HP-SEE

HP-SEE receives EC support through FP7
under the "Research Infrastructures" action

HP-SEE TRAINING

30 NOVEMBER 2010, SOFIA, IICT-BAS

SPONSORED BY:

Future HPC Architectures

Robert Uebelmesser(rue@sgi.com)
HPC Director, SGI EMEA

Topics

- SGI Customers, Target Markets, and Product Focus
- A Key Challenge for Future HPC Systems – The System Interconnect
  - Applications Analysis
  - The SGI ICE Approach
  - The SGI UV Approach (solves the large memory problem at the same time)
- A Futuristic Data Center – SGI ICE Cube
- The Structure of large HPC Data Centers in Europe

sgi

SGI

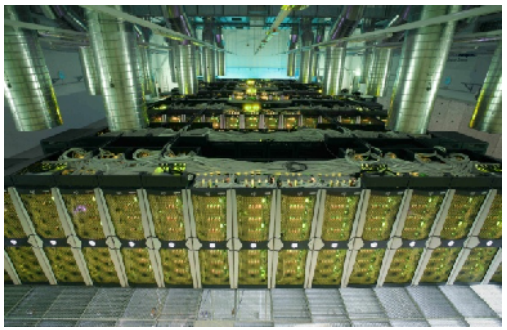
since May 2009

About SGI

- **Silicon Graphics Incorporated** was
  - A leading supplier of high-end compute, storage and visualization solutions for the technical and scientific market
- **Rackable Systems Inc.** was
  - the dominant supplier of servers to the Internet and cloud computing market (YouTube, Amazon, Microsoft, Nasdaq…..)

sgi

SGI Key Target Markets

HPC 	The „Old SGI“ Markets	**Media** 
Clouds 	The „Rackable“ Markets	**Internet** 

Experts @ Scale

Microsoft

NASDAQ

Amazon

Conoco Philips

LRZ

NASA

sgi

And many more…

Defense Intelligence	Manufacturing	National Labs	Digital Media
Raytheon, Lockheed Martin, Northrop Grumman, Classified	LG, Motorola, Volkswagen, Boeing	lrz, NASA, apac, Lawrence Livermore National Laboratory	National Geographic Channel, NBA, Disney, YouTube, PlayStation
Energy	**Internet - Cloud**	**Research**	**And More**
BP, Exxon Mobil, Total, Schlumberger, Marathon Oil Corporation	amazon, facebook, Microsoft, craigslist, shopzilla, myYearbook	NOAA, Meteo France, Yale, UCSC, BCM Baylor College of Medicine, National Cancer Institute	Sikorsky A United Technologies Company, AstraZeneca, Tata, Ford, 3M, Toyota, P&G

sgi

SGI Focus

A Trusted Leader in Technical Computing

IDC
Analyze the Future

"… As we enter the next generation, moving from petaflops to exaflops, ***SGI has emerged a trusted leader in technical computing.*** *Supporting some of the industry's most mission critical and large-scale computing applications."*

-Earl Joseph, program vice president, high-performance systems

sgi

SGI 5 Core Platforms

Scale Out

Scale Up

Storage

Datacenter Solutions

Software

Linux
Windows
Management
File Systems
Performance
Tools

sgi

SGI Compute Strategy

Leading solutions from scalable entry to hyper-scale and cloud

Large-scale Datacenter

CR FND XE

Cloud Inspired
Hyper-scale
Eco-Logical
Density
BTO

Shared Memory

UV 1000

UV 100

UV 10

Large Memory
Big Data
Fast I/O

Scalable Entry

O3 X2

Origin® 400

Office friendly
Self-contained
Scalable
Low IT needs

High Performance InfiniBand

XE ICE

Capability
Capacity
Cost-Optimized
Multi-Topology
Choice

sgi

SGI Storage Strategy

Leading solutions from cost to scalability and performance

Cloud Systems

IS1000

IS2000

Short-lived data
Cost optimized
Redundancy
Software RAID

RAID Systems

Entry Level (IS220/5000)
Price/Performance Leader
Enterprise RAID (IS4000)
Balanced Price/Performance
HPC RAID (IS6000/15000)
Ultimate Performance/Throughput

Integrated Storage Servers

NAS 50/100

IS3500

File serving
App Appliance
Cost or Performance optimized

Persistent Data

COPAN™

Spectra Logic

Long life data
Disk or Tape
Large Capacity
Eco-logical
High Density

Software: File systems (Lustre, CXFS™, XFS®), DMF, LiveArc™

sgi

Rackable / CloudRack

1. **Industry's Most Flexible & Configurable Platform**
   Supports low/high wattage Intel and AMD through SSD
2. **Built-to-Order**
   Configure the platform based on the customer's work load
3. **Datacenter Optimized**
   Cooling, power, layout and facility costs are top of mind
4. **FLOPS per SQ per Watt Optimized**
   High density and energy efficient are pre-requisites for scale
5. **Cloud Inspired (public and private)**
   Amazon EC2/S3, eBay, BT, Microsoft, Intuit, Shopzilla, NSA

sgi

(Sample) Rackable Customer Successes

Cloud e-Commerce

Delivering the world's largest cloud

Federal

Advanced Simulation for National Nuclear Simulation Agency

Commercial

Mission Critical Support, Powers Applications and Content Delivery

Cloud Storage

Powers Cloud Storage with SGI InfiniteStorage

sgi

Altix ICE

Altix ICE 8400

1. **World's fastest distributed memory computer**
   Base on SPECmpil. Up to dual IB channels per node.

2. **Scalable**
   Supports up to 131,072 nodes, 1 Million + Cores

3. **Open**
   Runs Standard Linux, Intel Xeon 5600 or AMD Opteron 6100 CPUs (a both strategy)

4. **New Topologies**
   Hypercube, enhanced hypercube, all-to-all, fat-tree

sgi

Altix UV

Altix UV 1000

1. **World's fastest shared memory computer**
   Base on SPECint and SPECfp
2. **Scalable**
   Single system image up to 2048 cores and 16TB memory
3. **Open**
   Runs Standard Linux, Intel Xeon 7500 Processors
4. **New Markets**
   HPC, Large Databases, Scalable I/O, RISC replacement

sgi

(Sample) Altix® UV Customer Successes

Education

東北大学
TOHOKU UNIVERSITY

Conducting research on flow science that protects the global environment.

Research

COSMOS CONSORTIUM
UK National Cosmology Supercomputer

Accelerating space research headed by Professor Stephen Hawking

Life Sciences

Focused on chemical dynamics, bioinformatics and computational biology, and biomedical imaging

Life Sciences

ICR
The Institute of Cancer Research

Processing hundreds of terabytes of data for the advancement of the study of cancer

sgi

COPAN

"MAID" Technology – Massive Array of Idle Disks

COPAN 400M

1. **Long-life persistent data storage**
   Disk is better than tape
2. **Eco-logical**
   High density (up to 3x the capacity per sq ft)
   Energy Efficient (up to 10x the power savings)
3. **Open**
   Runs Linux, Industry standard VTL/D2D packages, and uses standard SATA technology
4. **Wide Appeal**
   Every customer needs one !
5. **Best in Market**

Modular Data Centers

1. **Self-contained datacenter**
   Power distribution, cooling, safety
2. **Eco-logical**
   Achieving PUEs of 1.1 or better
3. **Eco-nomical**
   1/5 the cost of a traditional datacenter
4. **Simple and easy to deploy**
   Live in 5 days
5. **Universal**
   Support for SGI, DELL, HP, SUN, IBM, CISCO, EMC and others

sgi

Topics

- SGI Customers, Target Markets, and Product Focus
- A Key Challenge for Future HPC Systems – The System Interconnect
  - Applications Analysis
  - The SGI ICE Approach
  - The SGI UV Approach (solves the large memory problem at the same time)
- A Futuristic Data Center – SGI ICE Cube
- The Structure of large HPC Data Centers in Europe

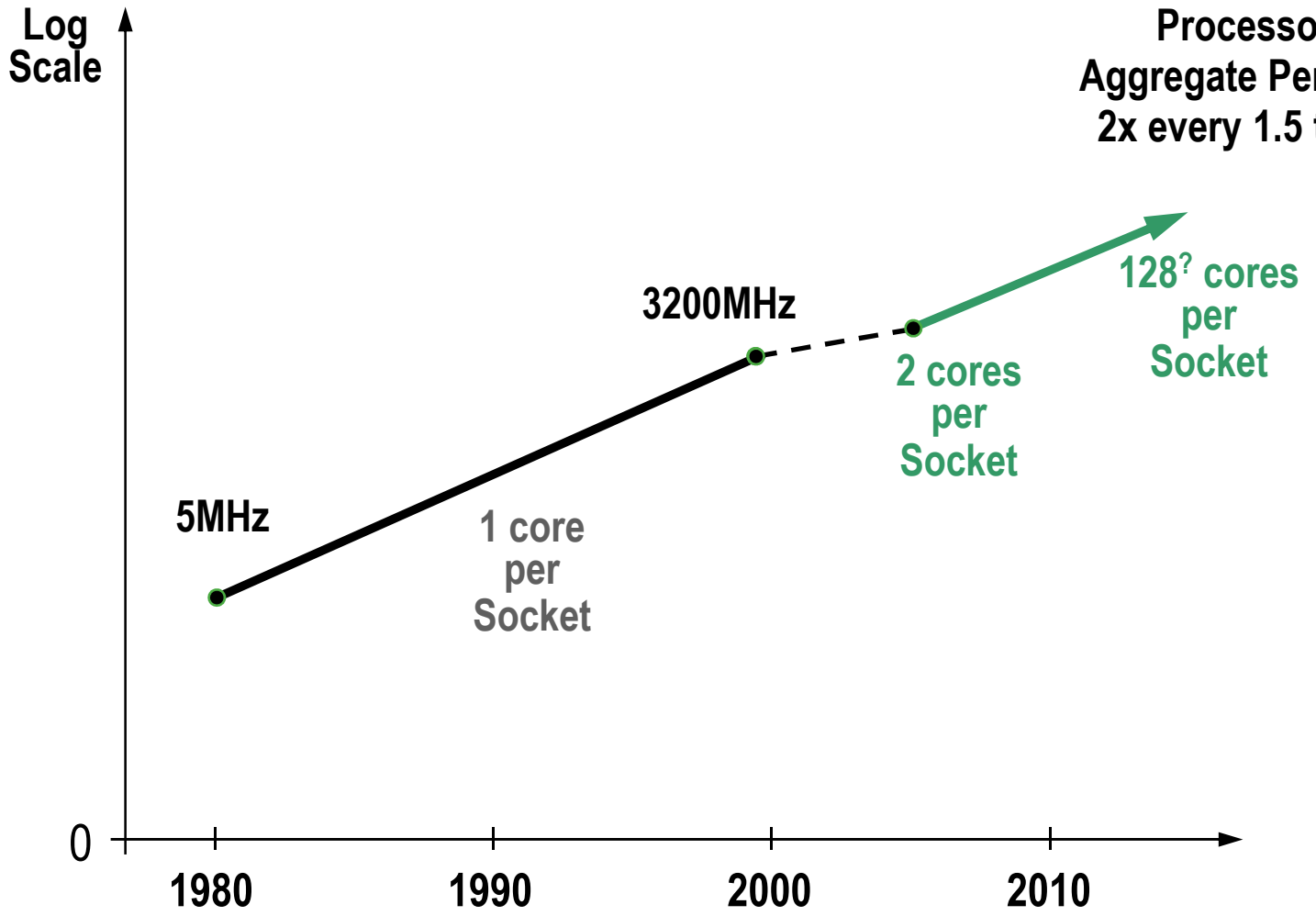
sgi

(Some of the) Big Challenges for Big Next Generation HPC Systems

- Architectural Challenges
  - Many core processors
  - Very large memories
  - System interconnects
  - Accelerators
- Physical Challenges
  - Power
  - Floorspace
  - Cooling
- Software Challenges
  - OS
  - Languages
  - Applications

sgi

The HPC Problem: from clock speed to Multicore



sgi

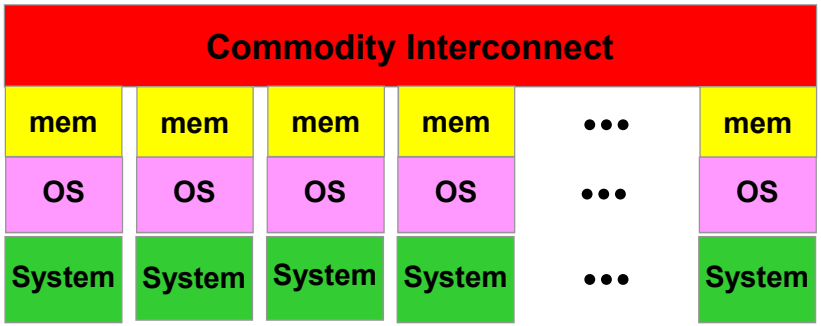
(Some of the) Big Challenges for Big Next Generation HPC Systems

- Architectural Challenges
  - Many core processors
  - Very large memories
  - System interconnects
  - Accelerators
- Physical Challenges
  - Power
  - Floorspace
  - Cooling
- Software Challenges
  - OS
  - Languages
  - Applications

sgi

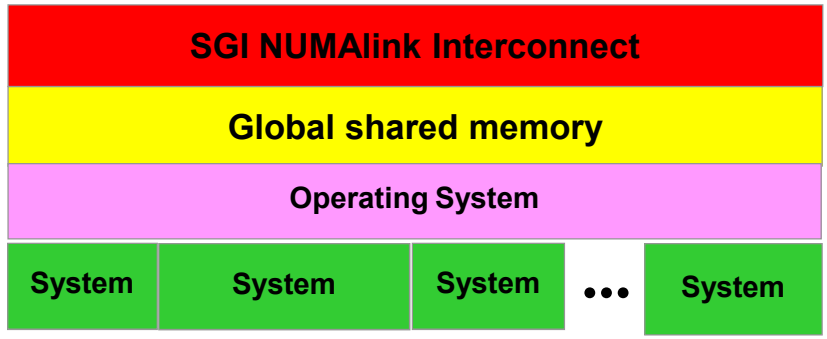
Supercomputer System Approaches

Integrated Cluster Systems



± Each system has own memory and OS
- Node bandwidth and latency issues
– More network interfaces
– Lower application efficiency
+ lower hardware cost
+ Heterogeneity
+ Node autonomy
+ Increased reliability

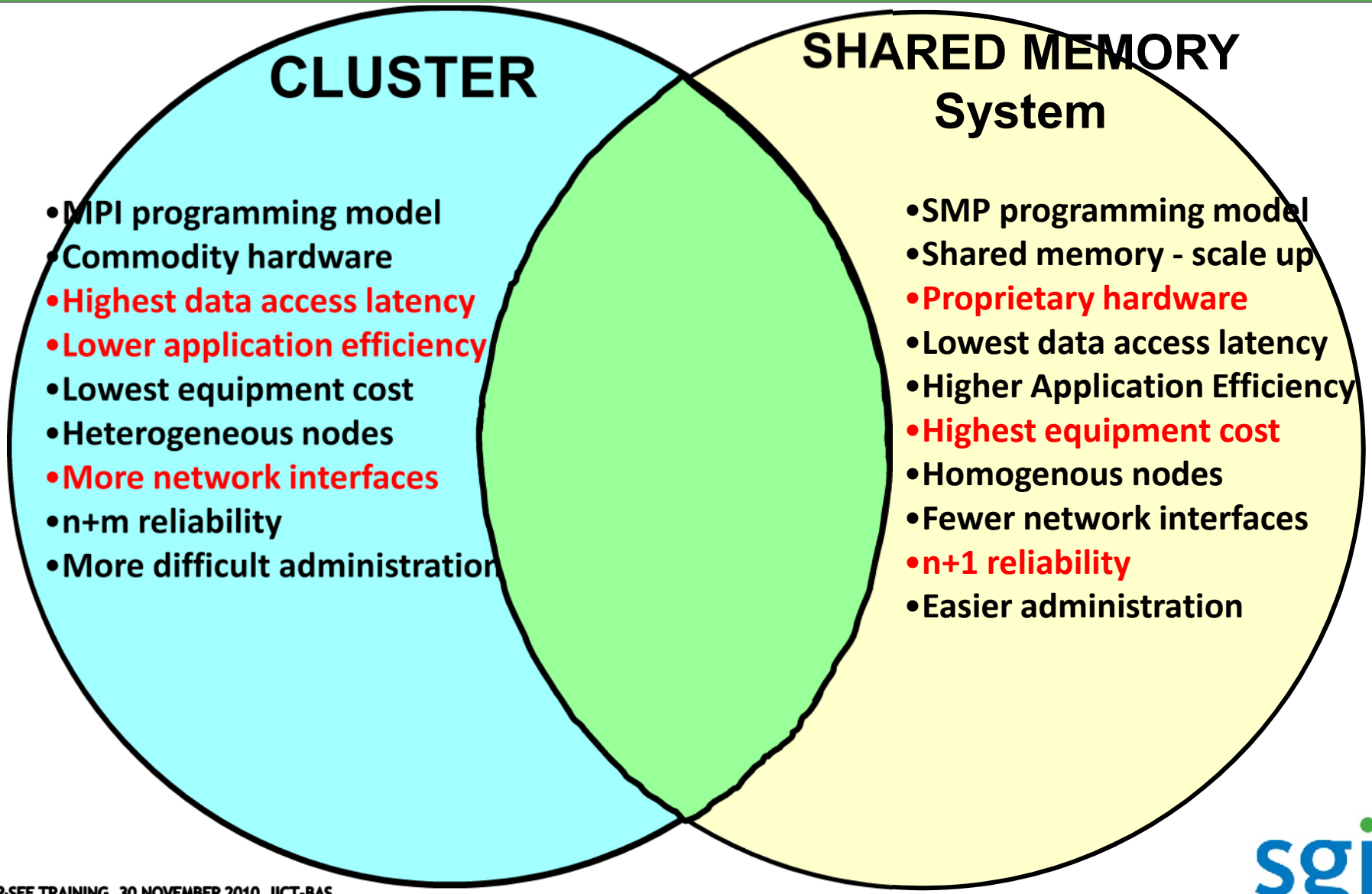
Globally Shared Memory Systems



+ All nodes operate on one large shared memory space
+ Eliminates data passing between nodes
+ Big data sets fit in memory
+ Less memory per node required
+ Higher Application Efficiency
+ Easier to Deploy and Administer
– More expensive hardware
– Considered less scalable

sgi

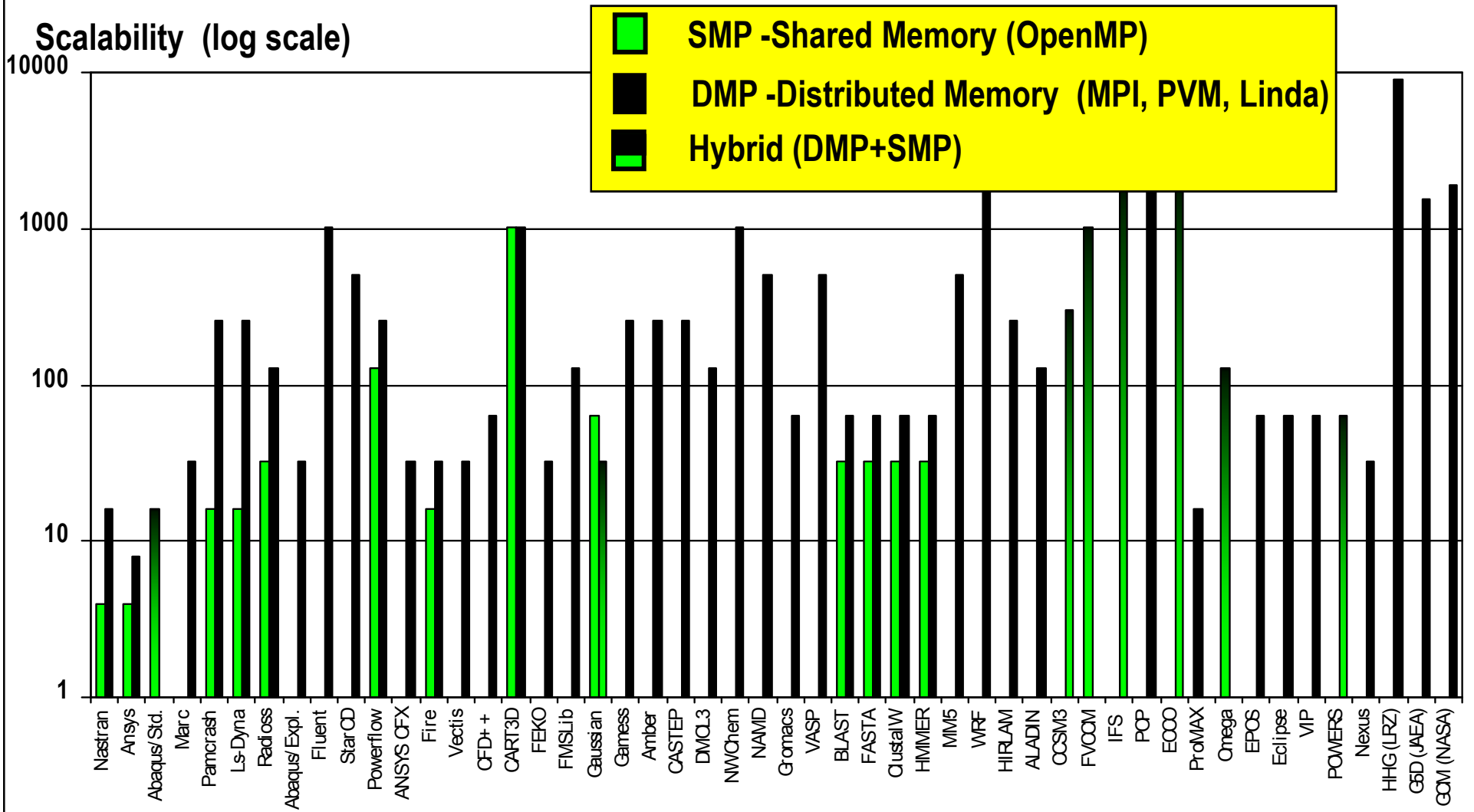
MPI and SMP Programming Models



sgi

Key Applications

Analysis of Scalability and Implementation



Scalability at 50% efficiency

Hybrid programming model is growing

sgi

SGI System Architecture and Performance Model

SGI Application Performance Model

$$T_{total} = T_{Core} + T_{mem\_lat} + T_{mem\_BW} + T_{comm\_lat} + T_{comm\_BW} + T_{IO}$$

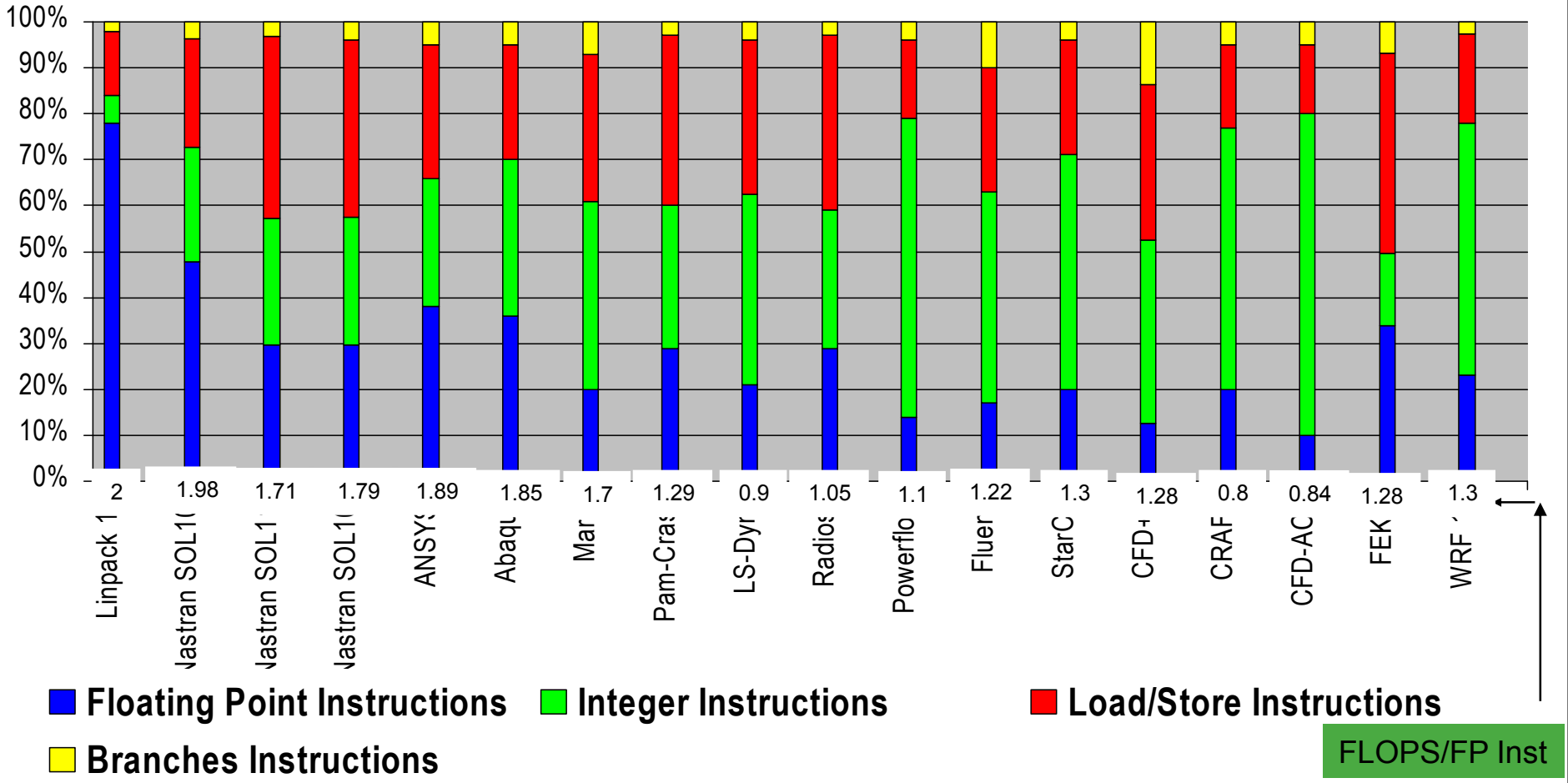
-Use SGI PerfSuite to estimate T_{Core}, T_{mem_lat}, T_{mem_BW}

-Use SGI MPInside to estimate: T_{comm_lat}, T_{comm_BW}

sgi

Key Applications on Altix 4700

Instruction Mix (w/o NOPs)

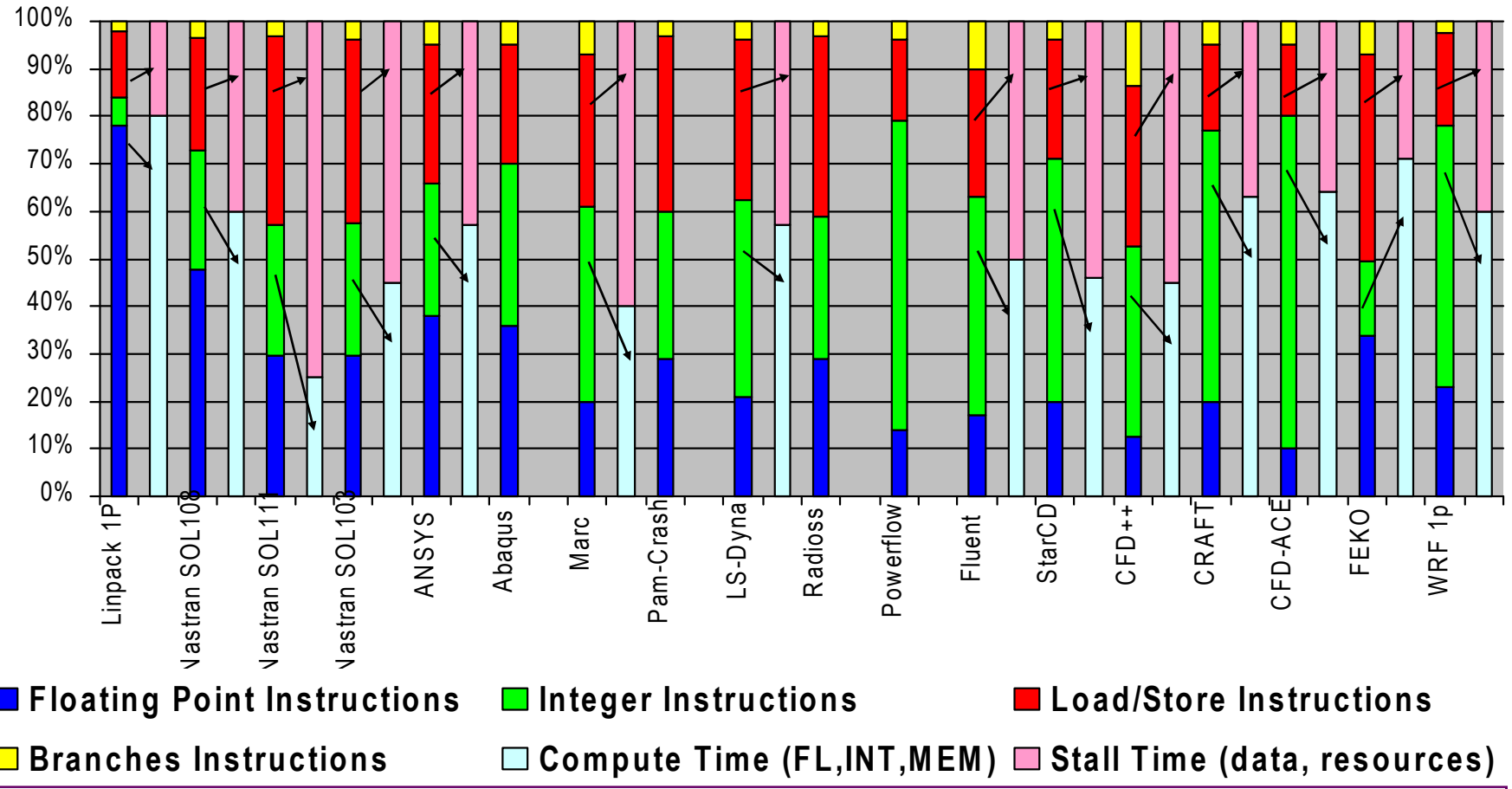


- Applications have in average of 23% FP and 28% memory instructions -
Computational intensity (flops/word) Linpack : 7 Applications: 0.8

sgi

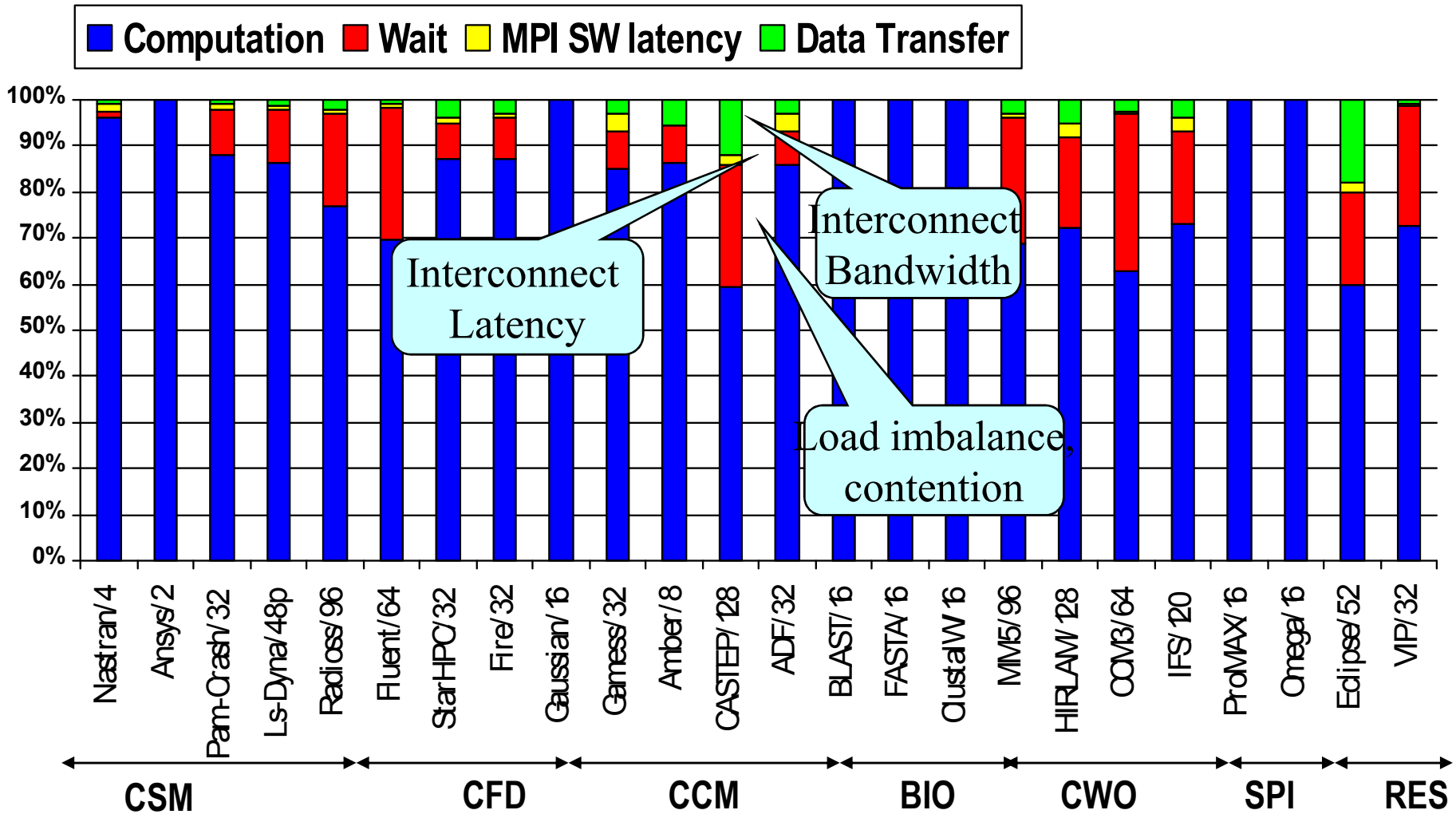
Key Applications on Altix 4700

Memory access vs. Stall time (w/o NOPs)



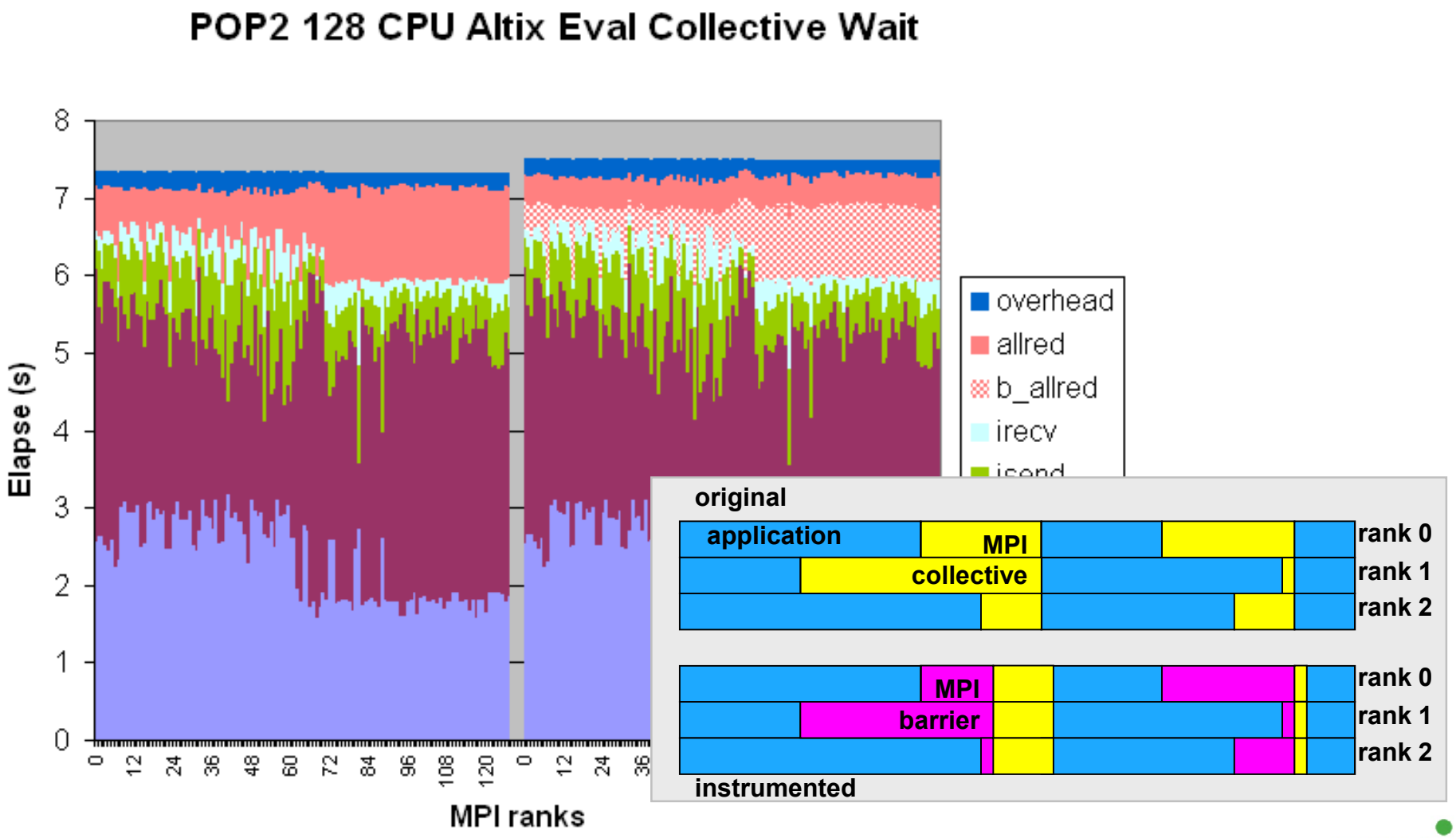
- ***Compute*** **means CPU is executing Instructions**
- ***Stalls*** **means CPU is waiting for data from memory or for other resources**

Communication vs. Computation Ratio
measured with MPInside



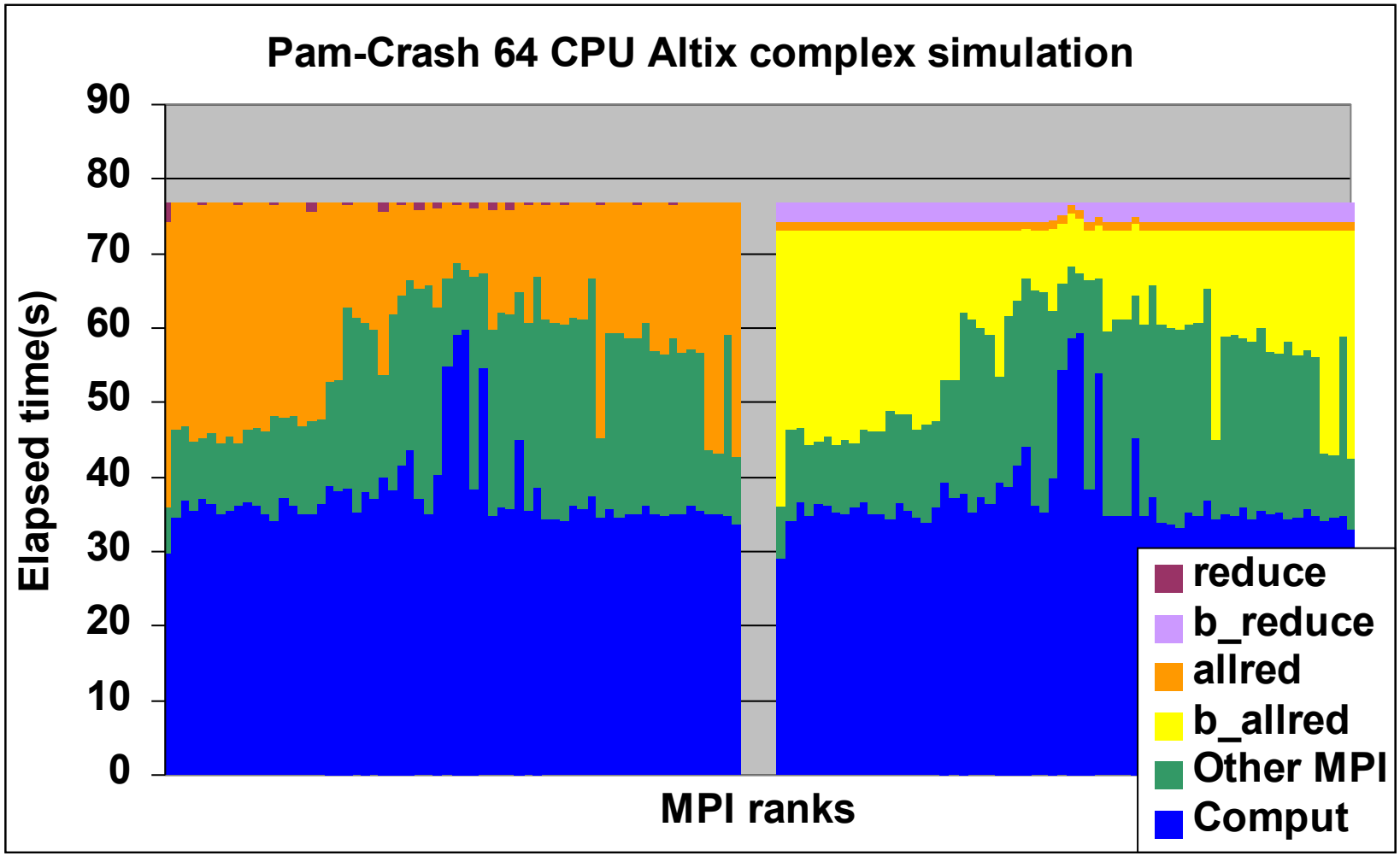
MPInside:

Evaluating collective operation synchronization time



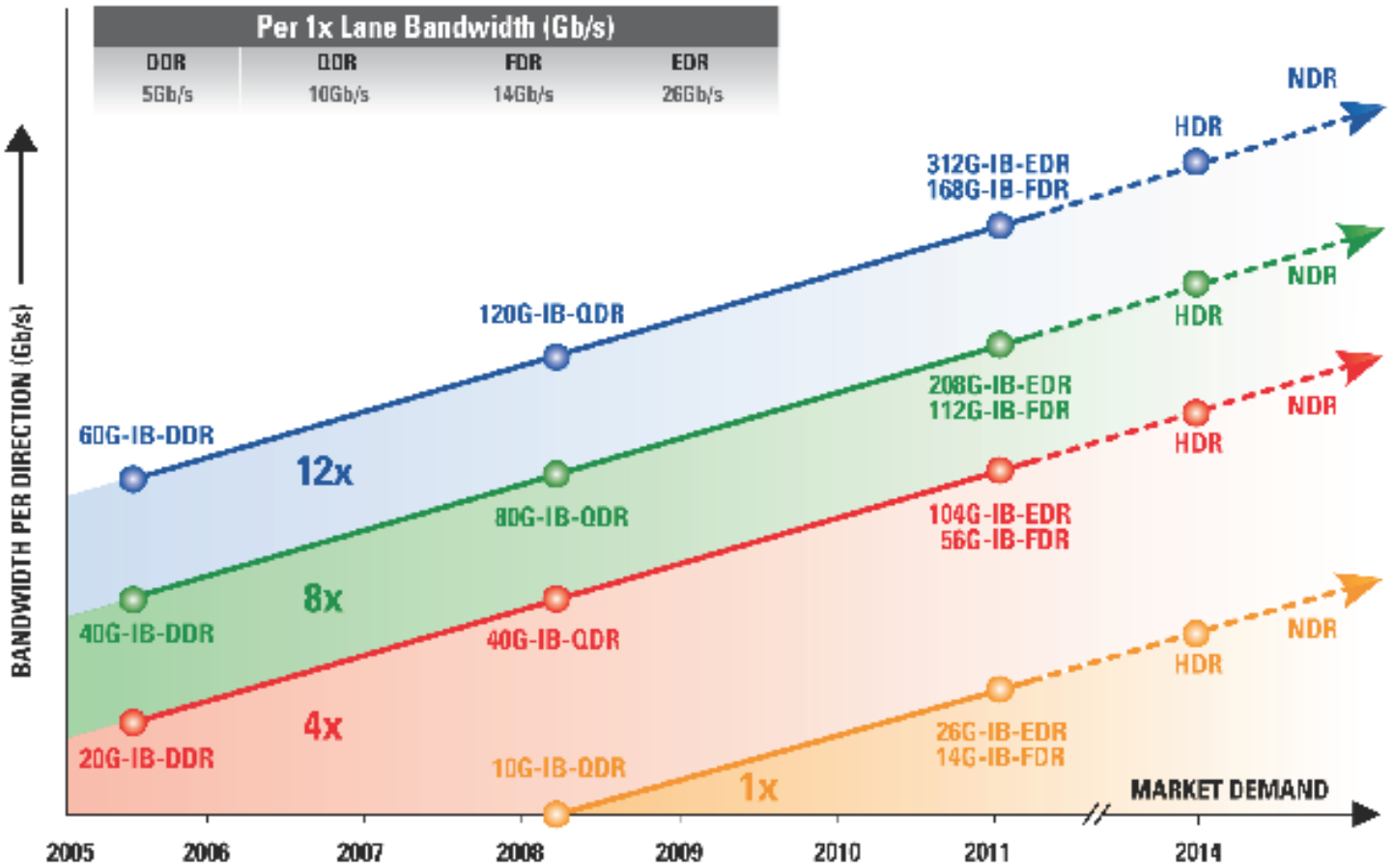
MPInside:

Evaluating collective operation synchronization time



sgi

InfiniBand Performance Roadmap – FDR/EDR



SGI Supercomputer Lines

SGI ICE 8400

Highly Integrated Cluster System

SGI Altix® UV

Partitioned Globally Addressable Memory System

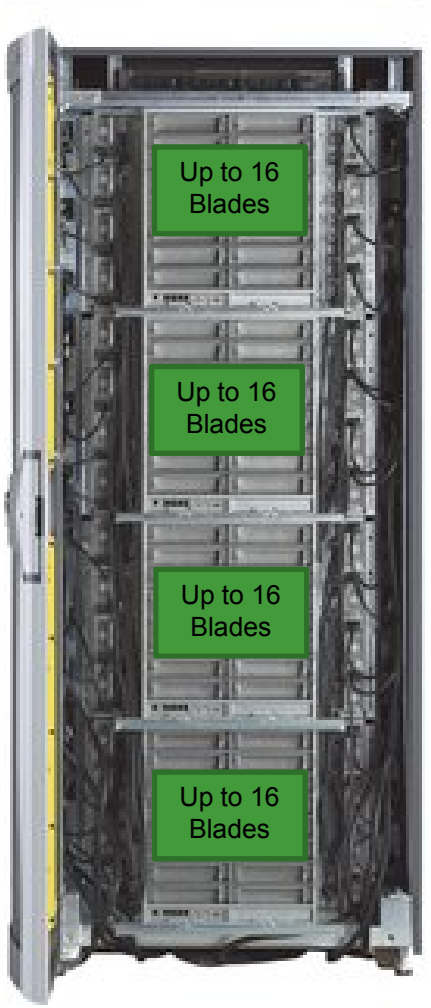
SGI Altix ICE 8400

SGI Altix ICE 8400

- Blade-based architecture
- AMD and Intel based processor blades
- Diskless blades operation
- Integrated Management network
- Hierarchical System Management
- single-plane or dual-plane 4xQDR Infiniband interconnect
- SGI enhanced Hypercube, Fat Tree or other network topologies
- Integrated switch topology simplifies scaling from 32 to 65,536 nodes (1024 racks)
- up to 128 processor sockets per rack
- 4 or more Dimms per socket
- Optional 2.5” SSD, HD for local storage

SGI Altix ICE 8400

Designed for High-Performance Computing



Breakthrough Performance Density:
Up to 128 sockets per Rack

SGI® Altix® ICE Compute Blade
Up to 12-Core, 96GB, 2-IB

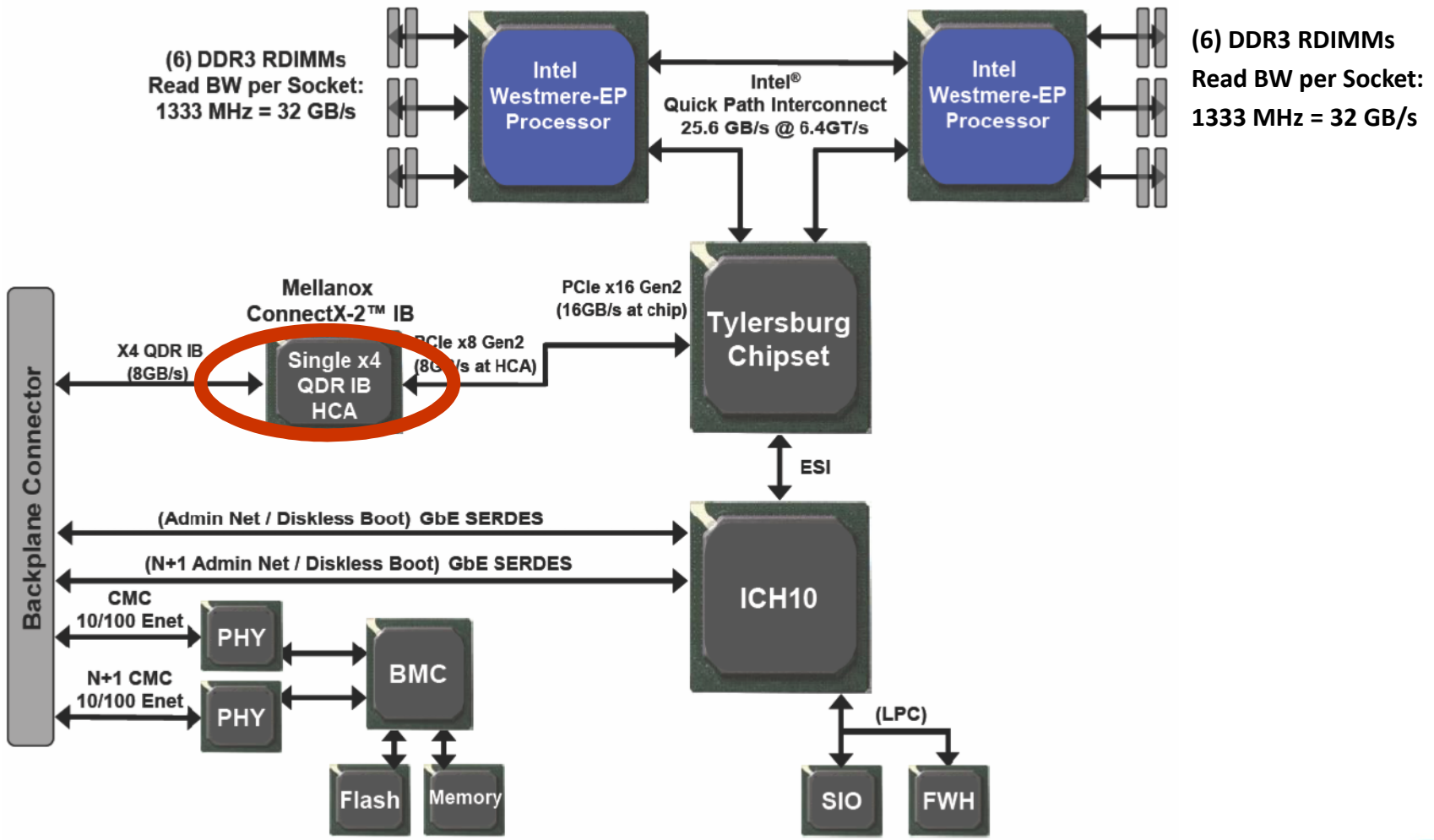
SGI® Altix® ICE Compute Blade
Up to 24-Core, 128GB, 2-IB

Altix ICE Rack:

- 42U rack (30” W x 40” D)
- 4 Cable-free blade enclosures, each with up to 16 2-Socket nodes
- Up to 128 DP AMD Opteron or Intel® Xeon® sockets
- Single-plane or Dual-plane IB 4x QDR interconnect
- Minimal switch topology simplifies scaling to 1000s of nodes

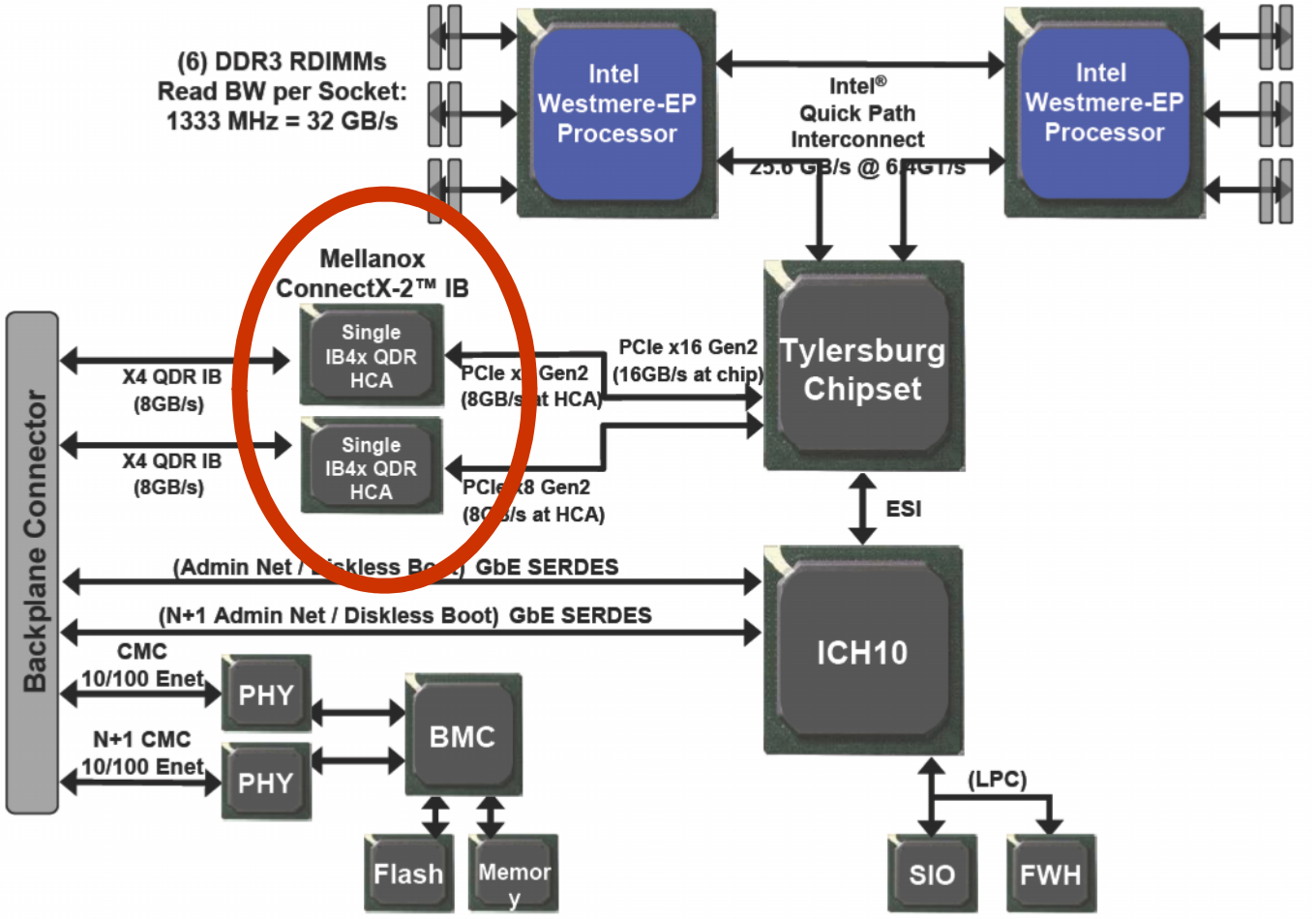
SGI ICE Blade for Intel Westmere-EP (Type 1)

IP101: One Single Port QDR HCA



SGI ICE Blade for Intel Westmere-EP (Type 2)

IP-105: Two Single Port QDR HCAs (Dual Plane)

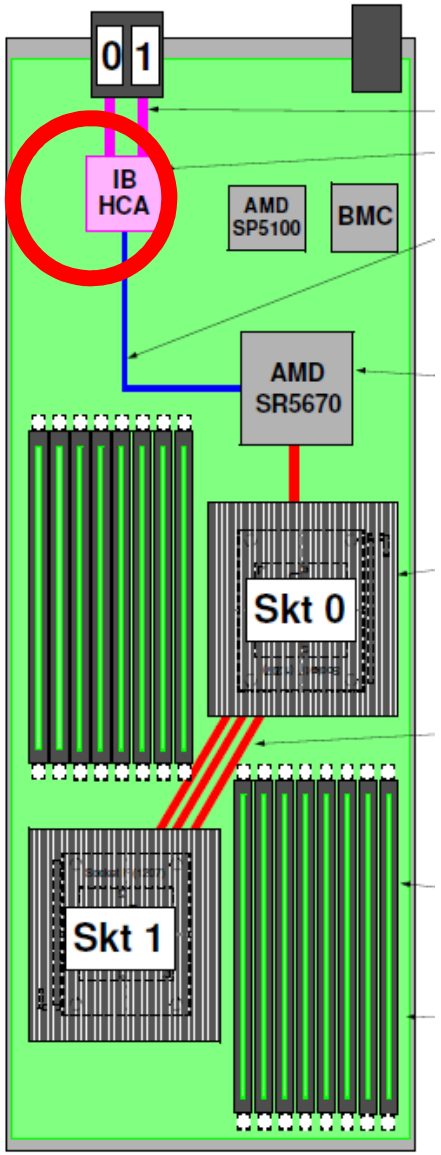


(6) DDR3 RDIMMs
Read BW per Socket:
1333 MHz = 32 GB/s

Two independent Mellanox ConnectX-2 HCAs for 2x off-blade, IB interconnect bandwidth.

SGI ICE Node Blade for AMD Magny Cours

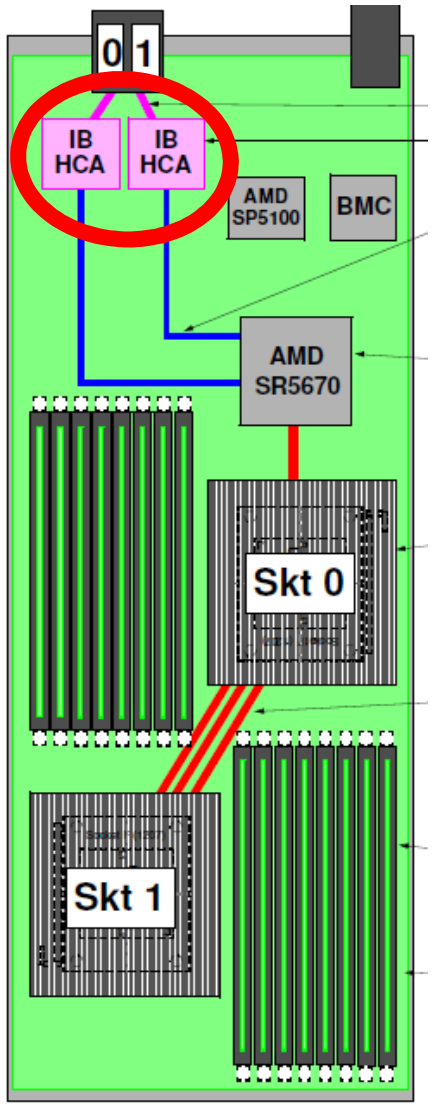
one single dual-ported port QDR HCA



SGI ICE Node Blade for AMD Magny Cours

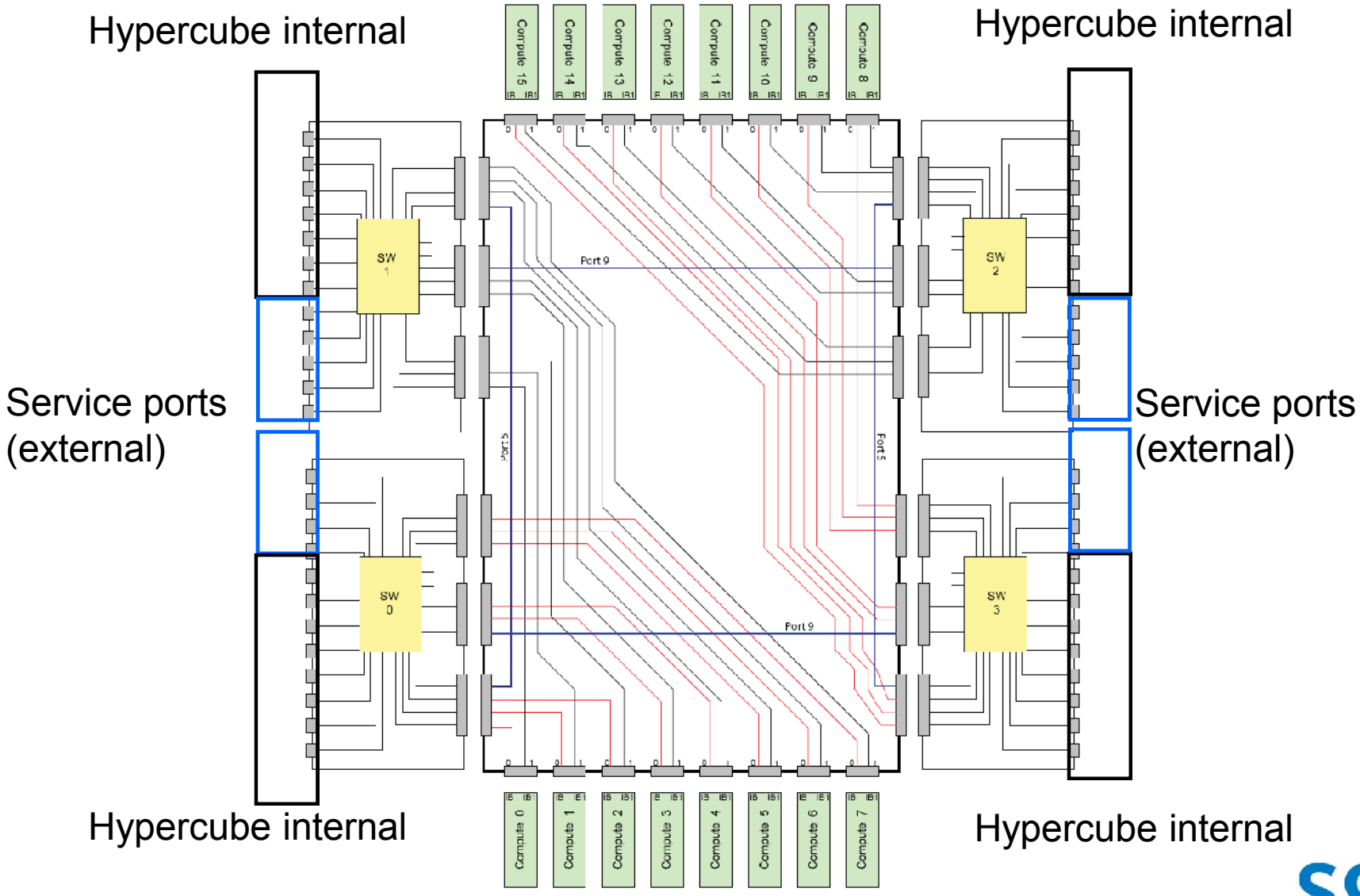
two single ported QDR HCAs

Two independent IB HCAs for 2x off-blade, IB interconnect bandwidth.



sgi

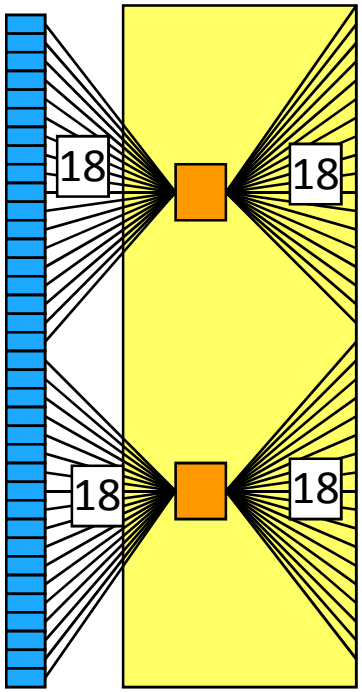
SGI Altix ICE 8400EX Blade Container



Comparison Fat Tree vs. Hypercube

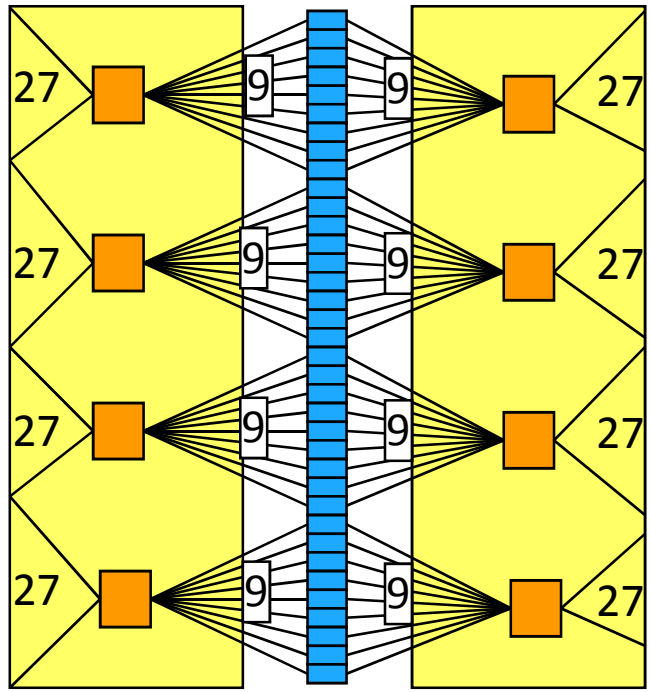
36 nodes

36 external IB ports



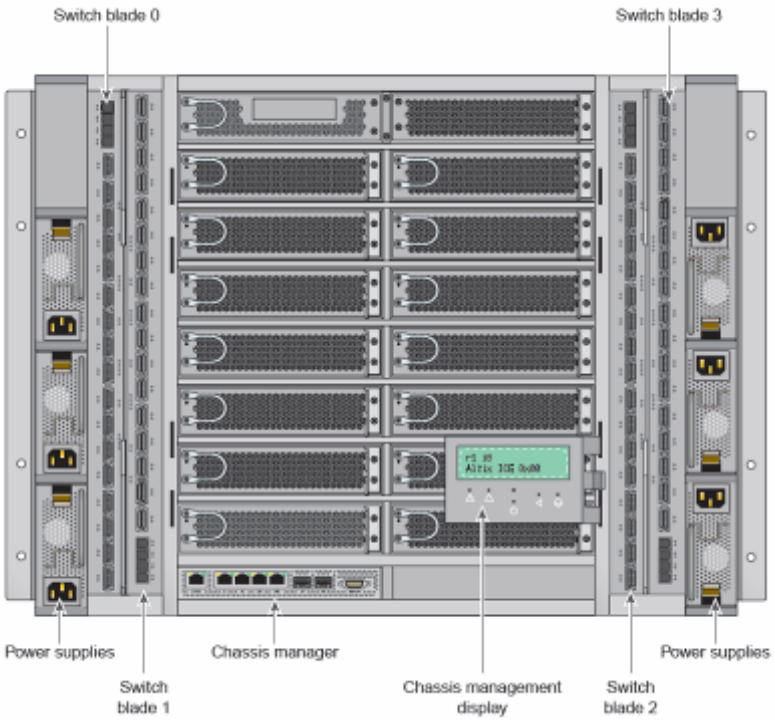
36 nodes

216 external IB ports (8*27)



sgi

Flexibility in Networking Topologies

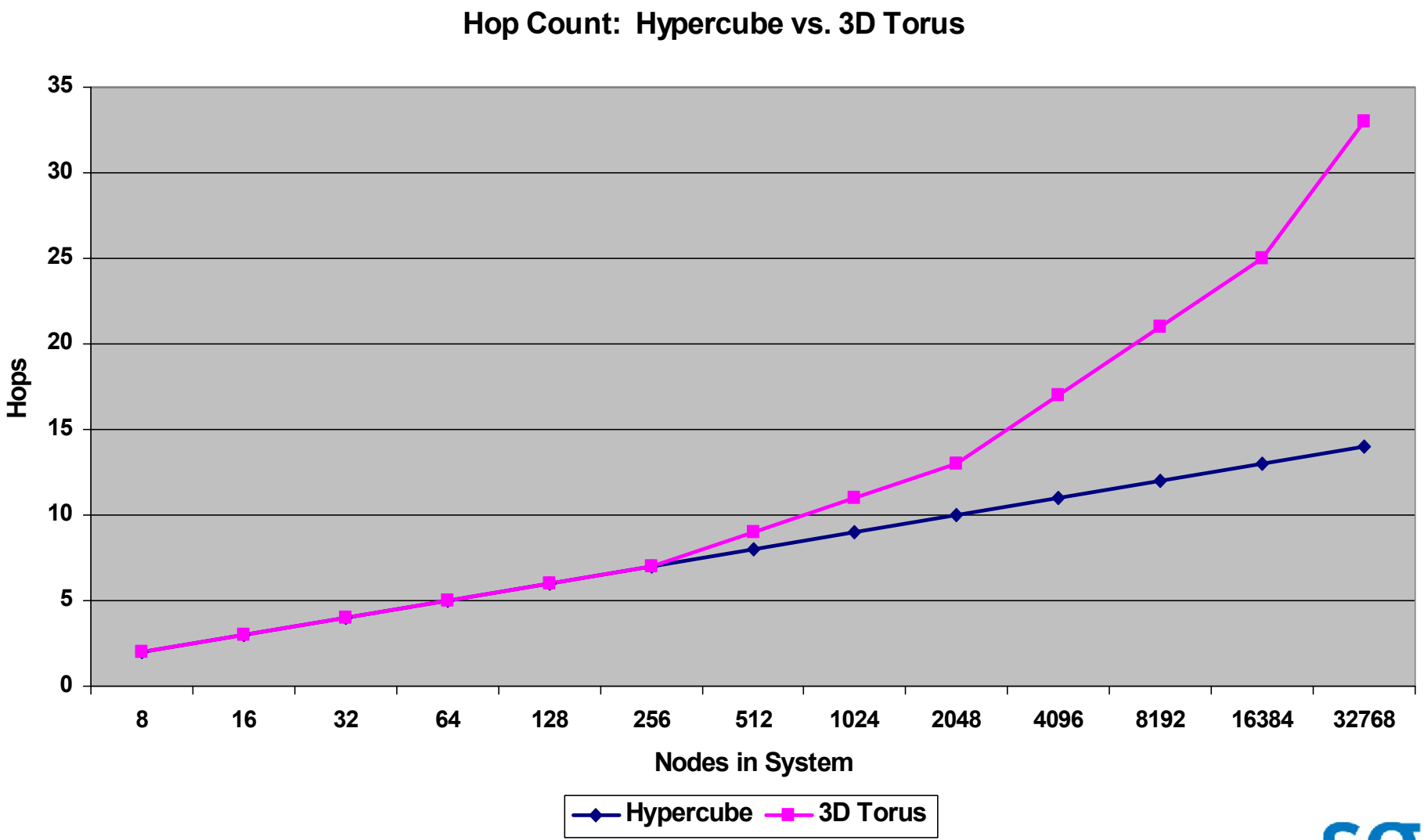


Robust integrated switch blade design enables industry-leading bisectional bandwidth at ultra-low latency!

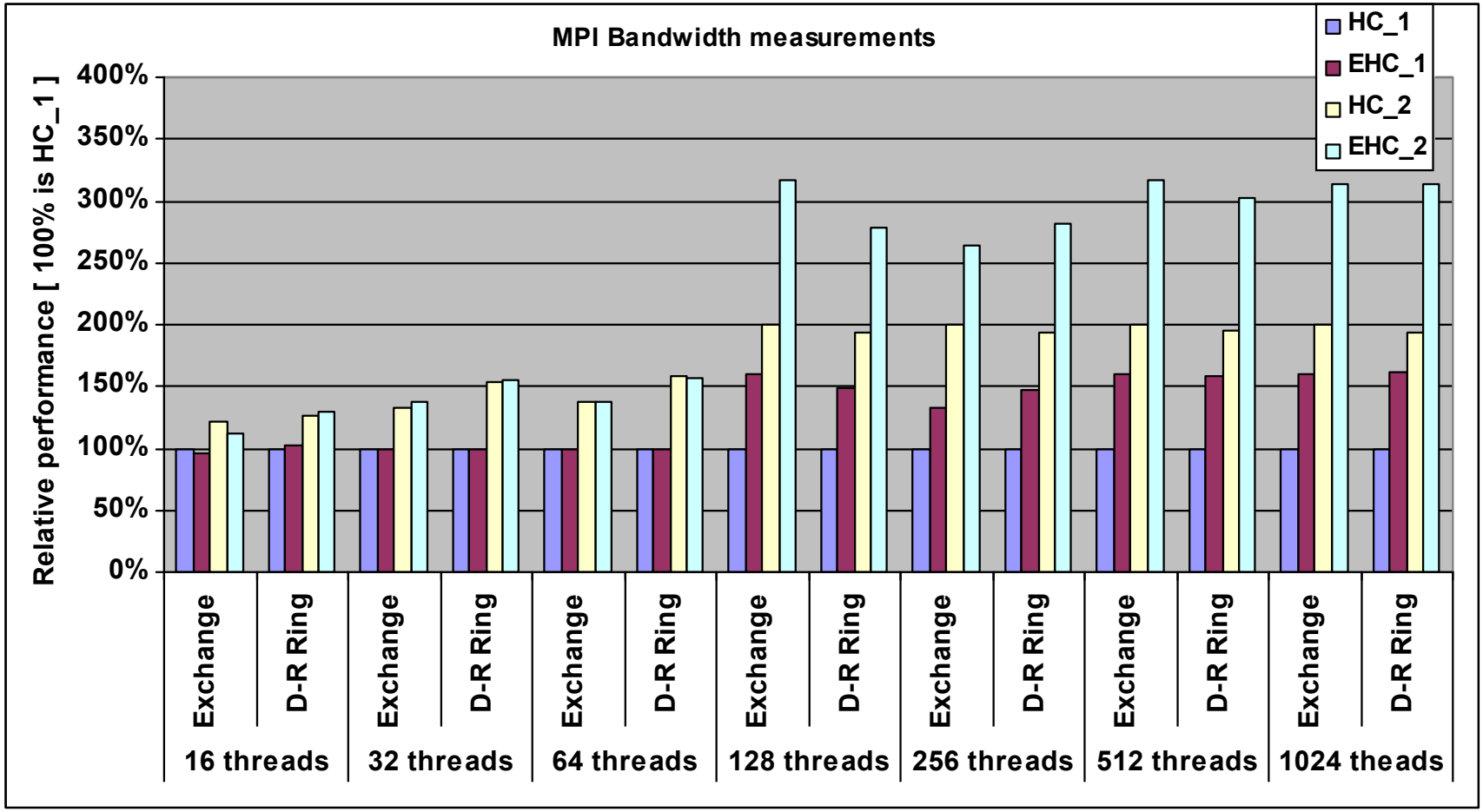
- **Hypercube Topology:**
  - Lowest network infrastructure cost
  - Well suited for "nearest neighbor" type MPI communication patterns
- **Enhanced Hypercube Topology:**
  - Increased bisectional bandwidth per node at only a small increase in cost
  - Well suited for larger node count MPI jobs
- **All-to-All Topology:**
  - Maximum bandwidth at lowest latency for up to 128 nodes
  - Well suited for "all-to-all" MPI communication patterns.
- **Fat Tree Topology:**
  - Highest network infrastructure cost. **Requires external switches.**
  - Well suited for "all-to-all" type MPI communication patterns

sgi

Hypercube vs. 3D Torus



MPI bisection bandwidth measurements



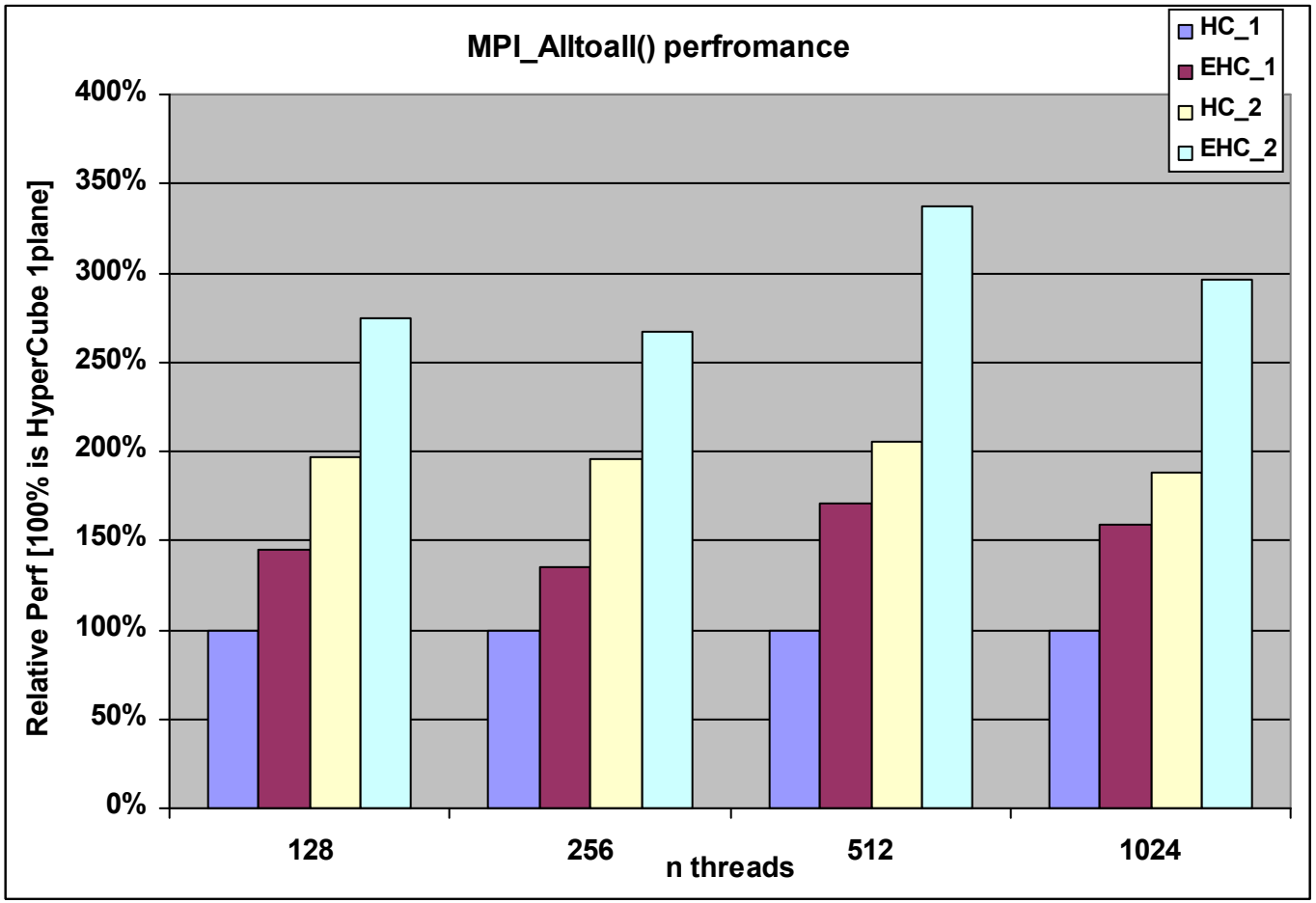
MPT 1.23

Exchange: Simple bisection BW experiment

D-R Ring: Double Random Ring – more accurate bisection BW estimation

sgi

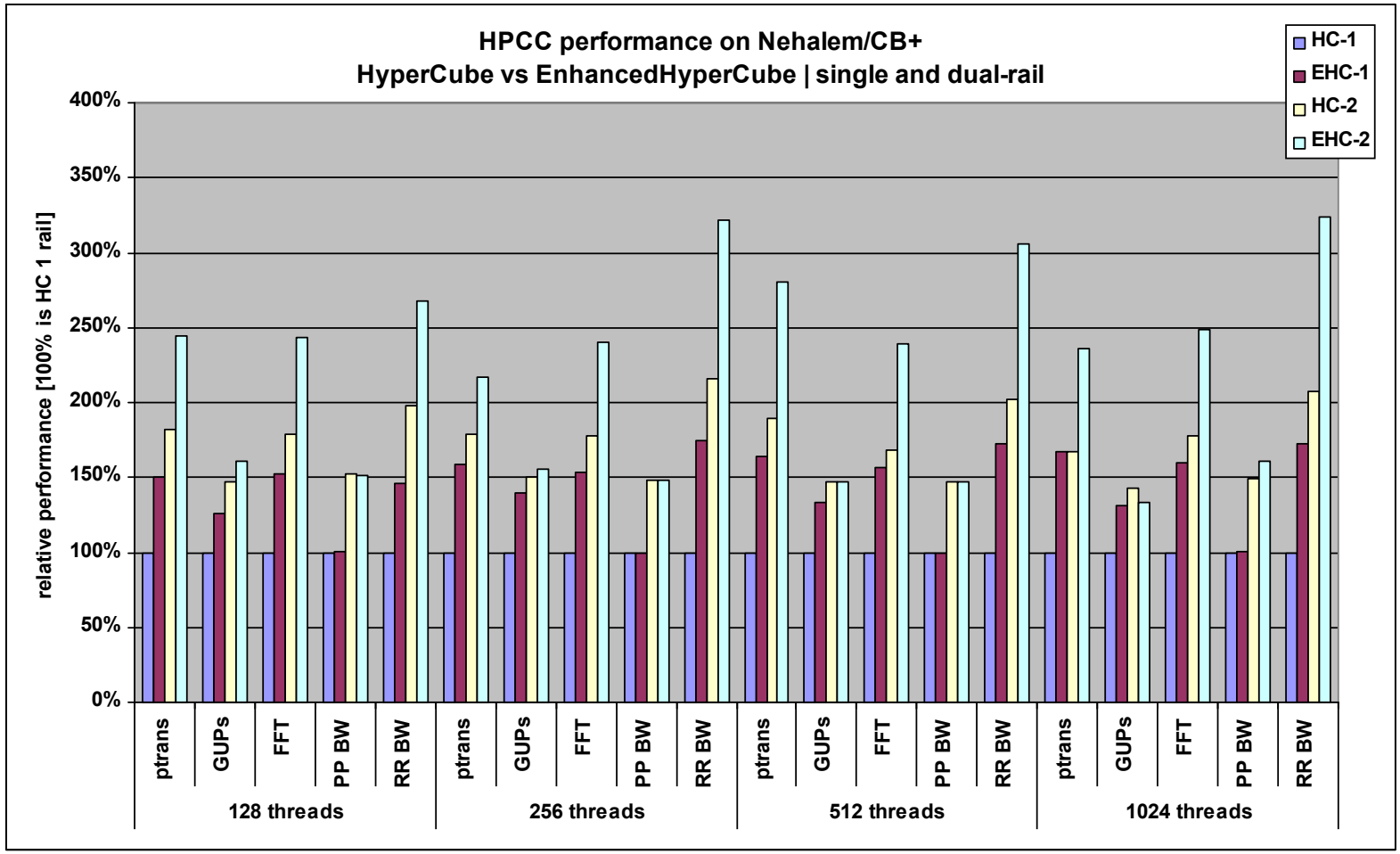
MPI_AlltoAll communications



MPT 1.23 - MPI_Alltoall() – buffer size = 700 KB

sgi

HPCC benchmarks

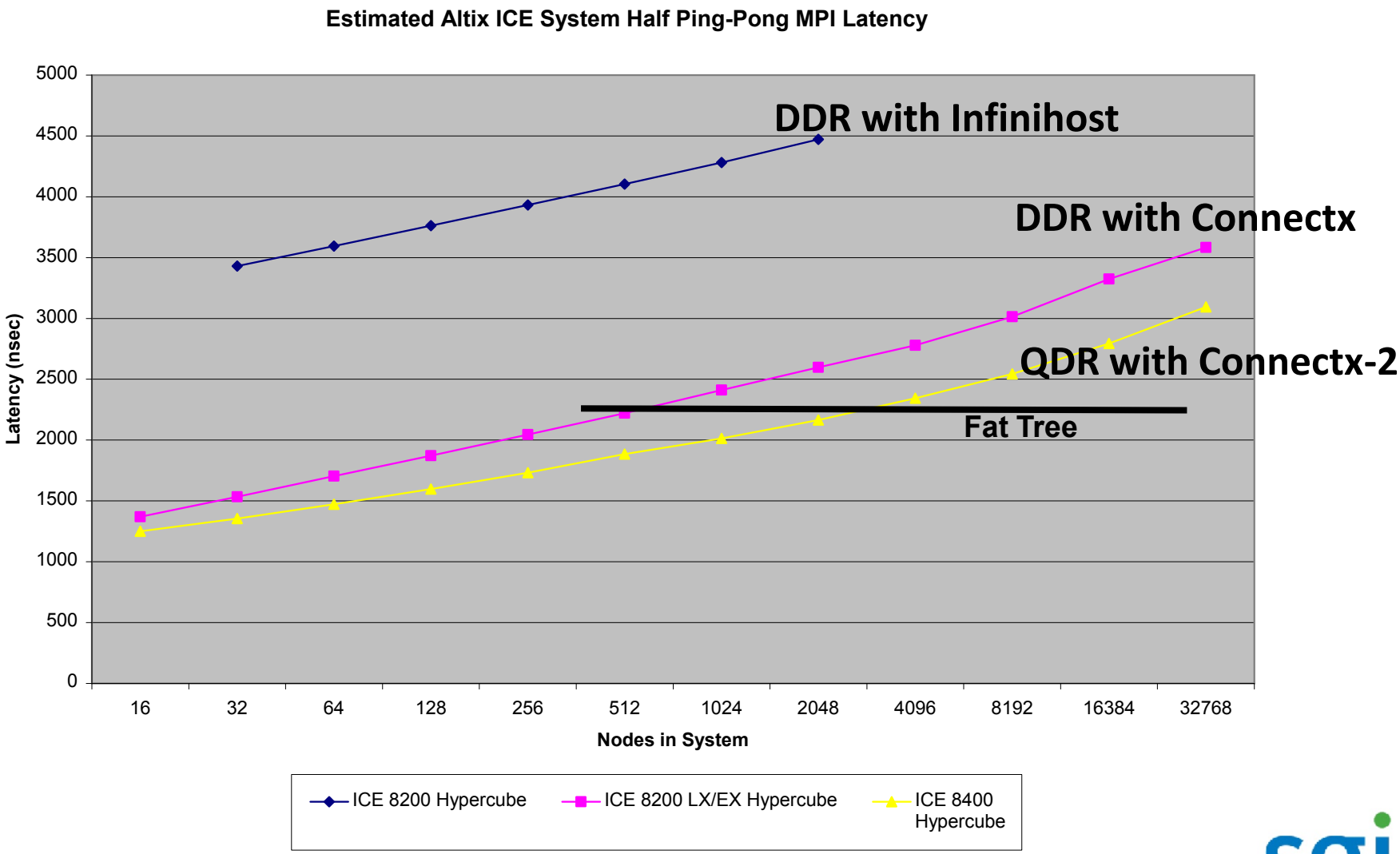


HPCC version 1.x – MPT 1.23

PP BW: PingPong Bandwidth

RR Bandwidth: Random Ring bandwidth

SGI Enhanced Hypercube vs FatTree Latencies



Altix® UV


sgi
Altix® UV
The World's Fastest Supercomputer
Open Platform with
Intel® Xeon® Processors

The SGI Altix Ultraviolet (UV) System

Evolution

from

ccNUMA Shared Memory (SGI Origin)

to

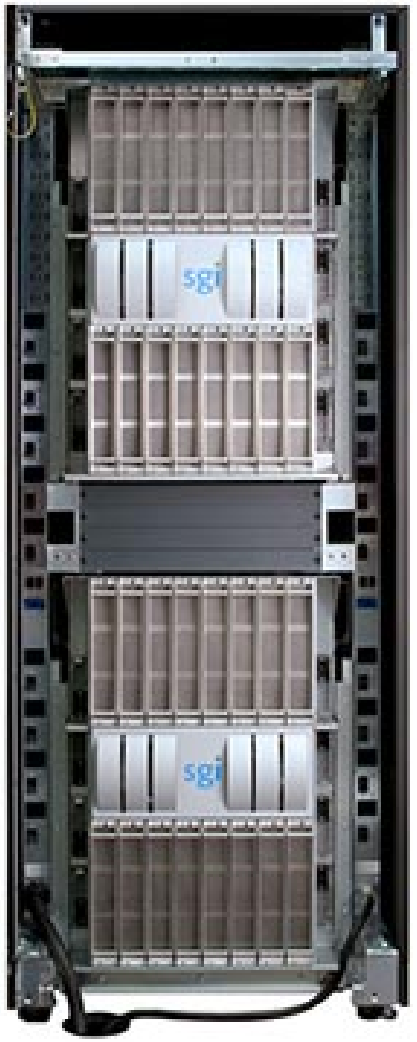
Partitioned Globally Addressable Shared Memory (SGI Altix 4700)

to

HW Accelerated Partitioned Globally Addressable System (SGI Altix UV)

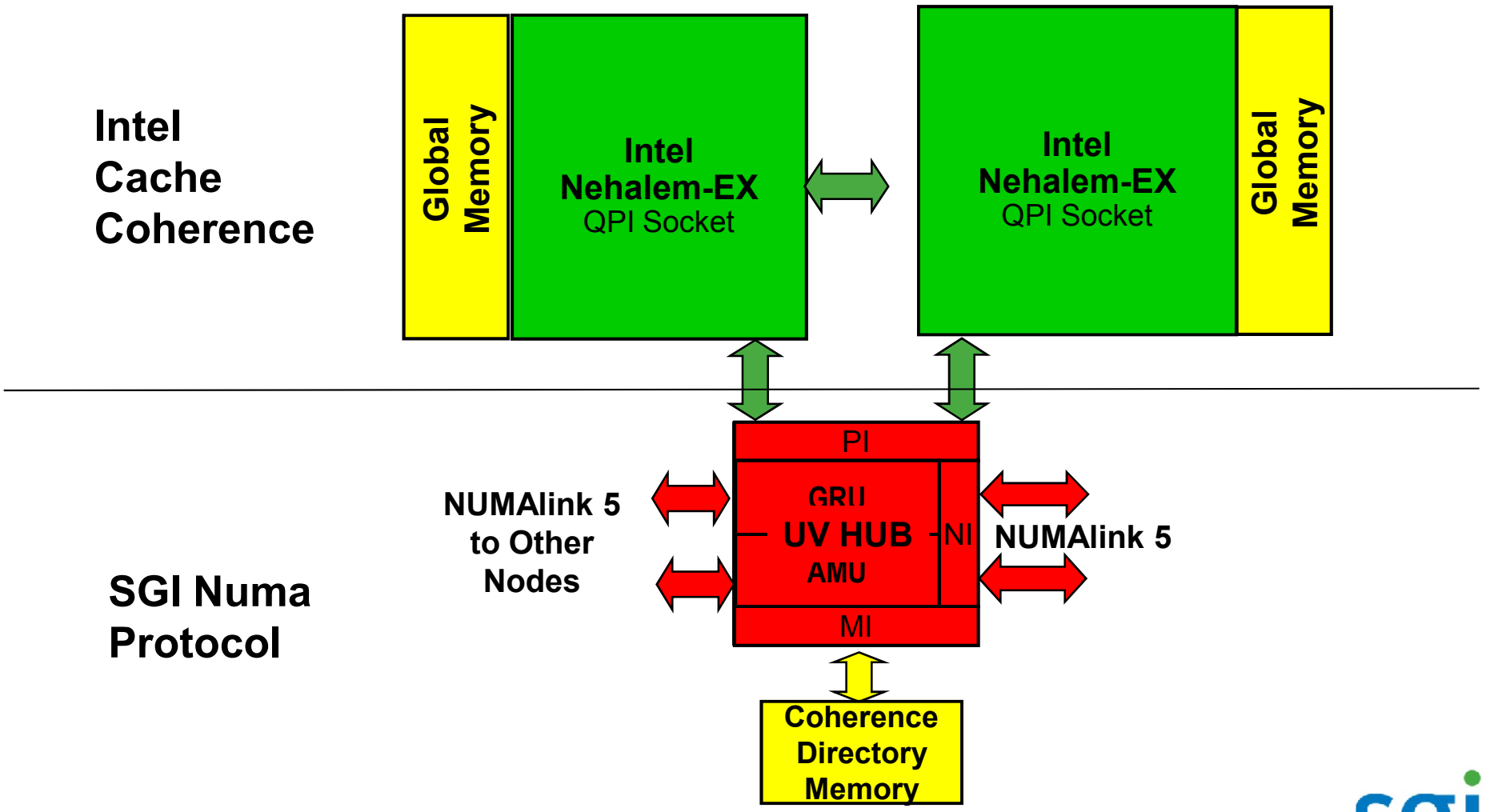
sgi

SGI Altix® UV

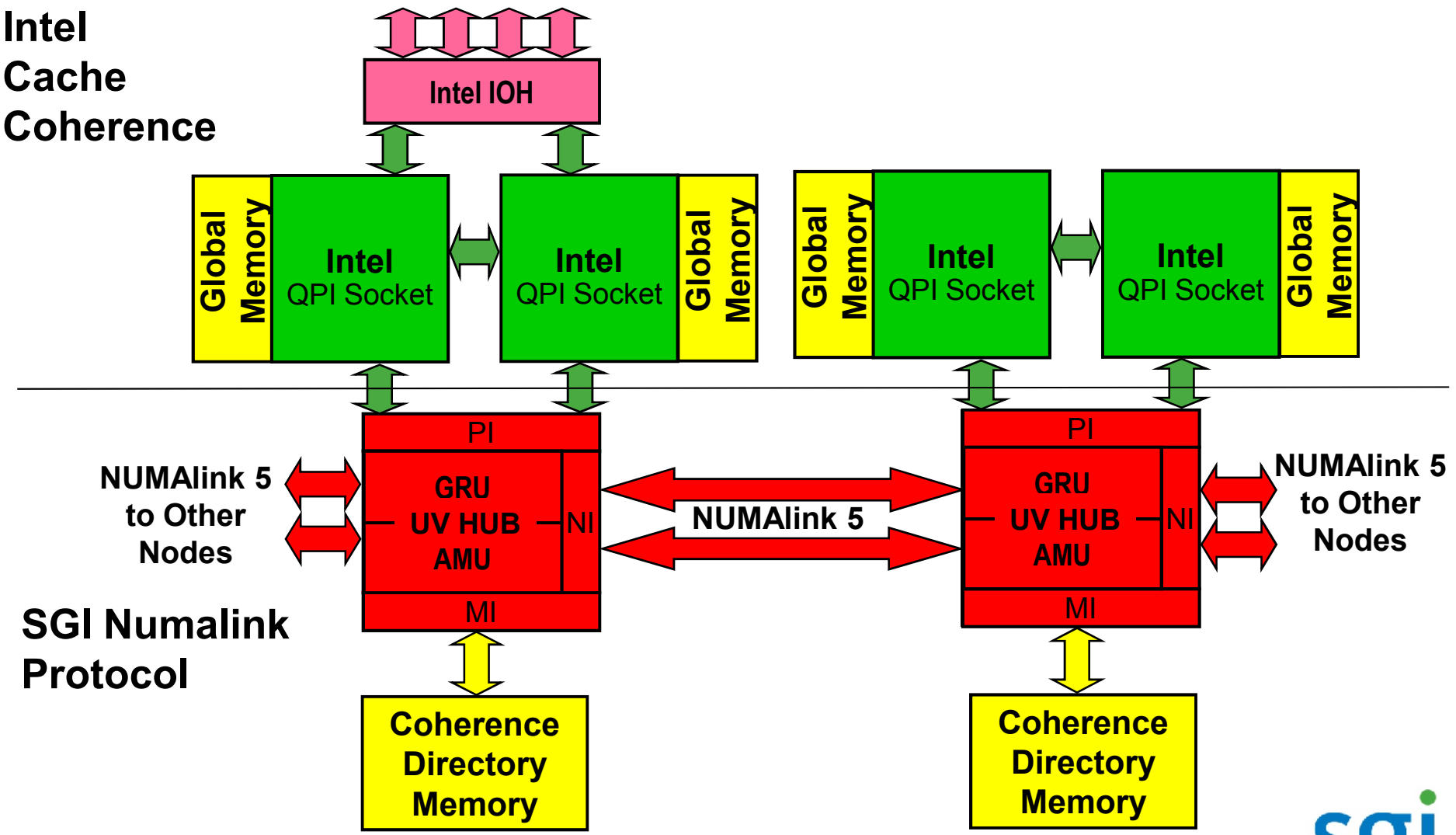


- Partitioned Globally Addressable Memory System
- Advanced, SGI-enhanced bladed architecture
- Intel Nehalem-EX processors
- SGI Numalink Interconnect for shared memory implementation
- Built-in MPI offload engine

UV Node Architecture

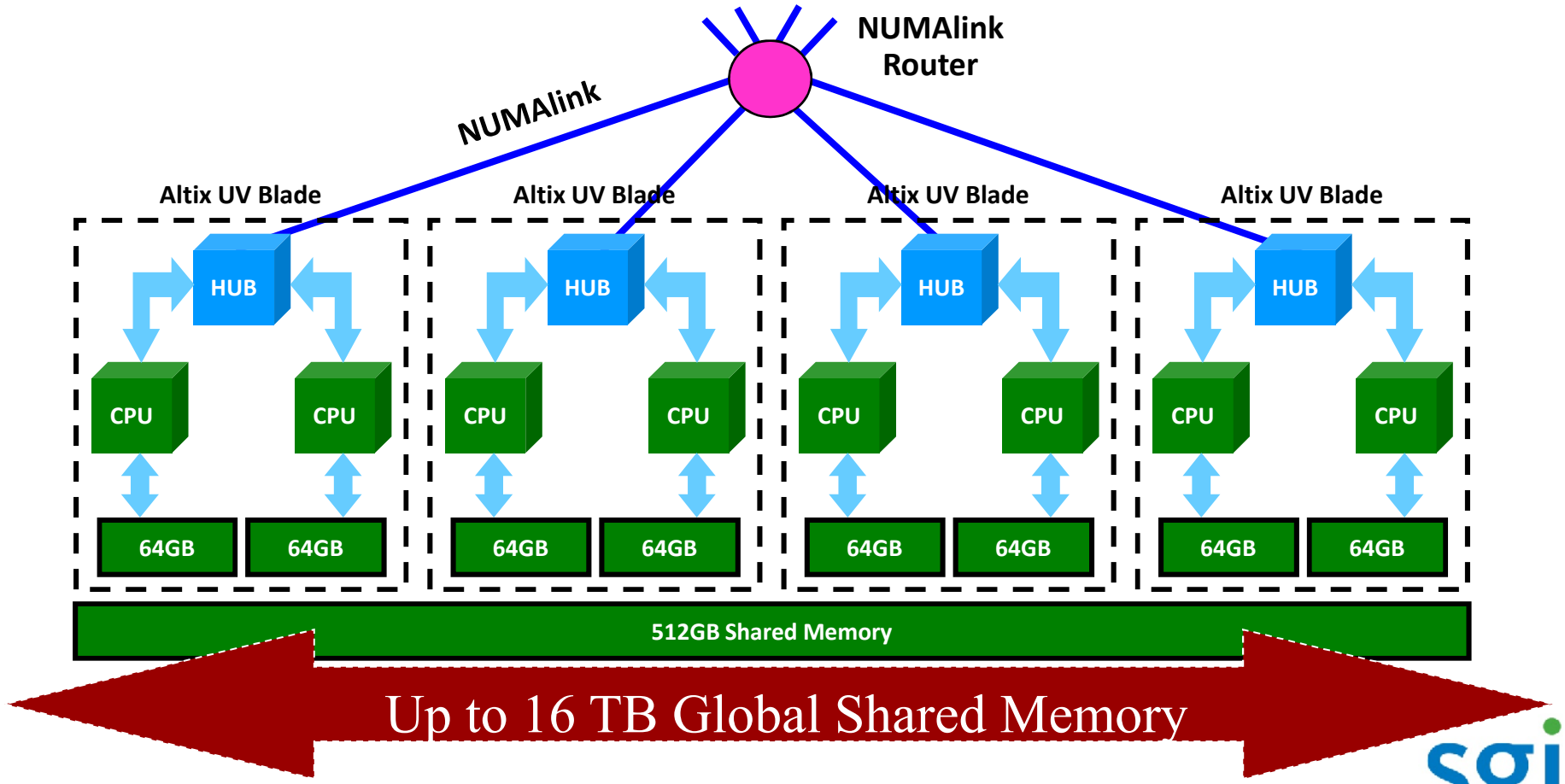


UV Interconnect Architecture



Globally Shared Memory System

- NUMAlink® 5 is the glue of Altix® UV 100/1000



UV Architectural Scalability

- 16,384 Nodes (scaling supported by NUMAlink 5 node ID)
  - 16,384 UV_HUBs
  - 32,768K Sockets / 262,144 Cores (with 8-cores per socket)
- Coherent shared memory
  - Xeon: 16TB (44 bits socket PA)
- 8 Petabytes coherent get/put memory (53 bits PA w/GRU)

sgi

UV_HUB/Node Controller Technologies

Globally Addressable Memory

- Large Shared Address Space (8 PB)
- Extremely Large Coherent Get/Put Space
- Atomic Memory Operations (AMOs) in Coherent Memory
- Coherence Directory

RAS

- X4 DRAM correction
- Redundant Real-Time Clock
- Failure Isolation Between Partitions
- Built-In Debug and Performance Monitors
- Internal/External Datapath Protection

Active Memory Unit

- Rich set of Atomic Operations (e.g. HW barrier support)
- Multicast
- Message Queues in Coherent Memory
- Page Initialization

GRU Global Reference Unit

- For MPI data movement
- For PGAS support
- High-BW, Low-Latency Socket Communication
- Update Cache for many AMOs
- Scatter/Gather Operations
- BCOPY Operations
- External TLB with Large Page Support

NOTE: UV HUB memory management functions do not interfere with fast on-node memory access

sgi

Altix® UV Characteristics

1. **Scalability**
2. **Performance**

sgi

SGI Altix UV Scalability

Single System Images

- **Single System Image scales to 256 Intel® Xeon® "Nehalem-EX" sockets (2048 cores) & 16TB memory**
  - **Intel coherence within blade**
  - **SGI coherence between blades**
  - **16 TB is the global shared memory limit of "Nehalem-EX" processor**

sgi

SGI Altix UV Scalability

Architectural Limits

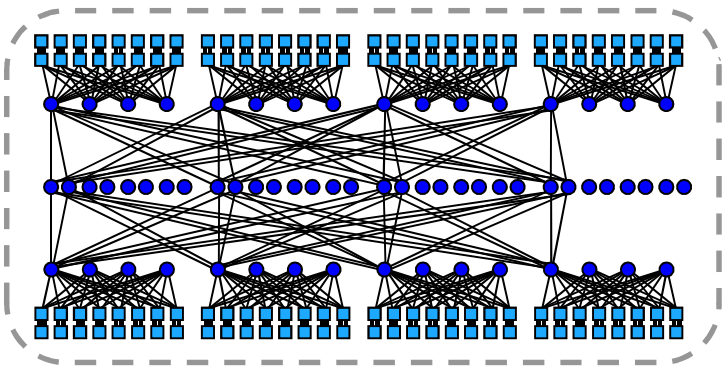
- Altix® UV's architecture supports scaling to Petaflop level
- Upper limit on scaling is the Altix UV hub, capable of connecting 32,768 sockets

Petaflop System

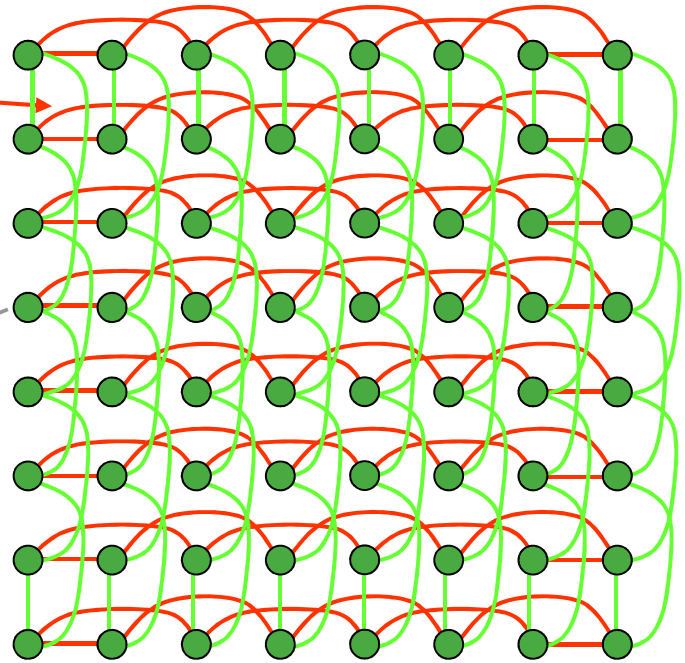
Each Red & Green Torus link shown is (2) links / L1R

(8) L1Rs per plane
(8 of 16) ports / L1R support Fat Tree
(8 of 16) ports / L1R support 2-copies of Torus
(16) copies of Torus per plane

Green Links (Interleaved across the aisles)



256-Socket Fat Tree Building Block (4 racks)



Red Links (Interleaved down the ranks)

= 4-Rack Group

SGI Altix UV Performance

SPEC Benchmarks

- **World record SPECint_rate and SPECfp_rate performance with only 64 sockets populated!**
  - **SPECint_rate_2006: #1 on any architecture**
  - **SPECfp_rate_2006: #1 on x86 architecture, #2 behind SGI Altix 4700 with eight times as many processors**

sgi

SGI Altix UV Performance

SPEC Benchmarks

SPECint_rate_base2006:

#1: SGI Altix UV 1000 512c Xeon X7560	**10400**
#2: SGI Altix 4700 Bandwidth System 1024c Itanium	9030
#3: Sun Blade 6048 Chassis 768c Opteron 8384 (cluster)	8840
#4: ScaleMP vSMP Foundation 128c Xeon X5570	3150
#5: SGI Altix 4700 Density System 256c Itanium	2890

SPECfp_rate_base2006:

#1: SGI Altix 4700 Bandwidth System 1024c Itanium	10600
#2: SGI Altix UV 1000 512c Xeon X7560	**6840**
#3: Sun Blade 6048 Chassis 768c Opteron 8384 (cluster)	6500
#4: SGI Altix 4700 Bandwidth System 256c Itanium	3420
#5: ScaleMP vSMP Foundation 128c Xeon X5570	2550

Source: www.spec.org (March, 2010)

sgi

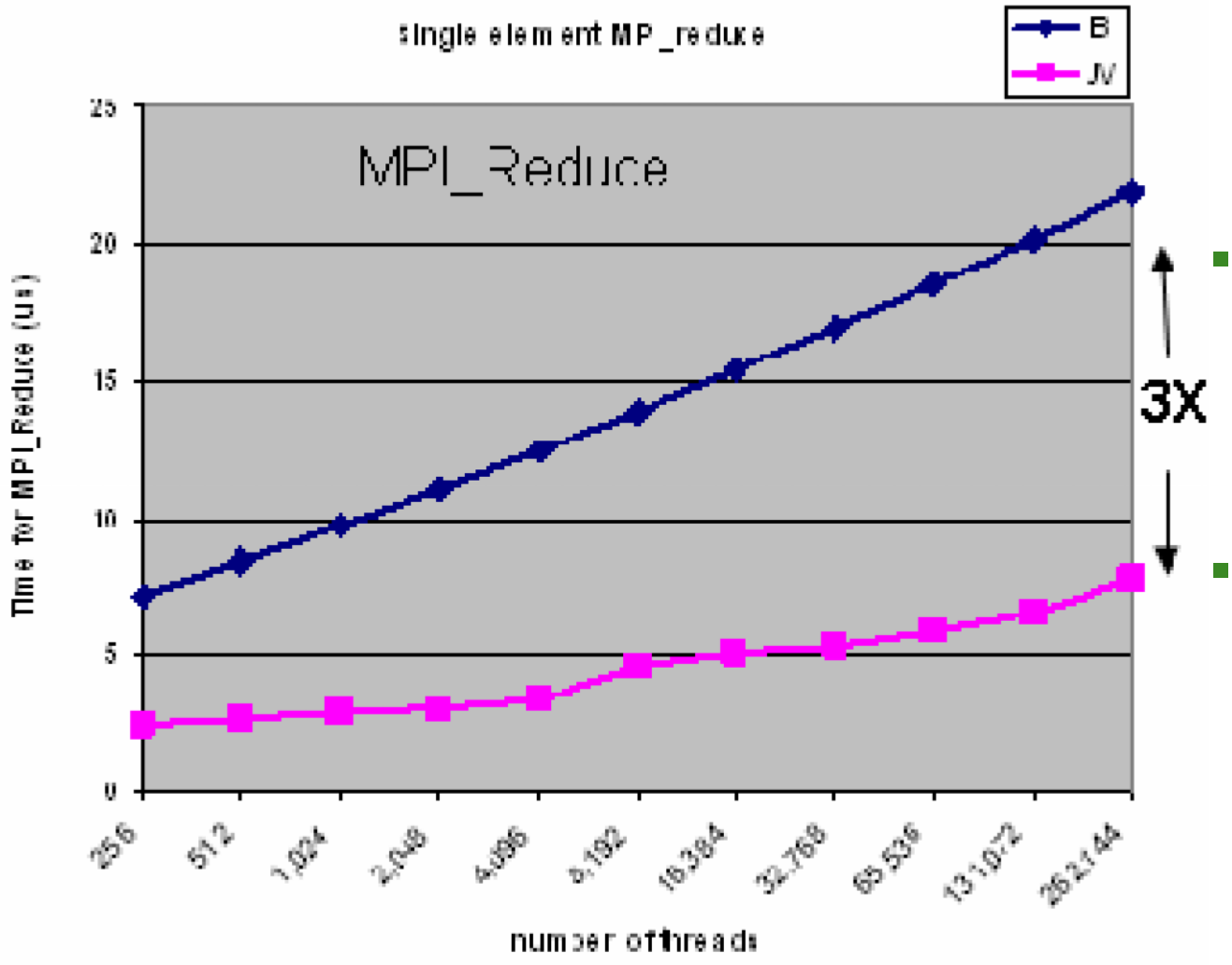
SGI Altix UV Performance

- **Shared memory capacity per SSI (max. 16TB)**
  - **Massive speed-up for memory-bound applications**
- **MPI Offload Engine (MOE)**
  - **frees CPU cycles and improves MPI performance**
  - **MPI reductions 2-3X faster than competitive clusters/MPPs**
  - **Barriers up to 80X+ faster**

sgi

SGI Altix UV

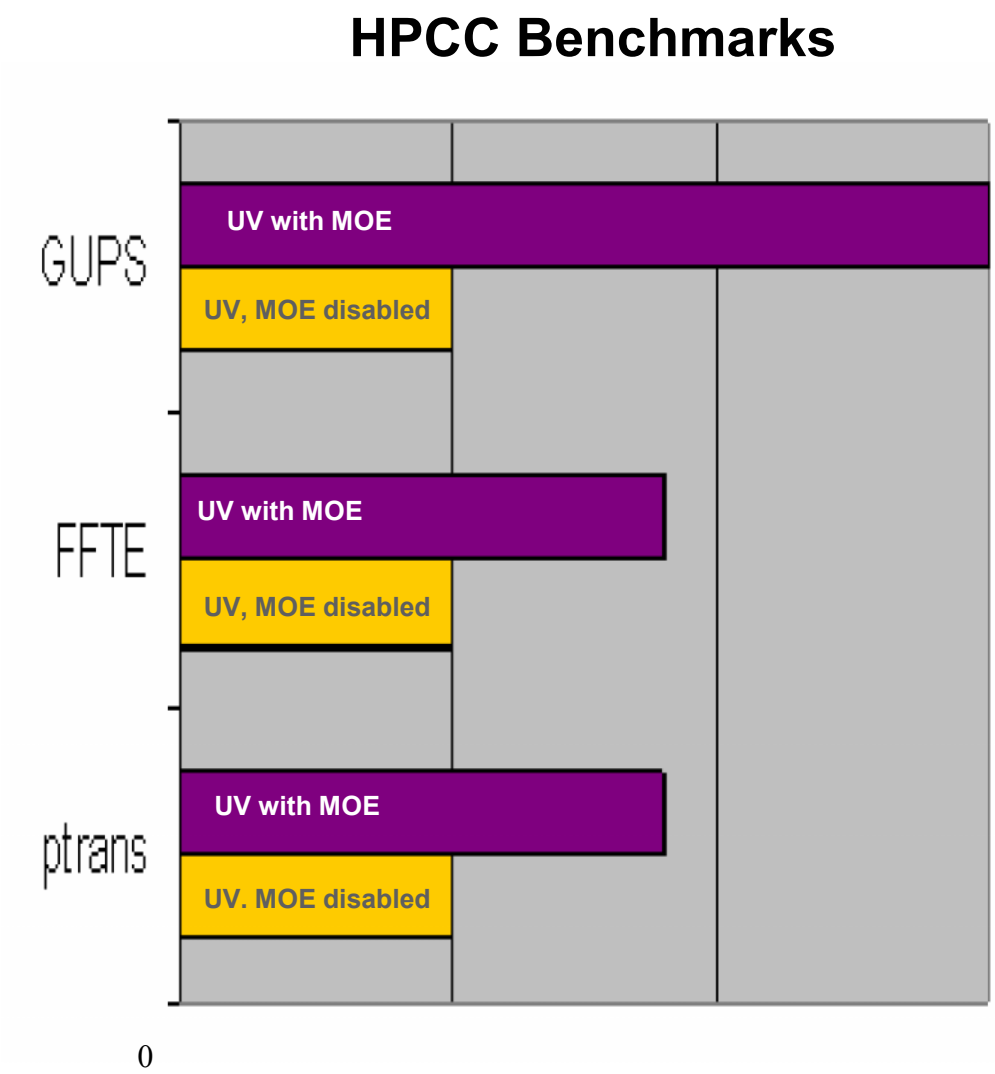
MPI Performance Acceleration



- Altix UV offers up to 3X improvement in MPI reduction processes over standard IB networks
- Barrier latency is dramatically better than competing platforms (up to 80 times)

SGI Altix UV

Performance Acceleration with MPI Offload Engine (MOE)



- HPCC benchmark show substantial improvements with MPI Offload Engine (MOE)

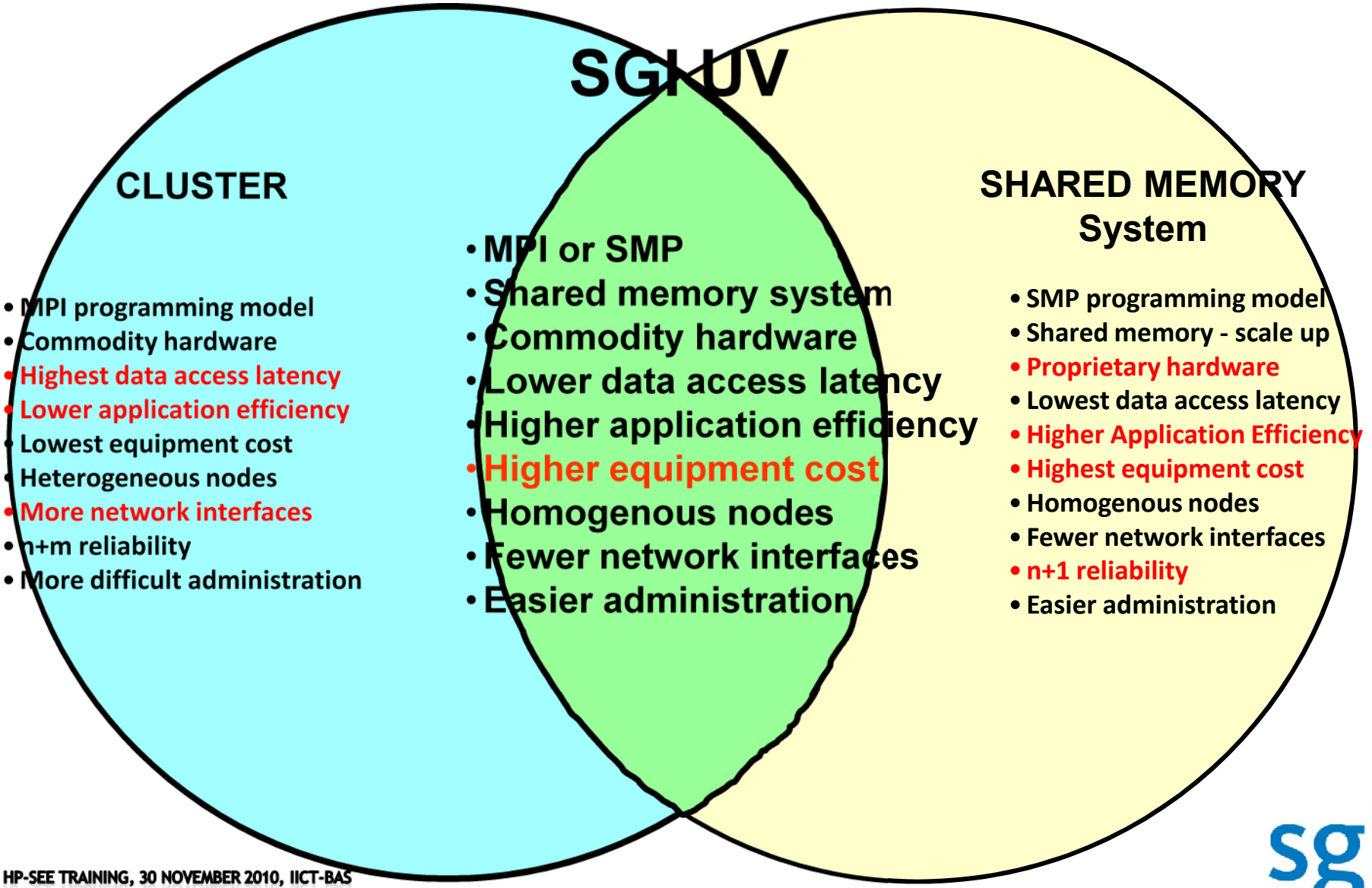
Source: SGI Engineering projections

SGI Altix® UV

- **World's Fastest** with
  - World Record SPECint_rate and SPECfp_rate Performance
  - High speed NUMAlink® 5 interconnect (15 GB/sec)
  - MPI offload engines maximize efficiency
- **World's Most Scalable** with
  - Single system image scales up to 2048 cores & 16TB memory
  - Direct access to global data sets up to 16TB
- **World's Most Flexible** in
  - Investment protection:
    - start with four sockets and scale up over time if needed
    - Start with 2048-core SSI and partition over time if needed
  - Compelling performance regardless of type of application
- **Open Platform** which
  - Leverages Intel® Xeon® 7500 ("Nehalem-EX") processors
  - Runs industry-standard x86 operating systems & application code

sgi

MPI and SMP Programming Models



A Key Challenge for Future HPC Systems
The System Interconnect

- The SGI ICE 8400 system and its IB-based, enhanced system interconnect
  - Significantly improves interconnect bandwidth without adding cost
- The SGI Altix UV and its Numalink-based system interconnect
  - Significantly improves interconnect latencies and performance on complex MPI operations
  - Allows very large memories without performance degradation

sgi

Topics

- SGI Customers, Target Markets, and Product Focus
- A Key Challenge for Future HPC Systems – The System Interconnect
  - Applications Analysis
  - The SGI ICE Approach
  - The SGI UV Approach (solves the large memory problem at the same time)
- A Futuristic Data Center – SGI ICE Cube
- The Structure of large HPC Data Centers in Europe

sgi