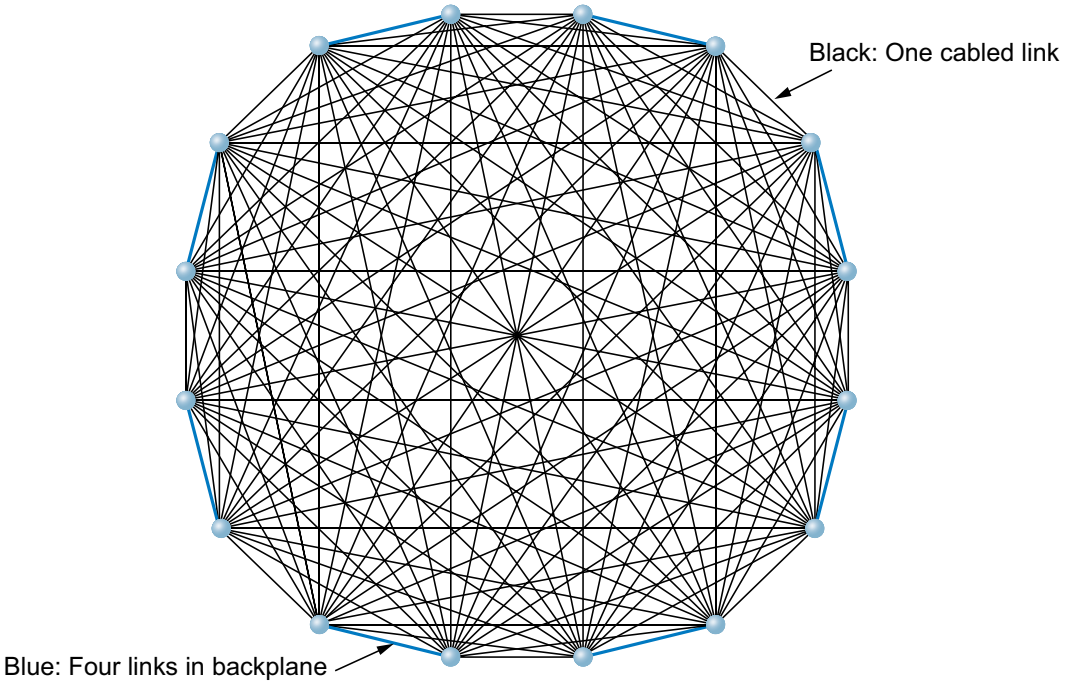

Figure 5. *A two-rack SGI Altix ICE 8400 All-to-All topology configuration employs quad backplane inks within each IRU.*

4.2 Configuring for Standard and Enhanced Hypercube

SGI Altix ICE system design lends itself well to a broad range of Standard Hypercube and Enhanced Hypercube topologies.

- A single IRU represents 1-dimensional (1D) Hypercube
- Two IRUs combine to form a 2D hypercube
- A 3D hypercube requires one rack of hardware (four IRUs)
- Two racks can be combined to form a 4D hypercube

Figure 6 illustrates hypercube dimensions one through four.

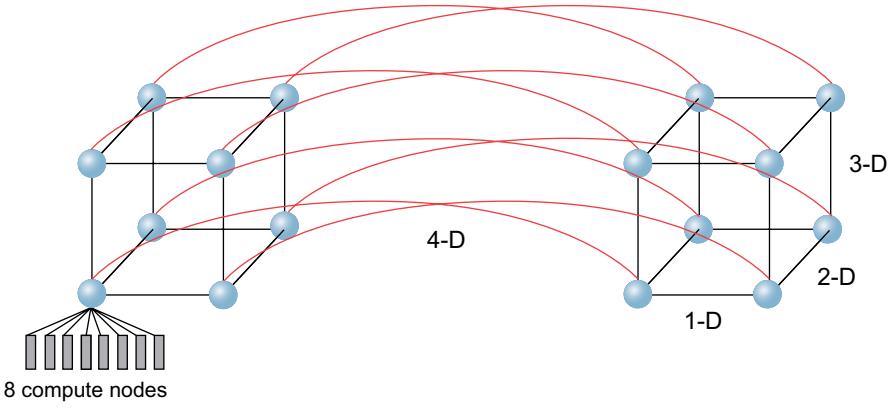

Figure 6. *Eight compute nodes exist at each vertex of a 4D Hypercube or Enhanced Hypercube topology.*

The dual-plane design of the SGI Altix ICE 8400 backplane means that this illustration can be configured as:

- A Standard Hypercube with 1D connections quad linked, and 2D, 3D, and 4D connections single-linked
- An Enhanced Hypercube with 1D and 2D connections oct-linked, 3D and 4D connections quad-linked

Hypercubes can be built to very large dimensions and very large numbers of nodes using the SGI Altix ICE 8400 system as shown in Table 2.

SGI Altix ICE IRUs/Racks	Hypercube Dimension	Number of Nodes
1 IRU	1D	16 Nodes
2 IRUs	2D	32 Nodes
1 Rack	3D	64 Nodes
2 Racks	4D	128 Nodes
4 Racks	5D	256 Nodes
8 Racks	6D	512 Nodes
16 Racks	7D	1024 Nodes
32 Racks	8D	2048 Nodes
64 Racks	9D	4096 Nodes
128 Racks	10D	8192 Nodes
256 Racks	11D	16,384 Nodes
512 Racks	12D	32,768 Nodes

Table 2. *SGI Altix ICE can be used to build up to 12-dimensional hypercubes.*

4.3 Configuring for Fat Tree

The different backplanes available in the SGI Altix ICE 8400 IRU mean that Fat Tree topologies can be built to suit the needs of the applications they serve. The Eight-Node Bristled IRU Backplane can be used to build a Fat Tree cluster as shown in Figure 7.

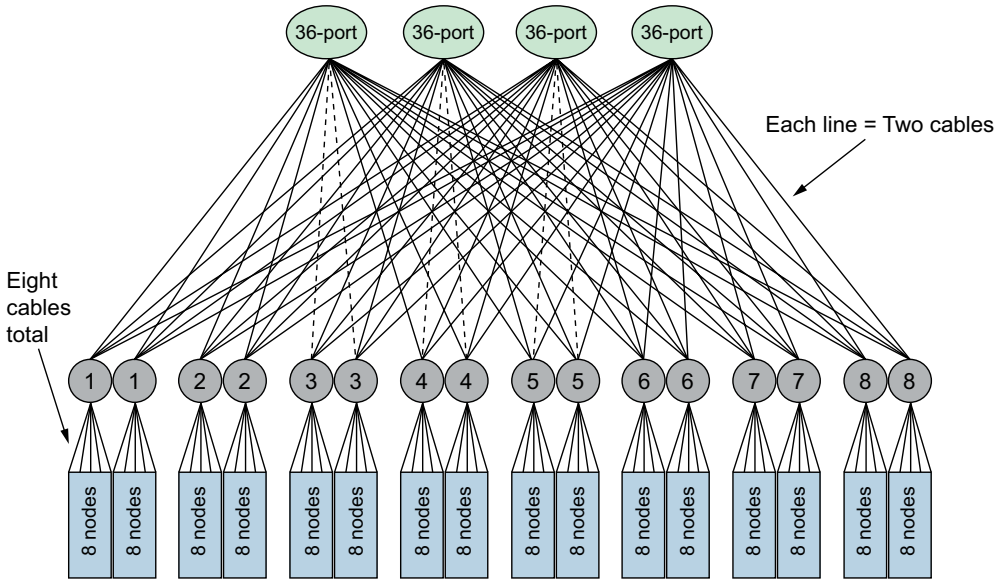

Figure 7. *A two-rack Fat Tree topology can be easily built with the Eight-Node Bristled IRU Backplane.*

As shown in Figure 8, the Sixteen-Node Bristled IRU Backplane can also be used to build a more cost-effective Fat Tree cluster.

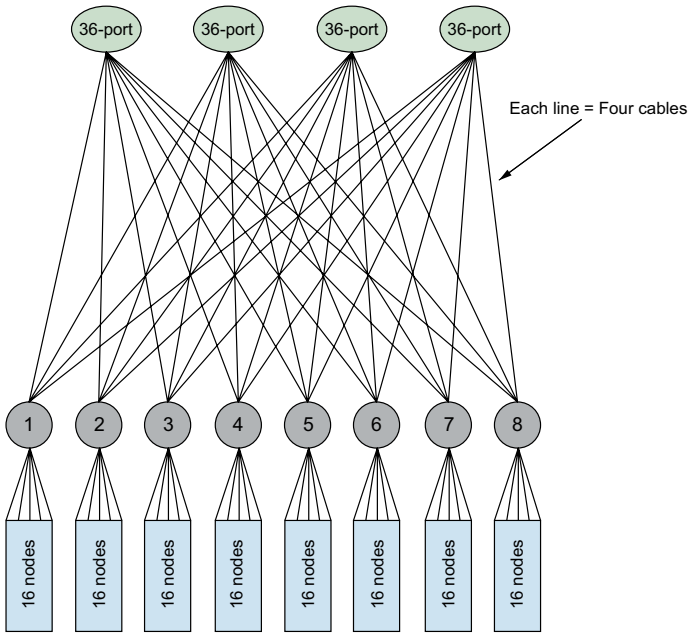

Figure 8. *The Sixteen-Node Bristled IRU Backplane can also be used to build a Fat Tree cluster.*

5.0 Topology Study Results

To evaluate the relative performance of various topologies, SGI has conducted extensive testing based on the SGI Altix ICE 8400 platform. Testing included a mixture of popular interconnect kernel benchmarks and applications. Tested topologies included:

- Simple Hypercube, single-rail and dual-rail
- Enhanced Hypercube, single-rail and dual-rail
- All-to-All, single-rail and dual-rail
- Fat Tree, single-rail

The system used for this testing was a two-rack, 128-node SGI Altix ICE 8400 system featuring SGI IP105 compute blades, each equipped with two six-core Intel Xeon Processor X5680 CPUs @ 3.33 GHz. Total memory for the configured cluster was 2.47TB.

Note: All of these results were conducted using SGI IP105 compute blades. The conclusions for the global interconnect-intensive benchmarks should remain valid for IP103 blades. However, relative to applications, the benefits of dual-rail configurations may be reduced when using SGI IP103 compute blades rather than SGI IP105 blades due to available PCIe bandwidth.

5.1 Interconnect Kernel Benchmarks

To evaluate topology differences, SGI ran a wide range of kernel benchmarks on different topologies. Kernel benchmarks tested included:

- HPCC (PTRANS, HPL, and MPI Random Access (GUPs), MPI FFT, and Maximum Ping Pong Latency, Minimum Ping Pong Bandwidth, Random Ring Latency, and Random Ring Bandwidth)
- Intel MPI Benchmarks (Large Message All to All and 8 Byte All Reduce)
- IMB Bisection Bandwidth (modified to utilize all-to-all links if present)

Selected samples from these tests are discussed in the sections that follow, along with a geometric mean across all of the kernel benchmarks. All test results are shown normalized to Standard Hypercube, Single-Rail performance.

IMB Bisection Bandwidth

The IMB Bisection Bandwidth test provides measured numbers for bisection bandwidth, modified to utilize all-all express links if they are present As shown in Figure 9, the results for this test produced a clear distinction between different topologies, showing a clear benefit for dual-rail topologies. Traditional bisection bandwidth metrics only stress a single dimensional link and only look at traffic across a single dimension. In contrast, this benchmark generates more random and orthogonal traffic, revealing the pronounced differences in bandwidth available from different topologies and multiple rails.

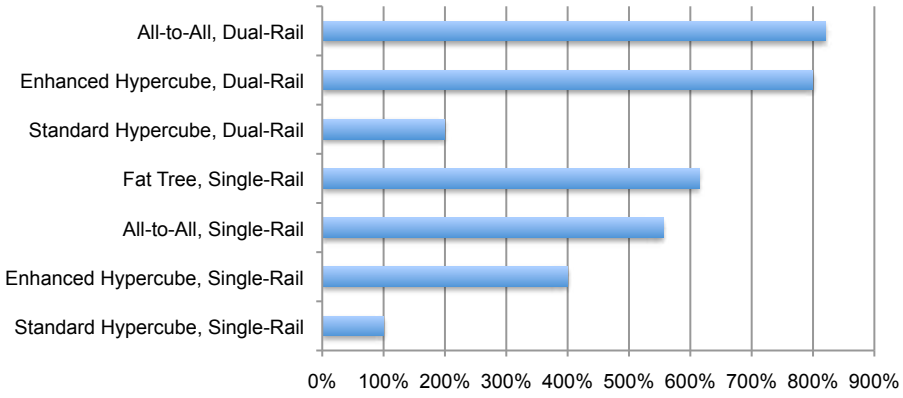

Figure 9. *The IMB Bisection Bandwidth Test shows a clear distinction between multiple topologies, and between single-rail and multi-rail deployments (normalized to Standard Hypercube, Single-Rail, larger is better).*

HPPC – PTRANS

The HPPC-PTRANS kernel benchmark produces variable traffic in multiple dimensions. Here too significant variability is shown between the multiple topologies as shown in Figure 10.

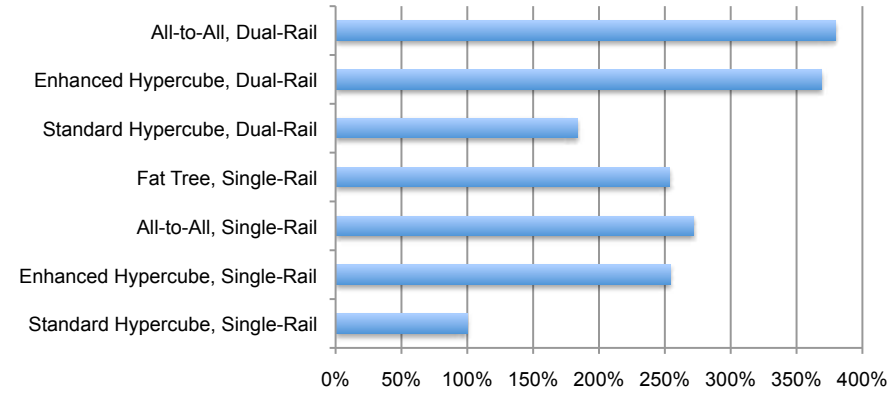

Figure 10. *The HPPC - PTRANS kernel benchmark demonstrates differences between topologies, and single- and multi-rail fabrics (normalized to Standard Hypercube, Single-Rail, larger is better).*

Geometric Mean Over All Interconnect Kernels

Figure 11 illustrates the geometric mean of all of the kernel benchmark tests run across the various topologies for clusters of 16, 32, 64, and 128 nodes, normalized to Standard Hypercube single-rail performance. The kernel benchmarks respond differently to the various topologies, depending on how sensitive they are to available global bandwidth. It is also clear that the differences between topologies become more pronounced as the cluster size grows. With clusters larger than 128 nodes, the scalability advantages of Standard and Enhanced Hypercube only increase.

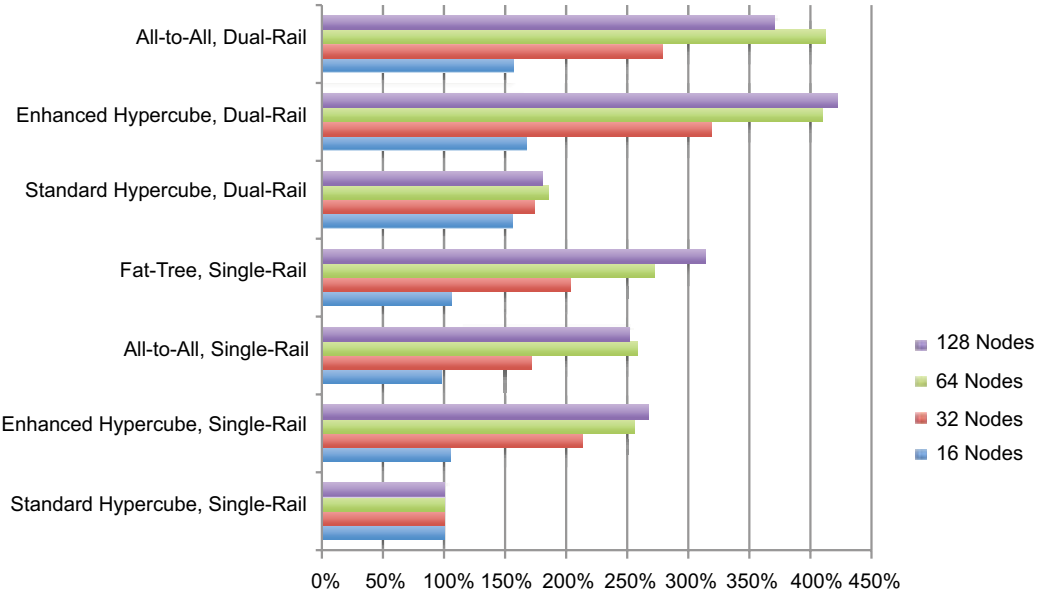

Figure 11. *The geometric mean across all conducted kernel benchmark tests shows differentiation between different InfiniBand topologies (shown normalized to Standard Hypercube, single-rail performance, larger is better).*

5.2 Application Benchmarks

In SGI's topology testing, application benchmarks were also run across various topologies, in both single-plane and dual-plane configurations. The application benchmarks run included:

- SPEC MPIL2007
- GAMESS (Standard and Large)
- LS-DYNA 3 Cars Model
- FLUENT - Truck 14M Case
- Cart3D
- WRF-Conus 12 KM
- LAMMPS - Scaled LJ, Scaled Rhodopsin

Selected samples from these tests are discussed in the sections that follow, along with a geometric mean across all of the application benchmarks.

WRF – Conus 12KM

As a weather modeling benchmark, WRF-Conus is not particularly sensitive to global bandwidth as shown in Figure 12. For example, single-rail Standard Hypercube provides similar performance to single-rail Fat Tree. However, using the SGI IP105 compute blade, benefit can be derived by splitting large point-to-point messages across multiple rails, resulting in roughly an 8-9% performance improvement for multi-rail topologies over single-rail topologies.

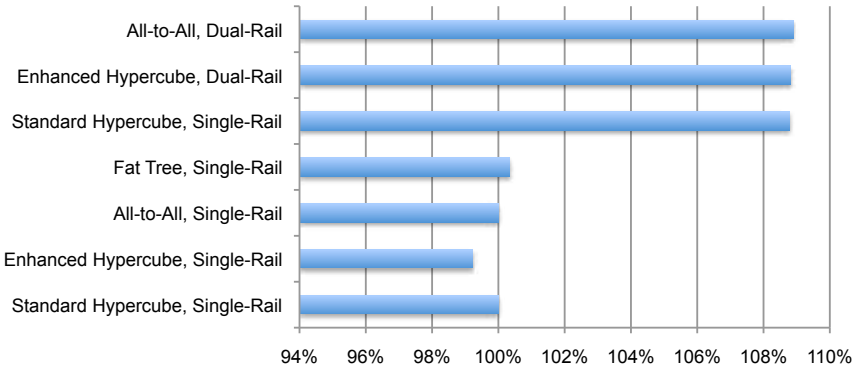

Figure 12. *With the Conus-WRF 12KM benchmark, dual-rail topologies provide an 8-9% advantage by distributing point-to-point messages across multiple rails (normalized to Standard Hypercube, Single-Rail, larger is better).*

FLUENT

As a computationally-intensive fluid dynamics solution, FLUENT provides consistent behavior across all topologies (Figure 13). Dual-rail Hypercube improves the bandwidth for point-to-point communications, and provides a substantial performance increase. Though not tested, dual-rail All-to-All and dual-rail Standard Hypercube are expected to deliver similar results to the tested dual-rail Enhanced Hypercube topology.

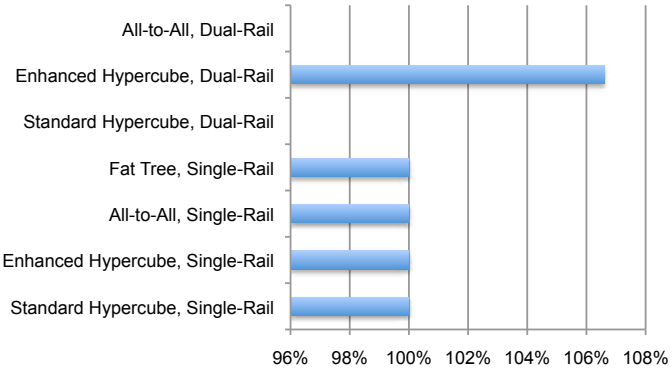

Figure 13. *FLUENT provides consistent results across all topologies (normalized to Standard Hypercube, Single-Rail, larger is better).*

LAMMPS – Scaled Rhodopsin

As a molecular dynamics code, LAMMPS is primarily composed of Fast Fourier Transforms (FFTs) that perform all-to-all communications. As a result, LAMMPS has a larger degree of global bandwidth sensitivity as shown in Figure 14.

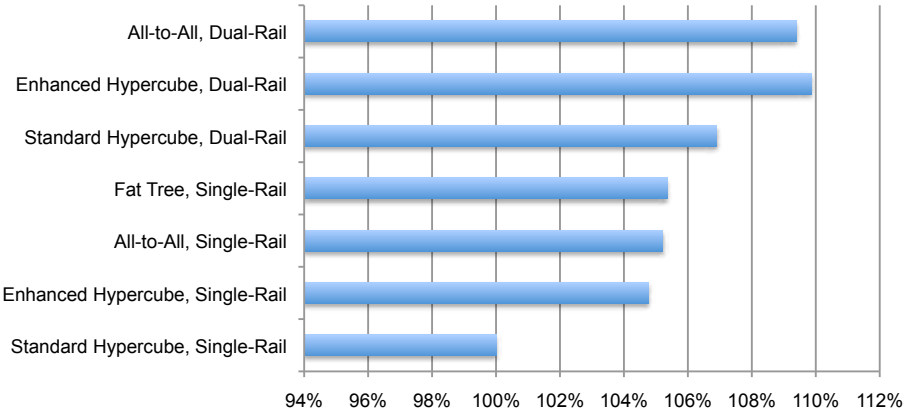

Figure 14. *The LAMMPS computational chemistry code demonstrates global bandwidth sensitivity.*

Geometric Mean Over All Applications

Figure 14 compares geometric means across all of the applications tested, for clusters of 16, 32, 64, and 128 nodes, normalized to single-rail Standard Hypercube performance. One of the clear conclusions is that the gains or differences for application benchmarks are much more modest when compared to the kernel benchmarks presented earlier in this document that are more globally bandwidth intensive. Stated another way, application performance is much less dependent on topology for performance. Given these smaller performance differences, topology choice may hinge on the simplest, or least expensive configuration to deploy.

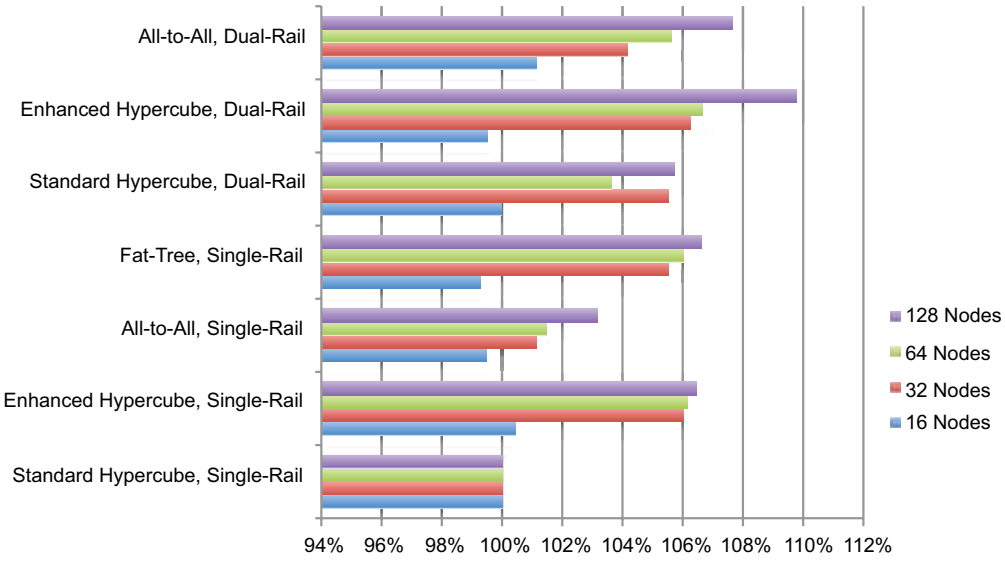

Figure 15. *Geometric means of tested applications across all topologies show less pronounced performance differences than kernel benchmarks, but still grow with cluster size (shown normalized to single-rail Standard Hypercube performance).*

As with the kernel benchmarks, it is clear that the differences in performance between topologies, and between single- and dual-rail deployments grow more pronounced with the size of the cluster. These differences are expected to expand as clusters grow beyond 128 nodes. Hypercube-based topologies will have a distinct advantage given their ability to match performance and scale more effectively than either Fat Tree or All-to-All topologies.

6.0 Conclusion

Effective InfiniBand topology requires system architecture designed with scalability in mind. The SGI Altix ICE system was purposely designed for InfiniBand networking, and together with multi-core Intel Xeon processors, the platform is capable of achieving industry-leading scalability for a broad range of technical computing applications. With a choice of supported InfiniBand topologies, the SGI Altix ICE system is ideal for deploying InfiniBand clusters ranging from a single 16-node IRU to hundreds of racks and many thousands of nodes.

Selecting an appropriate InfiniBand topology requires careful consideration of applications, algorithms, and data sets, along with likely needs for scalability into the future. In the absence of benchmark data, having some basic knowledge of the application characteristics may be enough to guide topology choices. Extensive testing done by SGI has shown that applications are generally less sensitive to topology than kernel benchmarks, but that differences in performance become more pronounced as clusters grow in size. When global interconnect bandwidth is important, Enhanced Hypercube dual-rail is the raw performance leader. For smaller single-rail topologies, Fat Tree is often the most economical choice. As clusters grow, hypercube topologies gain scalability, performance and cost advantages, avoiding the external switching and cabling that is required for Fat Tree and All-to-All topologies.

Having deployed some of the world's largest open systems InfiniBand networks and clusters, SGI has the experience and expertise to help organizations choose the right equipment and networking topology to meet their most challenging computational problems.

7.0 Appendix A: Blocking Ratios and Bisection Bandwidth

In choosing an InfiniBand topology, it is important to apply metrics objectively across multiple topologies. The tables in the sections that follow provide equivalent blocking factors and bisection bandwidth numbers for various topologies when implemented by SGI Altix ICE systems.

7.1 Standard Hypercube, Single-Rail

Table 3 provides blocking ratios and Table 4 provides bisection bandwidth for single-rail Standard Hypercubes built on SGI Altix ICE. As illustrated by shading in the tables,

- Quad links yields a 2:1 blocking factor and a 2000 MB/s bisection bandwidth per node
- Single links yields an 8:1 blocking factor and a 500 MB/s bisection bandwidth per node

	Equivalent Blocking Factor for Each Set of Dimensional Links											
Nodes	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D	12D
16	2:1											
32	2:1	8:1										
64	2:1	8:1	8:1									
128	2:1	8:1	8:1	8:1								
256	2:1	8:1	8:1	8:1	8:1							
512	2:1	8:1	8:1	8:1	8:1	8:1						
1024	2:1	8:1	8:1	8:1	8:1	8:1	8:1					
2048	2:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1				
4098	2:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1			
8192	2:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1		
16384	2:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	
32768	2:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1	8:1

Table 3. *Blocking ratios for single-rail Standard Hypercube clusters based on SGI Altix ICE.*

	Bisection bandwidth (MB/s) per Node for Each Set of Dimensional Links											
Nodes	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D	12D
16	2000											
32	2000	500										
64	2000	500	500									
128	2000	500	500	500								
256	2000	500	500	500	500							
512	2000	500	500	500	500	500						
1024	2000	500	500	500	500	500	500					
2048	2000	500	500	500	500	500	500	500				
4098	2000	500	500	500	500	500	500	500	500			
8192	2000	500	500	500	500	500	500	500	500	500		
16384	2000	500	500	500	500	500	500	500	500	500	500	
32768	2000	500	500	500	500	500	500	500	500	500	500	500

Table 4. *Bisection bandwidth for single-rail Standard Hypercube clusters based on SGI Altix ICE.*

7.2 Standard Hypercube, Dual-Rail

Table 5 provides blocking ratios and Table 6 provides bisection bandwidth for dual-rail Standard Hypercubes built on SGI Altix ICE. As illustrated by shading in the tables,

- Dual quad links yield a 1:1 blocking factor and a 4000 MB/s bisection bandwidth per node
- Dual single links yields a 4:1 blocking factor and a 1000 MB/s bisection bandwidth per node

	Equivalent Blocking Factor for Each Set of Dimensional Links											
Nodes	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D	12D
16	1:1											
32	1:1	4:1										
64	1:1	4:1	4:1									
128	1:1	4:1	4:1	4:1								
256	1:1	4:1	4:1	4:1	4:1							
512	1:1	4:1	4:1	4:1	4:1	4:1						
1024	1:1	4:1	4:1	4:1	4:1	4:1	4:1					
2048	1:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1				
4098	1:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1			
8192	1:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1		
16384	1:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	
32768	1:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1

Table 5. *Blocking ratios for dual-rail Standard Hypercube clusters based on SGI Altix ICE.*

	Bisection bandwidth (MB/s) per Node for Each Set of Dimensional Links											
Nodes	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D	12D
16	4000											
32	4000	1000										
64	4000	1000	1000									
128	4000	1000	1000	1000								
256	4000	1000	1000	1000	1000							
512	4000	1000	1000	1000	1000	1000						
1024	4000	1000	1000	1000	1000	1000	1000					
2048	4000	1000	1000	1000	1000	1000	1000	1000				
4098	4000	1000	1000	1000	1000	1000	1000	1000	1000			
8192	4000	1000	1000	1000	1000	1000	1000	1000	1000	1000		
16384	4000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	
32768	4000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000

Table 6. *Bisection bandwidth for dual-rail Standard Hypercube clusters based on SGI Altix ICE.*

7.3 Enhanced Hypercube, Single-Rail

Table 7 provides blocking ratios and Table 8 provides bisection bandwidth for single-rail Enhanced Hypercubes deployed on SGI Altix ICE. As illustrated by shading in the tables,

- Octal linking yields a 1:1 blocking factor and 4000 MB/s bisection bandwidth per node
- Quad linking yields a 2:1 blocking factor and 2000 MB/s bisection bandwidth per node
- Dual linking yields a 4:1 blocking factor and 1000 MB/s bisection bandwidth per node
- Single linking yields an 8:1 blocking factor and 500 MB/s bisection bandwidth per node

	Equivalent Blocking Factor of Each Set of Dimensional Links											
Nodes	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D	12D
16	1:1											
32	1:1	1:1										
64	1:1	1:1	1:1									
128	1:1	1:1	2:1	2:1								
256	1:1	2:1	2:1	2:1	2:1							
512	2:1	2:1	2:1	2:1	2:1	2:1						
1024	2:1	2:1	2:1	2:1	2:1	4:1	4:1					
2048	2:1	2:1	2:1	2:1	4:1	4:1	4:1	4:1				
4098	2:1	2:1	2:1	4:1	4:1	4:1	4:1	4:1	4:1			
8192	2:1	2:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1		
16384	2:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4.1	
32768	2:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	4:1	8:1	8:1

Table 7. *Blocking ratios for single-rail Enhanced Hypercube clusters based on SGI Altix ICE.*

	Bisection bandwidth (MB/s) per Node for Each Set of Dimensional Links											
Nodes	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D	12D
16	4000											
32	4000	4000										
64	4000	4000	4000									
128	4000	4000	2000	2000								
256	4000	2000	2000	2000	2000							
512	2000	2000	2000	2000	2000	2000						
1024	2000	2000	2000	2000	2000	1000	1000					
2048	2000	2000	2000	2000	1000	1000	1000	1000				
4098	2000	2000	2000	1000	1000	1000	1000	1000	1000			
8192	2000	2000	1000	1000	1000	1000	1000	1000	1000	1000		
16384	2000	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	
32768	2000	1000	1000	1000	1000	1000	1000	1000	1000	1000	500	500

Table 8. *Bisection Bandwidth for single-rail Enhanced Hypercube clusters based on SG Altix ICE.*

7.4 Enhanced Hypercube, Dual-Rail

Table 9 provides blocking ratios and Table 10 provides bisection bandwidth for dual-rail Enhanced Hypercubes deployed on SGI Altix ICE. As illustrated by shading in the tables,

- Dual octal linking yields a 1:2 blocking factor and 8000 MB/s bisection bandwidth per node
- Dual quad linking yields a 1:1 blocking factor and 4000 MB/s bisection bandwidth per node
- Dual dual linking yields a 2:1 blocking factor and 2000 MB/s bisection bandwidth per node
- Dual single linking yields a 4:1 blocking factor and 1000 MB/s bisection bandwidth per node

	Equivalent Blocking Factor for Each Set of Dimensional Links											
Nodes	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D	12D
16	1:2											
32	1:2	1:2										
64	1:2	1:2	1:2									
128	1:2	1:2	1:1	1:1								
256	1:2	1:1	1:1	1:1	1:1							
512	1:1	1:1	1:1	1:1	1:1	1:1						
1024	1:1	1:1	1:1	1:1	1:1	2:1	2:1					
2048	1:1	1:1	1:1	1:1	2:1	2:1	2:1	2:1				
4098	1:1	1:1	1:1	2:1	2:1	2:1	2:1	2:1	2:1			
8192	1:1	1:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1		
16384	1:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1	
32768	1:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1	2:1	4:1	4:1

Table 9. *Blocking ratios for dual-rail Enhanced Hypercube clusters deployed on SGI Altix ICE.*

	Bisection bandwidth (MB/s) per Node for Each Set of Dimensional Links											
Nodes	1D	2D	3D	4D	5D	6D	7D	8D	9D	10D	11D	12D
16	8000											
32	8000	8000										
64	8000	8000	8000									
128	8000	8000	4000	4000								
256	8000	4000	4000	4000	4000							
512	4000	4000	4000	4000	4000	4000						
1024	4000	4000	4000	4000	4000	2000	2000					
2048	4000	4000	4000	4000	2000	2000	2000	2000				
4098	4000	4000	4000	2000	2000	2000	2000	2000	2000			
8192	4000	4000	2000	2000	2000	2000	2000	2000	2000	2000		
16384	4000	2000	2000	2000	2000	2000	2000	2000	2000	2000	2000	
32768	4000	2000	2000	2000	2000	2000	2000	2000	2000	2000	1000	1000

Table 10. *Bisection bandwidth for dual-rail Enhanced Hypercube clusters deployed on SGI Altix ICE.*

7.5 All-to-All, Single-Rail

Table 11 provides blocking ratios and bisection bandwidth for single rail All-to-All topologies based on SGI Altix ICE. Blocking ratio numbers are given based on node-to-node communications.

Nodes	Blocking Ratio	Bisection Bandwidth (MB/s per Node)
16	1:1	4000
32	2:1	4000
48	2:1	6000
64	4:1	4000
80	4:1	5200
96	8:1	3000
112	8:1	3700
128	8:1	4000

Table 11. *Blocking ratios and bisection bandwidth for single-rail All-to-All clusters deployed on SGI Altix ICE.*

7.6 All-to-All, Dual-Rail

Table 12 provides blocking ratios and bisection bandwidth for single rail All-to-All topologies based on SGI Altix ICE.

Nodes	Blocking Ratio	Bisection Bandwidth (MB/s per Node)
16	1:2	8000
32	1:1	8000
48	1:1	12000
64	2:1	8000
80	2:1	10400
96	4:1	6000
112	4:1	7400
128	4:1	8000

Table 12. *Blocking ratios and bisection bandwidth for dual-rail All-to-All dual-rail deployed on SGI Altix ICE.*

Corporate Headquarters
46600 Landing Parkway
Fremont, CA 94538
tel 510.933.8300
fax 408.321.0293
www.sgi.com

Global Sales and Support
North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.927.8000
Asia Pacific +61 2.9448.1463