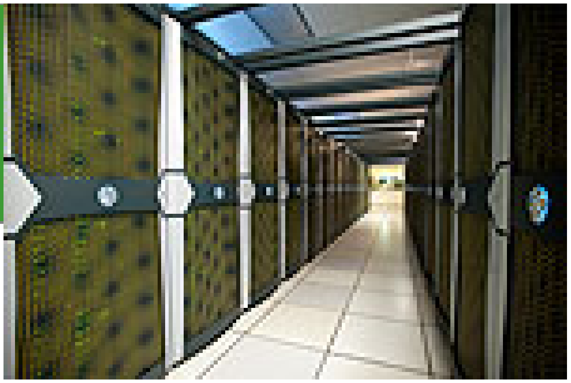
Nasa Pleiades Infiniband Communications Network

HPDC 2009 Munich

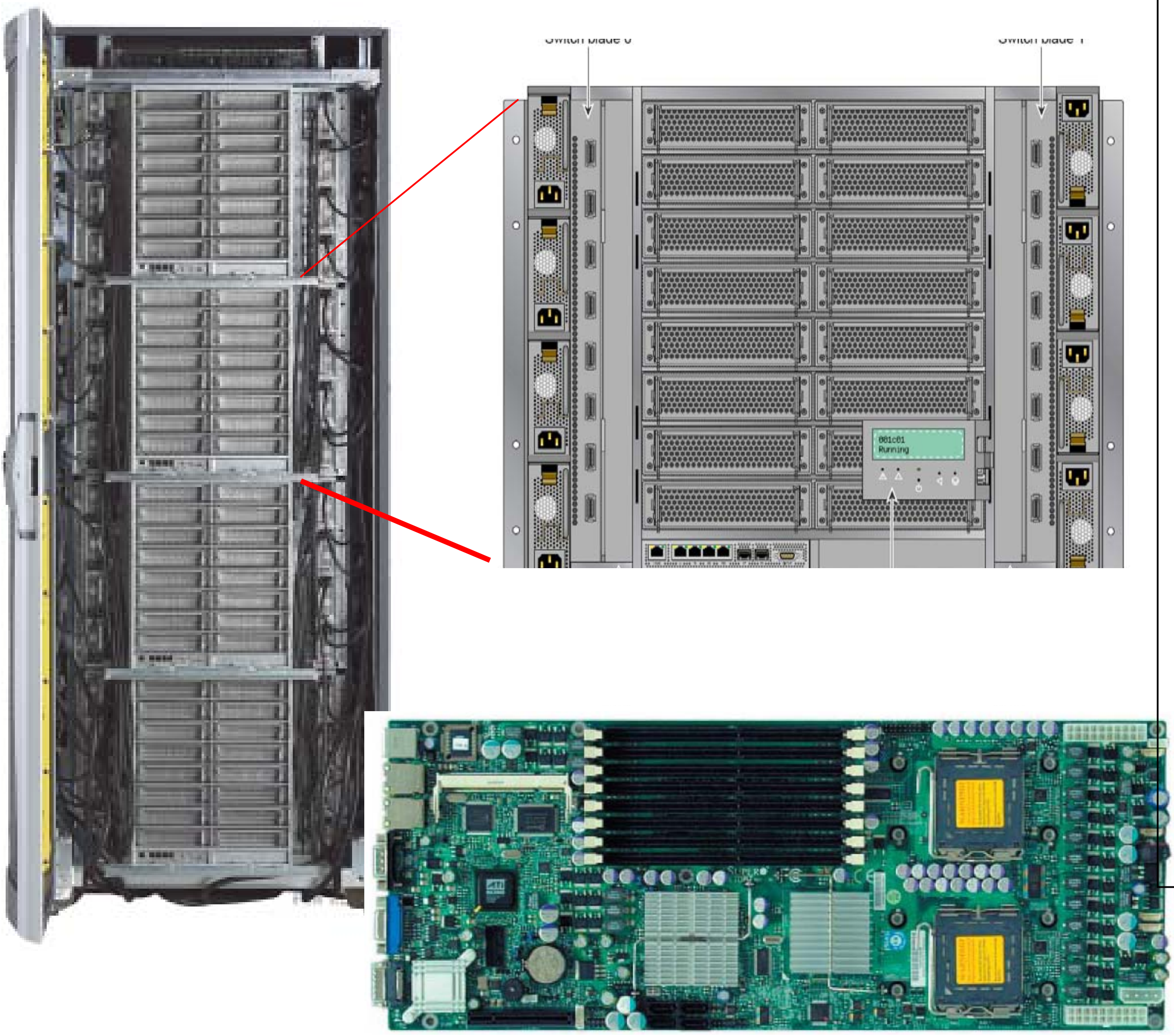
Ruediger Wolff
rgw@sgi.com

NASA Pleiades Supercomputer



- Top500 11/08: place number 3
- 51200 cores
- 608.83 TF/s Rpeak
- 487.01 TF/s Rmax 80% efficiency
- 100 Compute Racks
  - 64 nodes each
  - Intel Xeon E5472 (Harpertown, 3 GHz)
- Infiniband network
  - 10D Hypercube topology
  - Two independent network planes

AGI Altix ICE: Integrated Compute Environment Blades, Enclorures, Infiniband and Racks



- Blades
  - 2 Intel multicore chips
  - diskless blades
  - Remote management
- Enclosure
  - Big savings in cables through backplane
  - N+1 Fans, Powersuplies
- Rack
  - 4 Enclosures per rack
  - 16 Blades per enclosure
  - 64 blades per rack
  - 128 Intel chips p. rack
- Infiniband
  - HCA on Motherboard
  - Infiniband Backplane
  - Integrated IB “edge”-switches

sgi

Infiniband Network

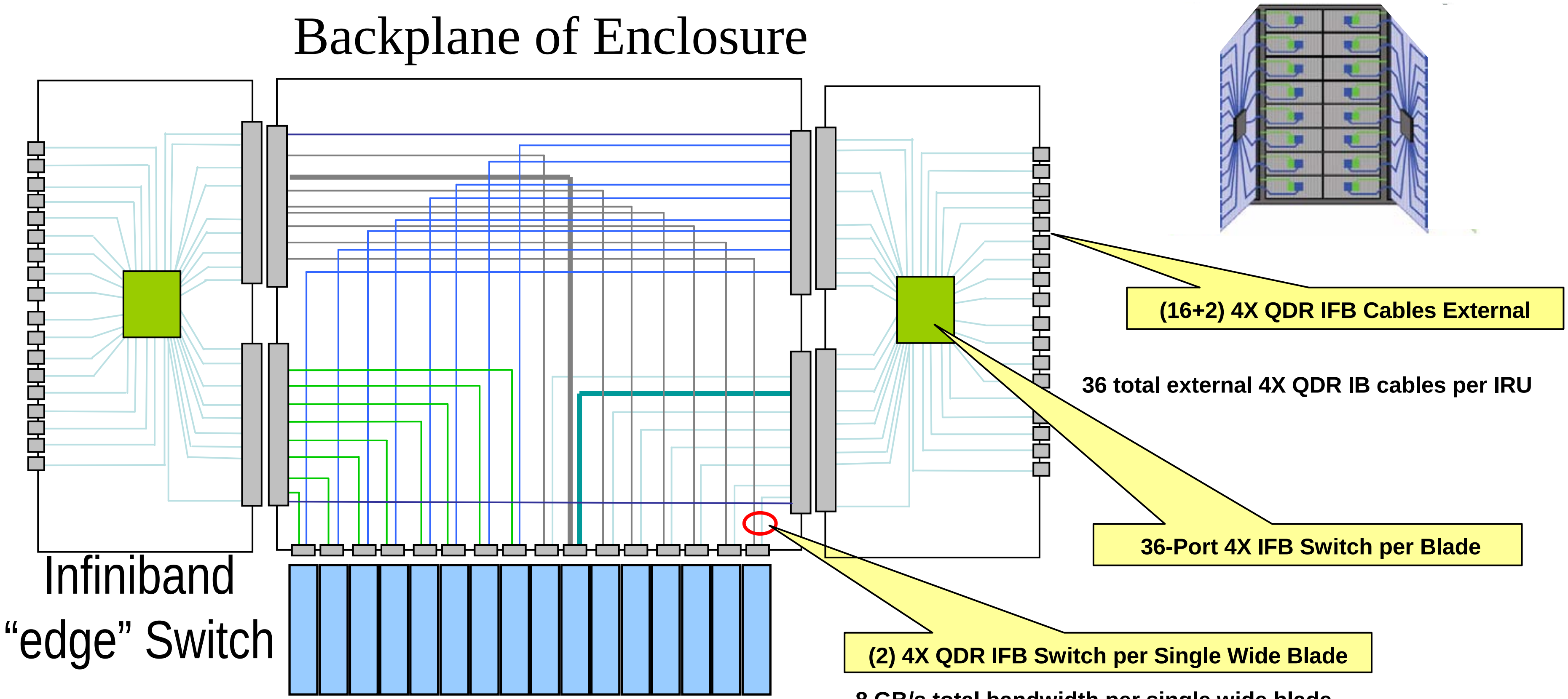
- Open Fabric and switch management software
  - OFED and OPENSM
- 4xDDR and 4xQDR supported
  - Static min-hop routing scheme
- Dual-port Infiniband HCAs enable
  - Two independent networkplanes
  - Used as two separate planes
    - MPI communications on one plane
    - I/O and TCP/IP on other plane
  - Dual-rail operation support in SGI MPI, Intel MPI and others
    - alternate messageblocks between network ports on HCA
    - Near linear scaling for larger messages
  - Redundant network

Infiniband Network

- Choice of Infiniband network topology
  - Clos Network using “big Infiniband Switches”
  - Hypercube network
- SGI enhanced Hypercube Network
  - No additional “big switches”
  - Good bisection bandwidth
  - Low latency across the system
  - Implementation does not need special length cables

SGI Altix ICE 4xQDR IFB Backplane Topology

Backplane of Enclosure



Infiniband “edge” Switch

(16+2) 4X QDR IFB Cables External

36 total external 4X QDR IB cables per IRU

36-Port 4X IFB Switch per Blade

(2) 4X QDR IFB Switch per Single Wide Blade

8 GB/s total bandwidth per single wide blade

16 x 8 = 128 GB/s total bandwidth for all nodes

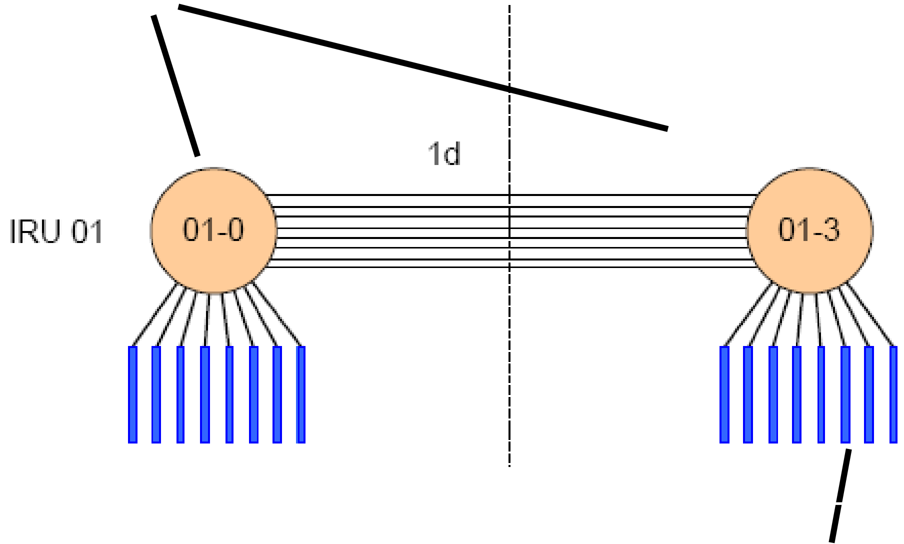
Enclosure capable of supporting 2 fully redundant, non-blocking topologies using single wide blades

(16) Blades

Edge switches part of Blade enclosure infrastructure

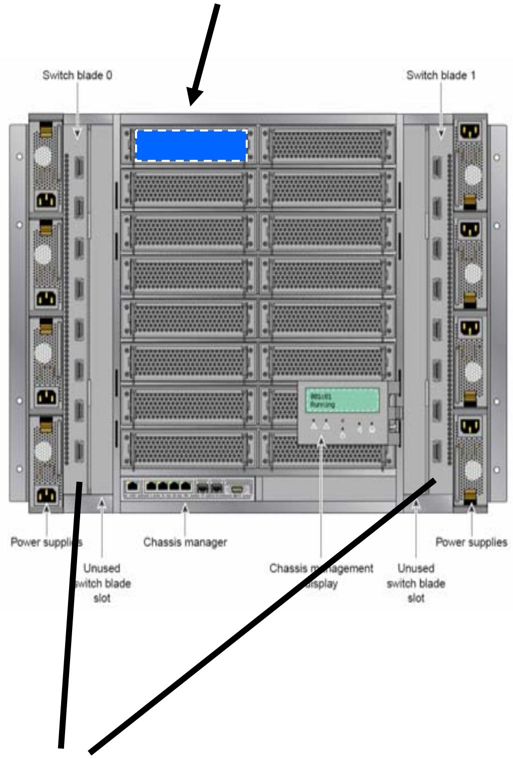
Construction of the single plane Hypercube

Infiniband switches



Processor blade

Processorblade



Infiniband switches

1D Hypercube, single Blade enclosure

16 Blades, 32 sockets, 128 cores

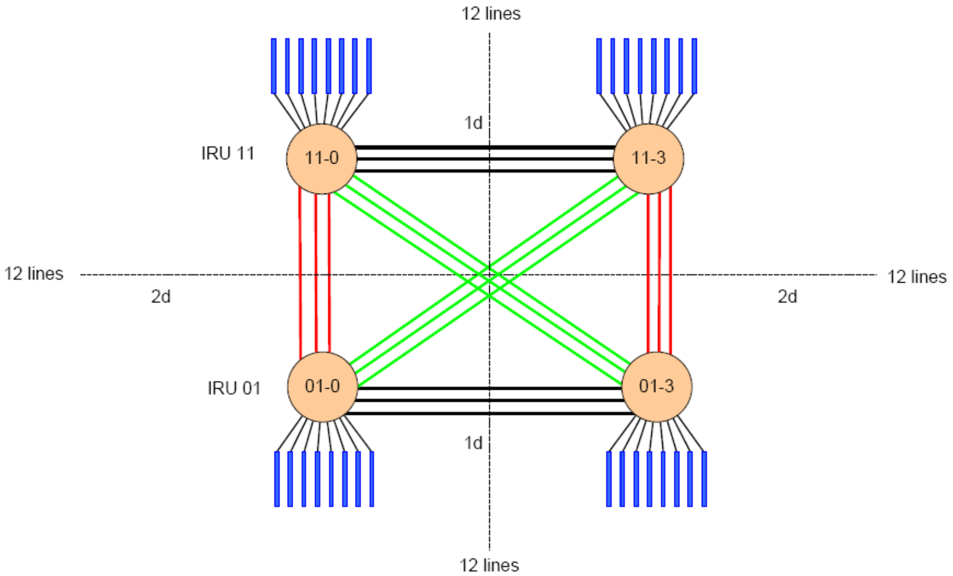
Hypercubes build from a single blade enclosure

are called regular hypercubes

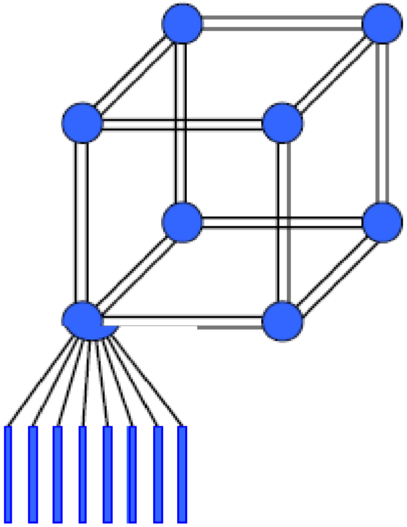
2D Hypercube

2D Hypercube, 2 Enclosures

32 Blades, 64 sockets, 256 cores



Single Rack – 3D Hypercube

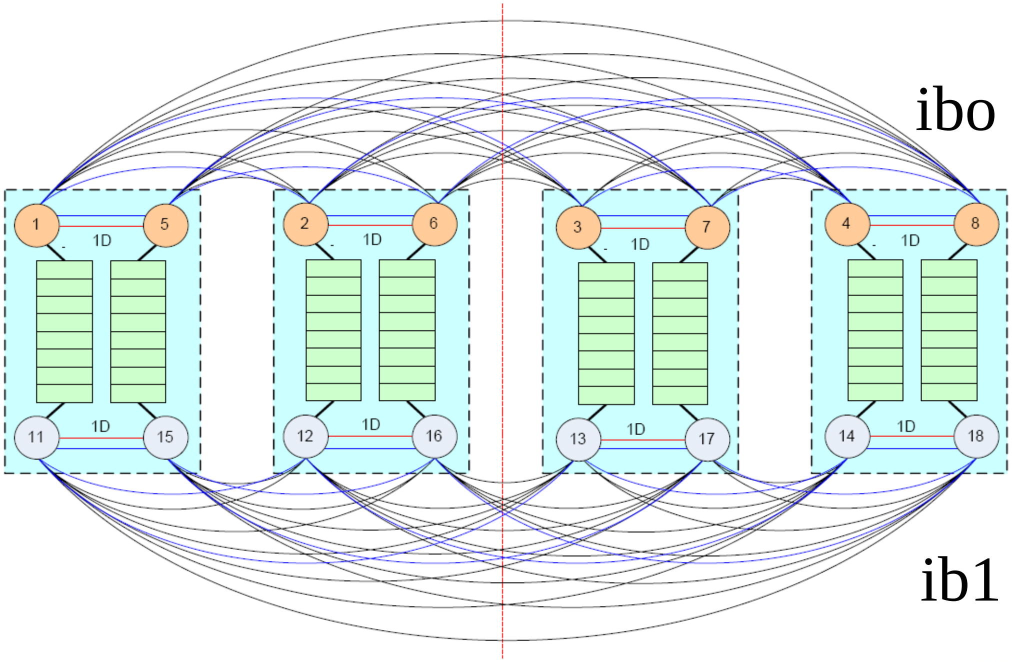


This 3d hypercube represents a single rack.

3D Hypercube

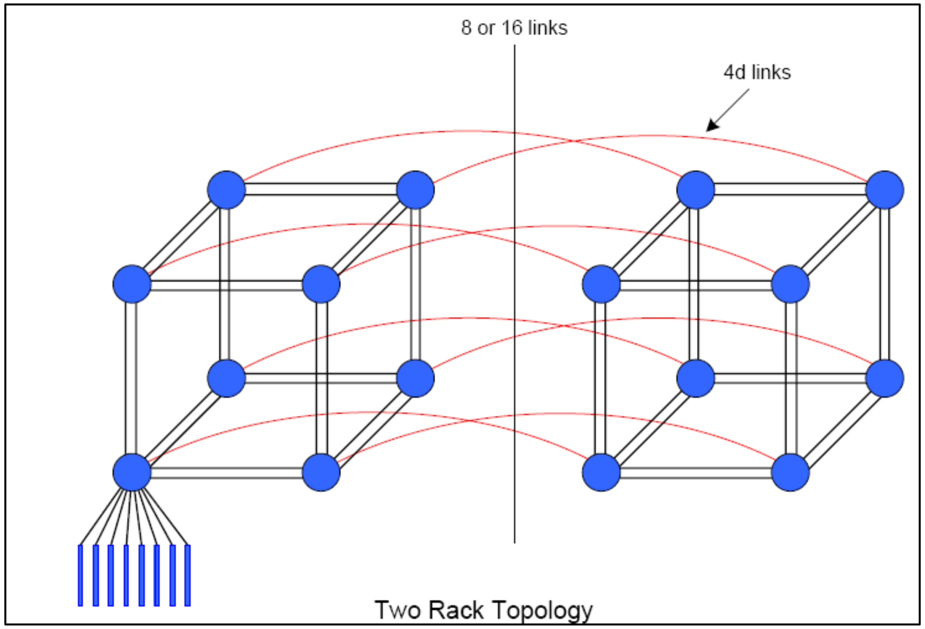
Two indendent parallel network planes

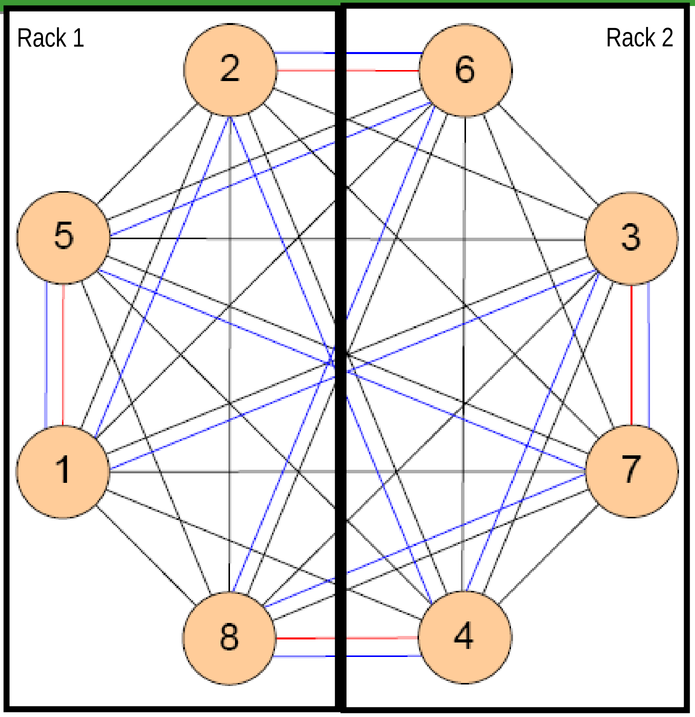
ibo



ib1

Two Racks – 4D enhanced Hypercube – Basic Cell

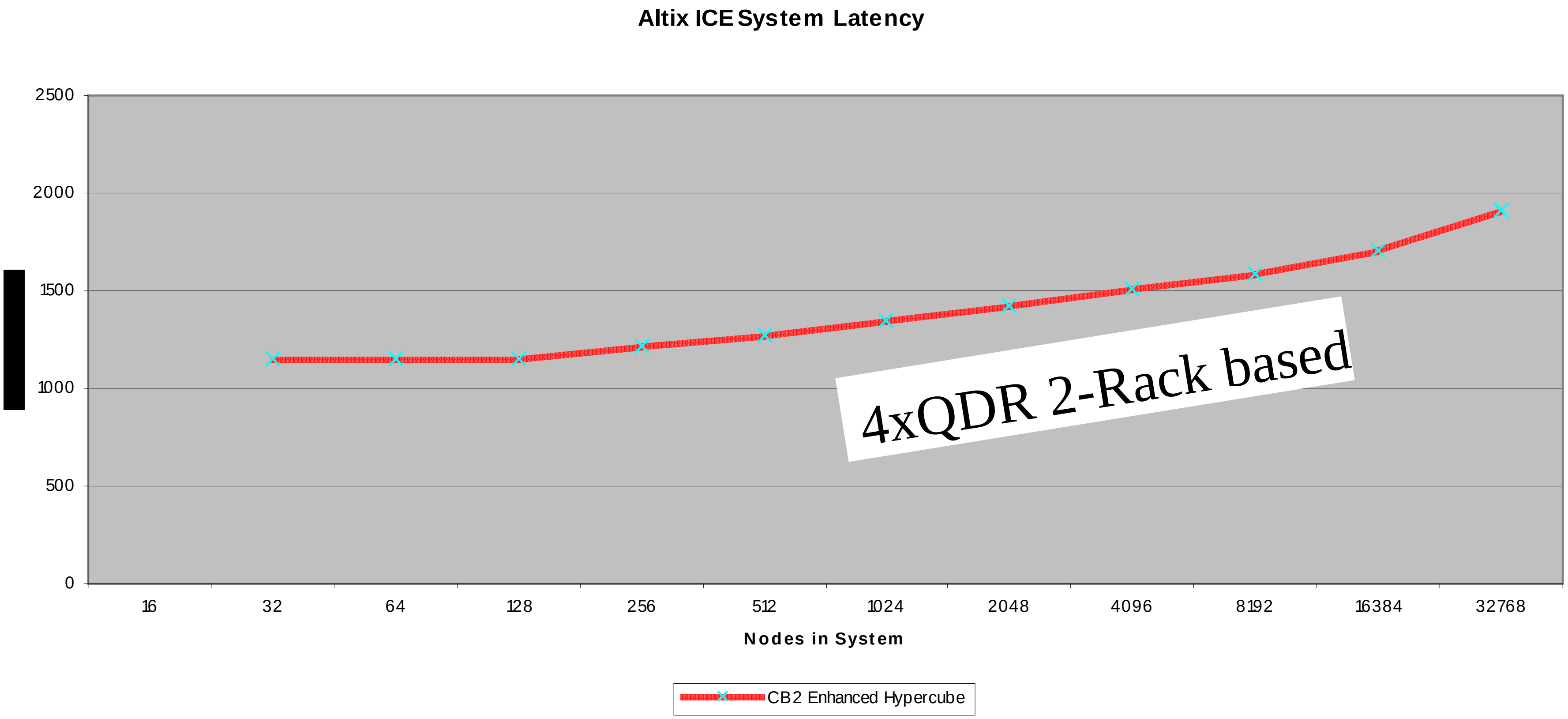




Larger configuration start from a two rack cell and form larger structures from this cell.

Doubling the number of racks rack increases the dimension of the hypercube.

Hypercube Topology Estimated MPI Latency



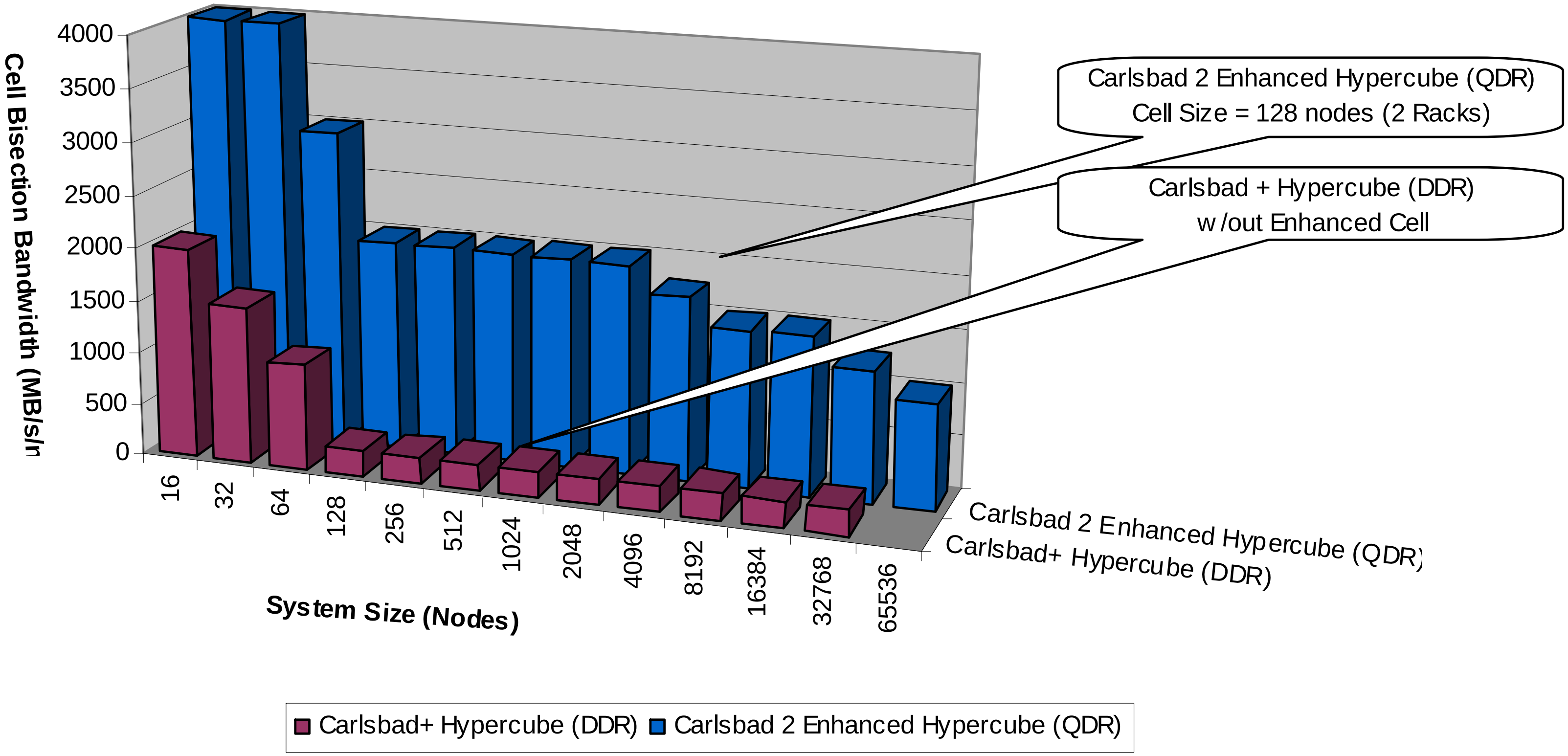
4xQDR 2-Rack based

Less than 2usec latency across the full system

In case of 4xQDR enhanced hypercube

Hypercube Topology Bisection Bandwidth

Cell Bisection Bandwidth vs. Node Count for Single Plane Configuration

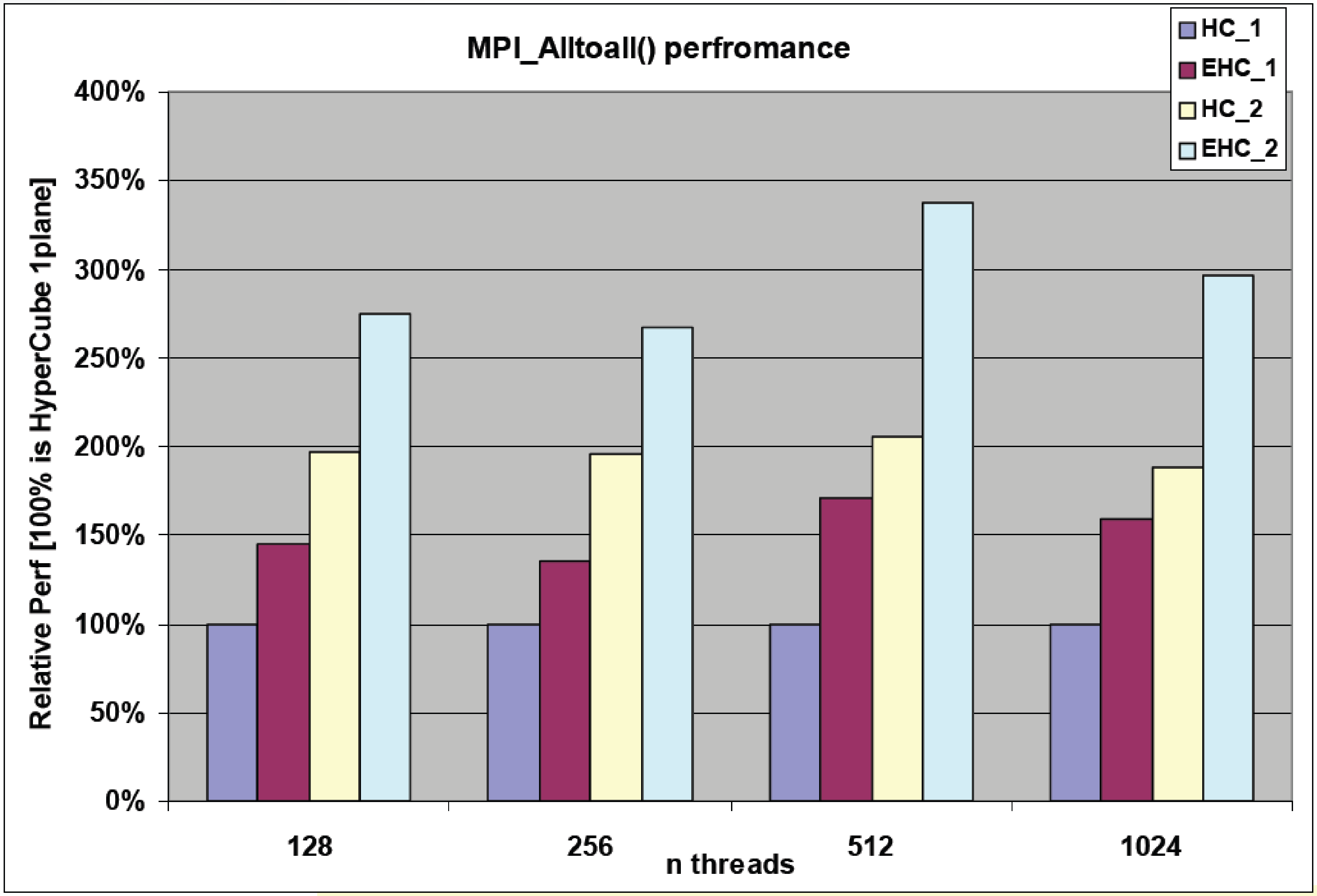


Carlsbad 2 Enhanced Hypercube (QDR)
Cell Size = 128 nodes (2 Racks)

Carlsbad + Hypercube (DDR)
w/out Enhanced Cell

Larger – 128 blade – basic cell results in significant higher bisection bandwidth

MPI All_to_All Comparison between hyper cube and enhanced hypercube cells.



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

Legend:

HC_1: single-rail hypercube

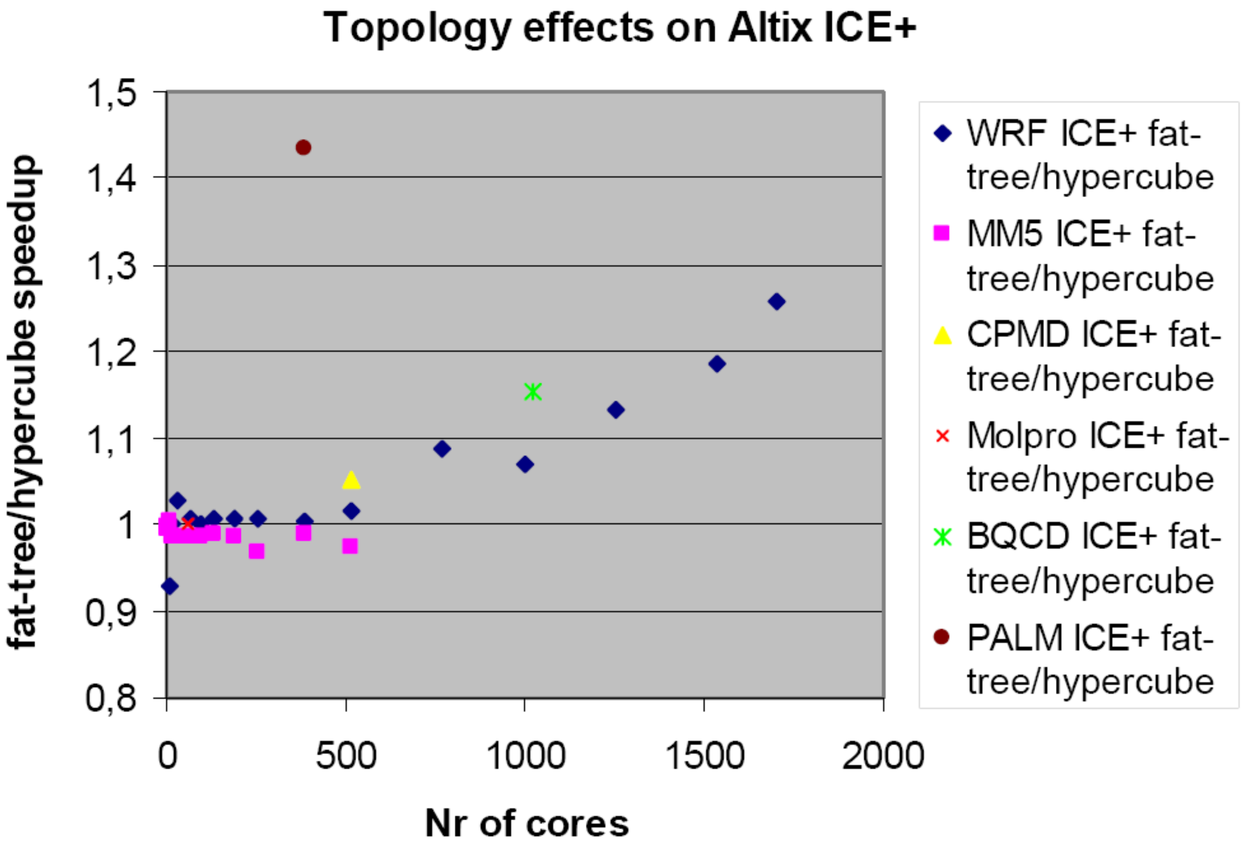
HC_2: dual-rail hypercube

EHC_1: single-rail enhanced hypercube

EHC_2: dual-rail enhanced hypercube

For MPI_Alltoall operations having more communication channels (HC-2) is more important for performance than having faster channels (EHC-1)

Application Performance on ICE8200EX in Hypercube v/s Fat-Tree Topology



- WRF, BQCD, PALM are interconnect BW sensitive

-MM5, CPMD and Molpro are interconnect latency sensitive

Summary

- SGI Altix ICE is a high performance, highly scalable compute system
- Infiniband options range from switchless hypercube topologies to Clos-Net type networks
- Hypercube topologies built around two rack building blocks offer high bandwidth and low latency

sgi®

designed. engineered. results.