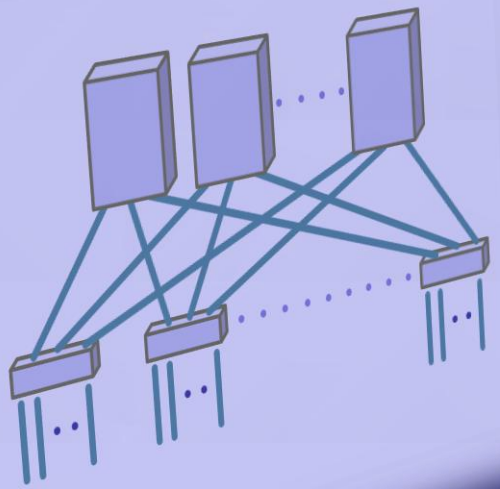
HyperTransport® CONSORTIUM

Architectural Challenges of Cluster Networks

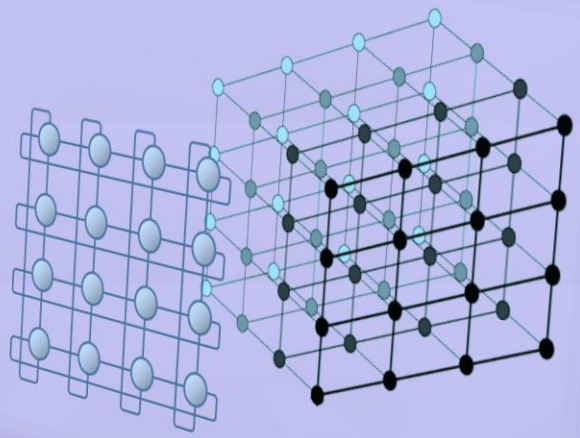
White Paper Analysis

Mainstream Network Topologies

Fat Tree



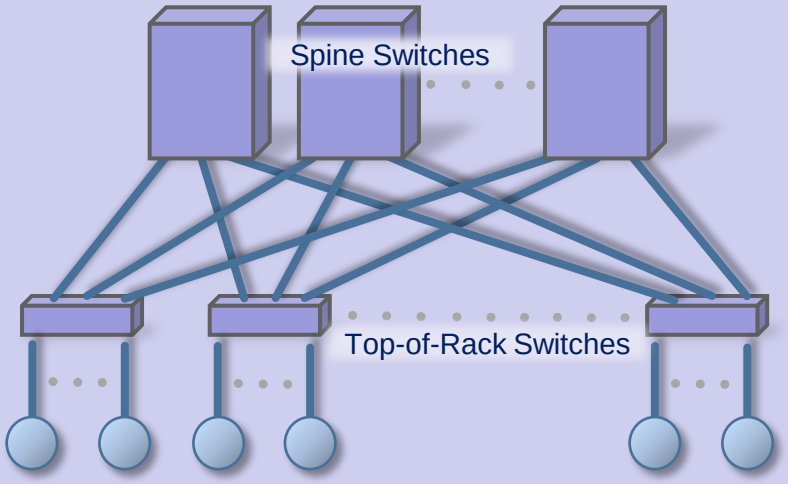
Torus



Fat Tree

Classic with Ethernet and InfiniBand

Multi-Stage Network Switching External to the Nodes

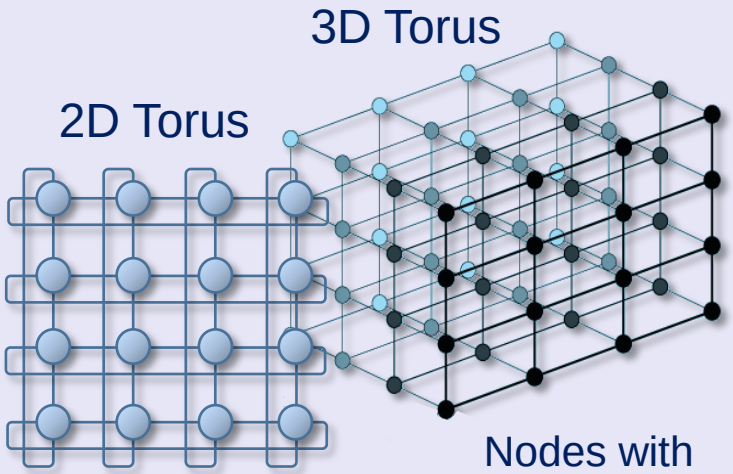


Nodes with Single-Port NICs

Torus

Classic in HPC and Supercomputing
(IBM, Cray, SGI, Etc.)

Network Switching Embedded in NICs

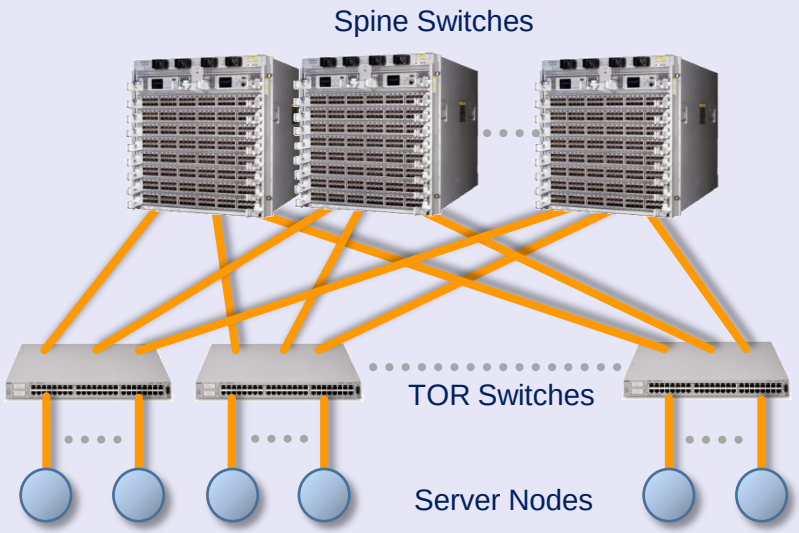


Nodes with 4-Port NICs

Nodes with 6-Port NICs

Fat Tree Cabling

Facts



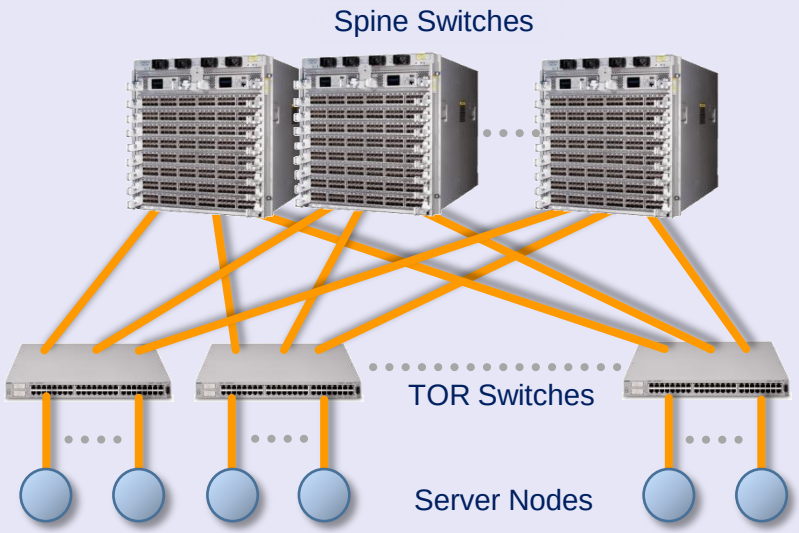
Spine Switches Interconnect with Every Top-of-Rack Switch in the Cluster, Close and Far

Scale-Out Fat Tree 10 Gb/s Ethernet Networks			
Cluster Configurations:	**Cluster 1**	**Cluster 2**	**Cluster 3**
Number of Server Nodes	576	4,224	10,752
10GbE 48-Port Switches (Stg. 1/TOR)	24	176	384
10GbE 48-Port Switches (Stg. 2/Spine)	12	-	-
10GbE 384-Port Switches (Stg. 2/Spine)	-	12	24
10GbE Cables	1,152	8,448	18,432
Network-Specific 42U Rack Enclosures	1	9	17

System Scaling Requires Cluster-Wide Network Re-Cabling

Fat Tree Cabling

CapEx



Multiple Cable Lengths, Regardless of Cluster Size

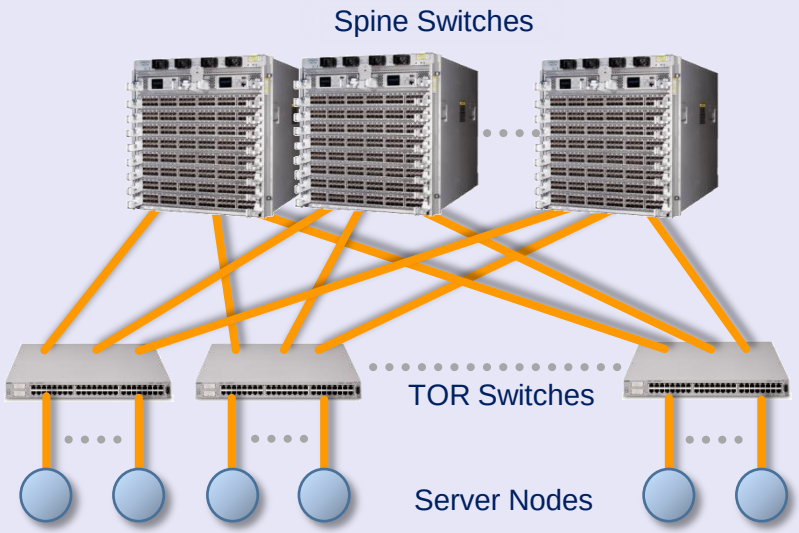
Many Long-Haul, Costly Optical Cables

Scale-Out Fat Tree 10 Gb/s Ethernet Networks CapEx					
		Cluster Nodes:	**576**	**4,224**	**10,752**
Arista	7050T-52	Switches:	$634,572	$3,102,352	$6,768,768
Arista	7508E-384	Switches:	$0	$5,744,952	$11,489,904
Mellanox	MCX311A-XCAT	NICs:	$109,440	$802,560	$2,042,880
		10GbE SFP+ Cables (Av. Lgth.):	$460,800	$3,379,200	$7,372,800
		Network-Specific 42U Rack Enclosures:	$1,420	$12,780	$24,140
		CapEx Per Node:	**$2,094**	**$3,088**	**$2,576**
		Total Network Capex:	**$ 1,206,232**	**$ 13,041,844**	**$ 27,698,492**

Multiple Cable Types and Lengths Fragmentize Quantities and Increase Purchase Price

Fat Tree Cabling

OpEx



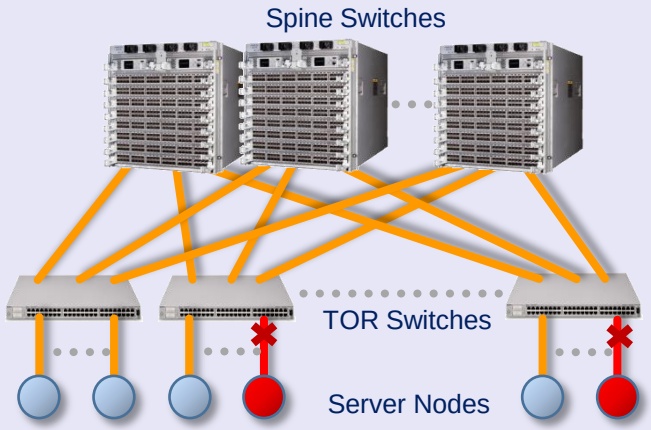
Optical Cables' Active Logic Increases Cluster's Power Consumption by 0.5W/Port, 1W/Cable

Cable's Active Logic → Reduced MTBF → Increased Cluster's Maintenance Rate and Down Time

Number of Cable Variants to be Stocked and Managed

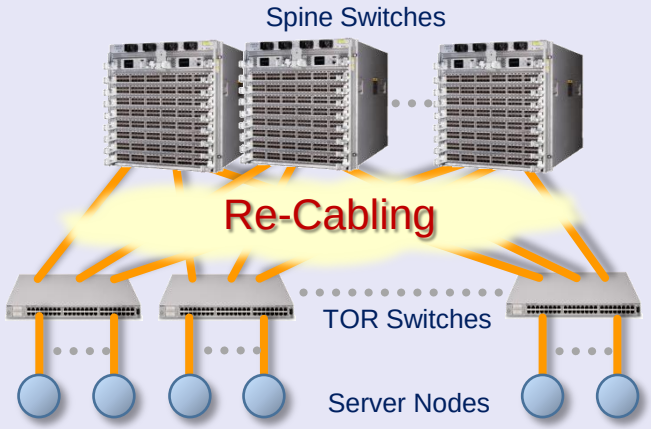
Fat Tree Cabling

Operation and Scaling



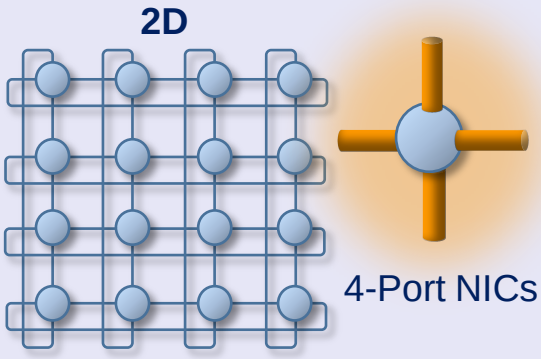
Node Cable Failures Cut-Off their Respective Server Nodes from the Cluster, Rendering them Inoperative

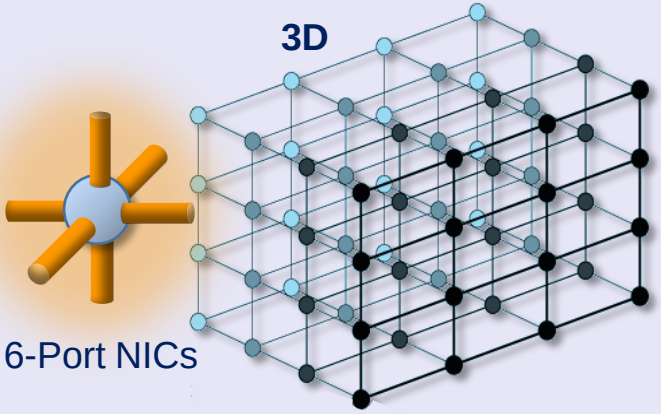
Fat Tree Networks Are Not Mission-Critical



Cluster Scaling Imposes Total Cluster Shut Down for TOR-to-Spine Switch Re-Cabling Across the Cluster → Extended Cluster Service Disruption

Torus Cabling





Facts

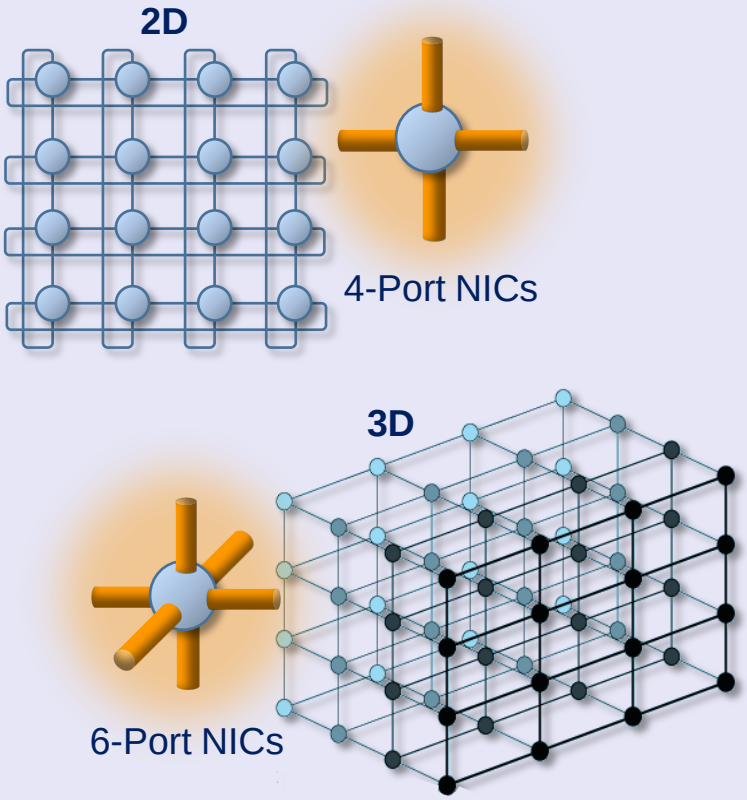
Network Switching Integrated in Each Cluster Node's NIC

Each Node Connects Directly to 2-6 Neighboring Nodes Based on Torus Topology

2D/3D Torus Networks				
Cluster Configurations:	**2D Torus**	**3D Torus 1**	**3D Torus 2**	**3D Torus 3**
Torus Nodes Geometry	**24x24**	**16x16x17**	**21x21x21**	**26x26x27**
Total Number of Server Nodes	576	4,352	9,261	18,252
External Switches	No External Switches - 100% Embedded in NIC			
Torus NIC	576	4,352	9,261	18,252
Cables	1,152	13,056	27,783	54,756
Network-Specific 42U Rack Enclosures	No Network-Specific Rack Enclosures			

Torus Cabling

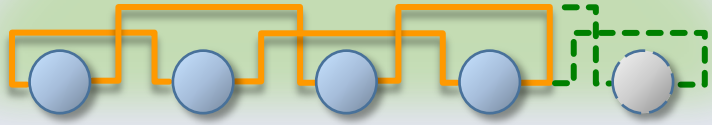
Facts (Cont.)



Same Short Lenth Cables Serve Whole Torus Cluster

Theoretical 1D Cabling Model

Practical 1D Cabling Model



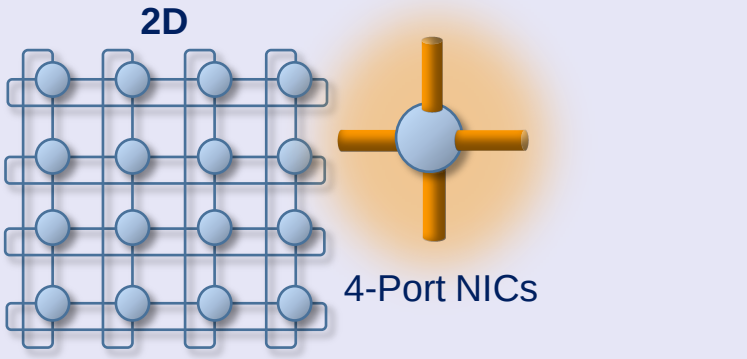
Each Node at Same Number of Hops from Any other Node like the Theoretical Model

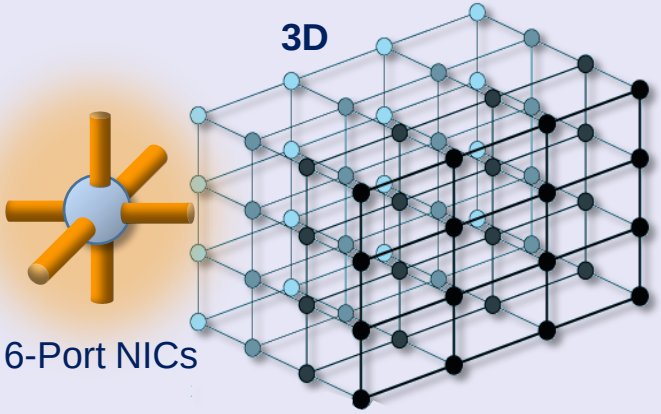
Incremental System Scaling

Same Configuration Applies to 2D and 3D Torus Y and Z Axis

Torus Cabling

CapEx





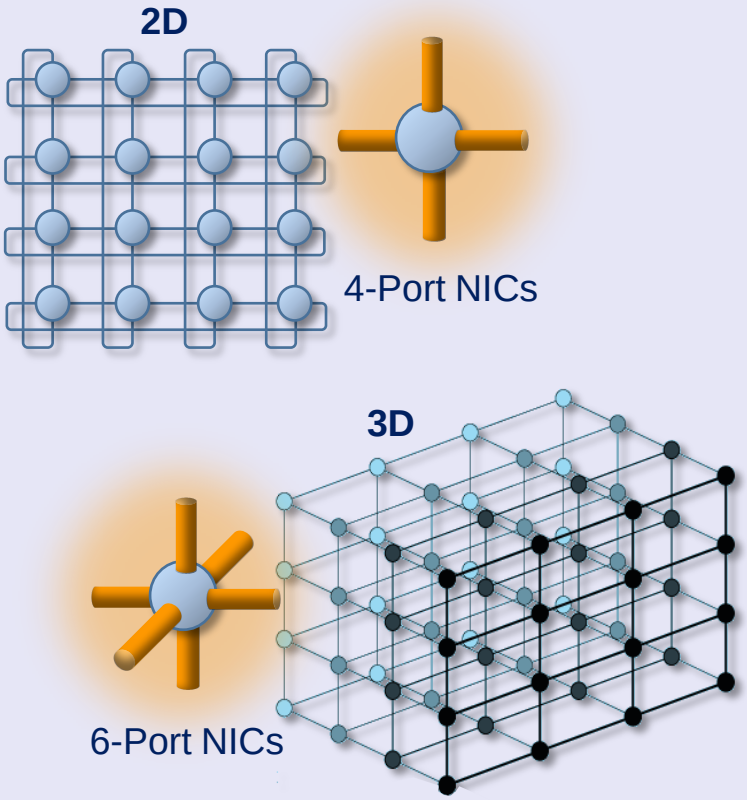
Only Short Haul, Passive (Copper) Cables, Regardless of Cluster Size and Topology (1D, 2D, 3D, Etc.)

Single Cable Length Optimizes Cabling Quantities and Purchase Pricing

2D/3D Torus Network CapEx				
Cluster Nodes:	**576**	**4,352**	**9,261**	**18,252**
External Switches	No External Switches - 100% Embedded in NICs			
Torus NICs	$230,400	$2,176,000	$4,630,500	$9,126,000
Cables (Avg. Length 1.5m)	$69,120	$130,560	$639,009	$1,259,388
Network-Specific 42U Rack Enclosures	No Network-Specific Rack Enclosures			
CapEx Per Node:	**$520**	**$530**	**$569**	**$569**
Cluster Total:	**$ 299,520**	**$ 2,306,560**	**$ 5,269,509**	**$ 10,385,388**

Torus Cabling

OpEx



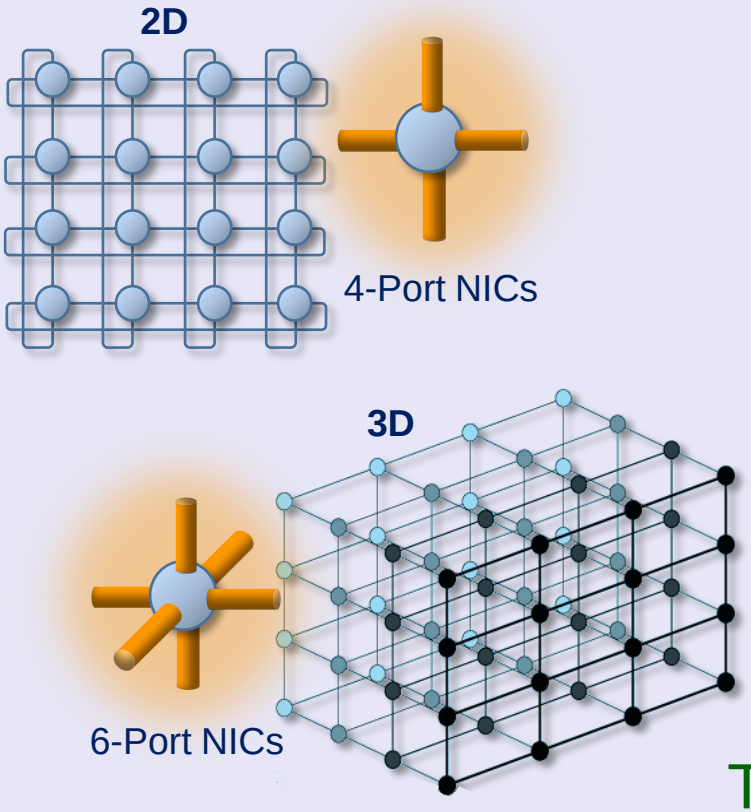
No Cabling Active Logic (Just Copper Cables) → No Power Consumption

No Active Logic → No Reduction of Cluster MTBF → No Increase in Cluster System's Maintenance Nor Down Time

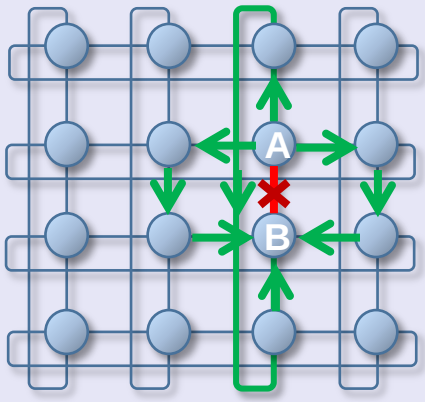
Single Cable Length Minimizes Stocking and Management Costs

Torus Cabling

Operations



Individual Cable Failures Do Not Isolate/Cut-Off Server Nodes. Affected Nodes Stay Connected with Rest of Cluster via Alternate Cable Links (3x/Node in 2D, 2-5x/Node in 3D) and Plenty Alternate Routes:

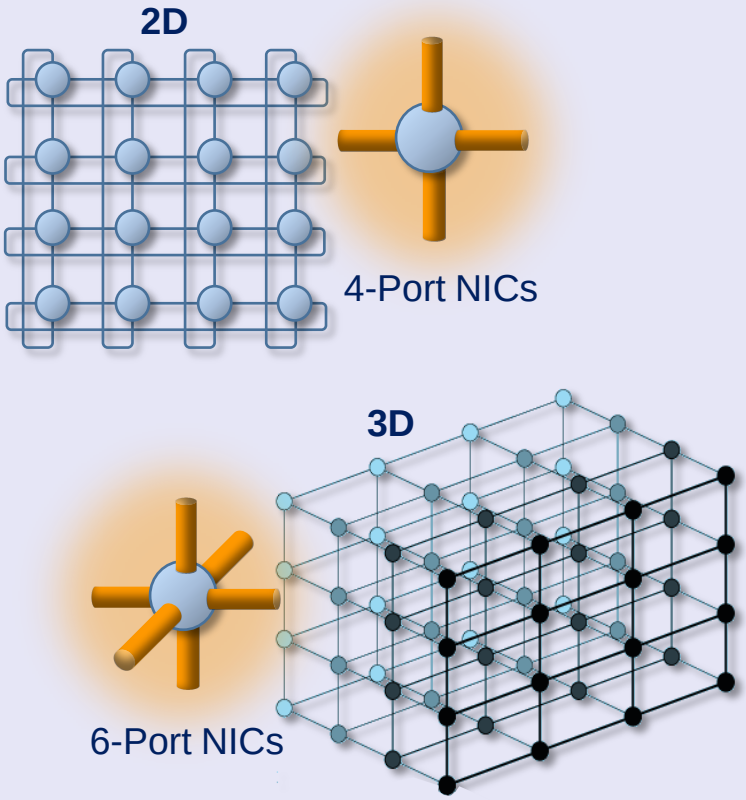


3x 3 Hops (Depicted)
5x 5 Hops
Etc.

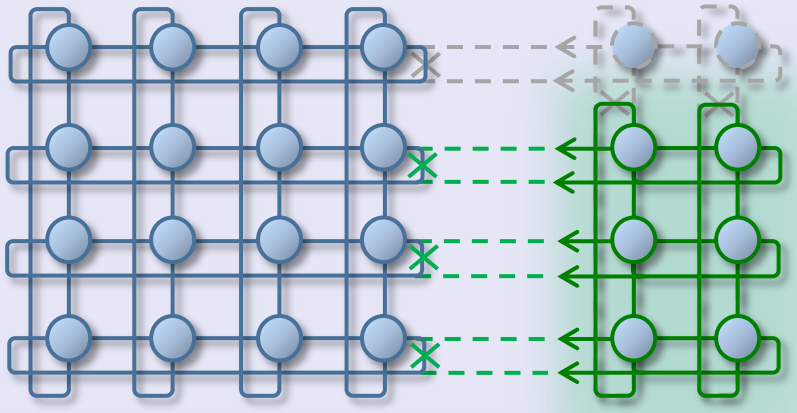
Torus Clusters Are Mission-Critical

Torus Cabling

Scaling



Incremental, Granular System Scaling Involves Just Specific Nodes of Specific Torus Network Outer Planes



Expansion Cluster Can be Built Aside and then Rapidly Connected

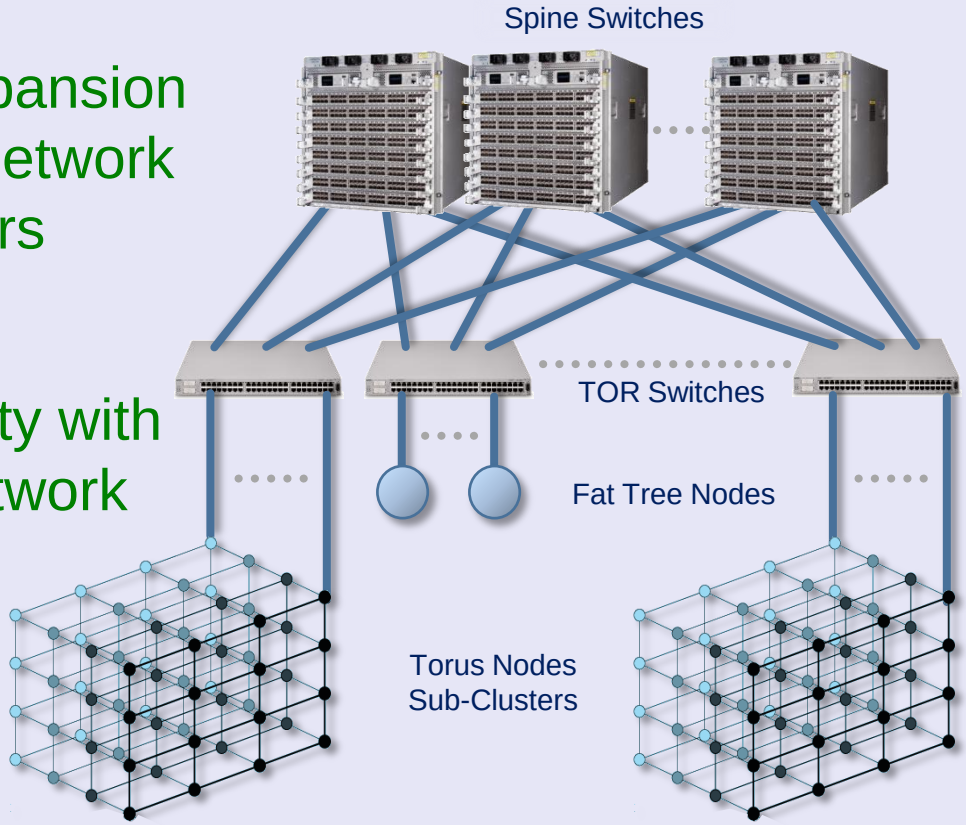
Main Cluster Remains Fully Operational

Same Configuration Applies to 2D and 3D Torus Y and Z Axis

Progress with Investment Preservation

Gradual Expansion Via Torus Network Clusters

Interoperability with Existing Network

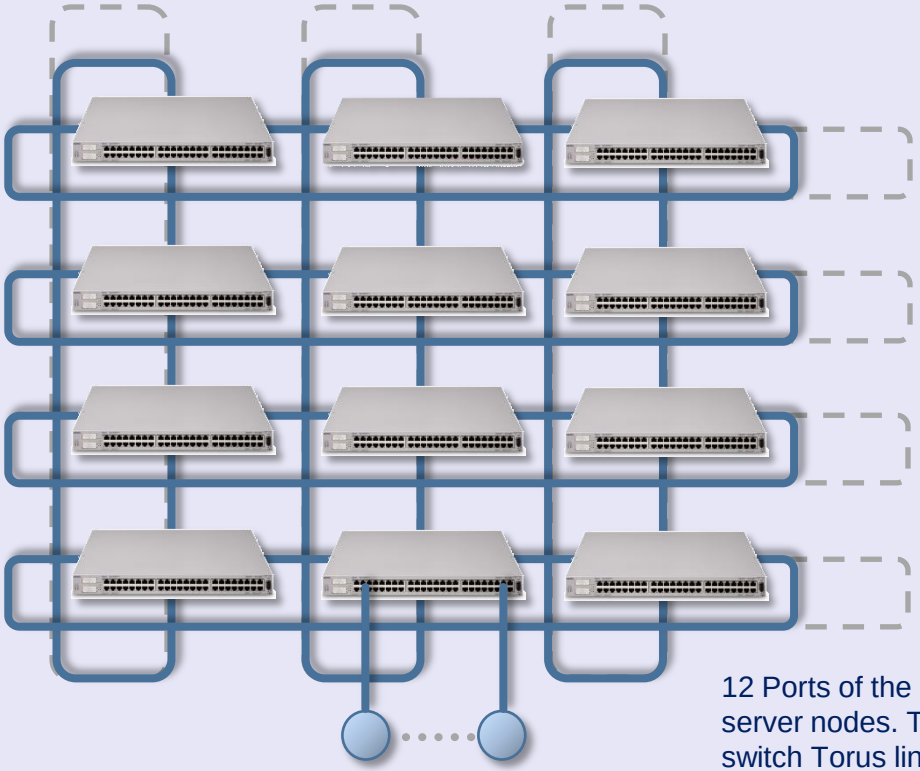


Removed Intrinsic Fat Tree Scalability Limitations

Incremental Latency Performance Improvement

Progress with Investment Preservation (cont.)

2D Torus-Linked 24-Port Fat Tree Switches Example



12 Nodes per Switch

Preserved Investment in TOR Switches

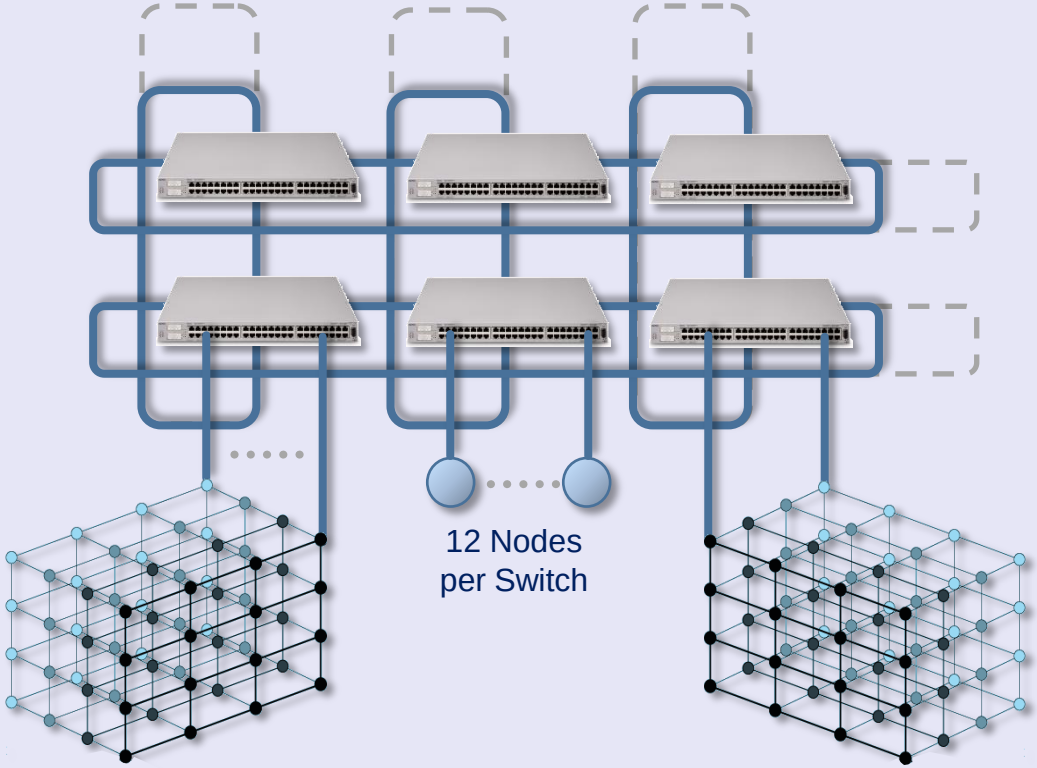
Limitless Cluster Scalability

Improved Latency Performance

12 Ports of the 24-Port Ethernet switches connect to server nodes. The other 12 ports serve the switch-to-switch Torus links, with 3 Ethernet Cables per link for adequate bandwidth

Progress with Investment Preservation (cont.)

2D Torus-Linked 24-Port Fat Tree Switches Example



Preserved Investment in TOR Switches

Limitless Cluster Scalability

Significantly Improved Latency Performance

Click Here to Learn Why

12 Ports of the 24-Port Ethernet switches connect to server nodes and/or to Torus sub-clusters. The other 12 ports serve the switch-to-switch Torus links, with 3 Ethernet Cabes per link for adequate bandwidth

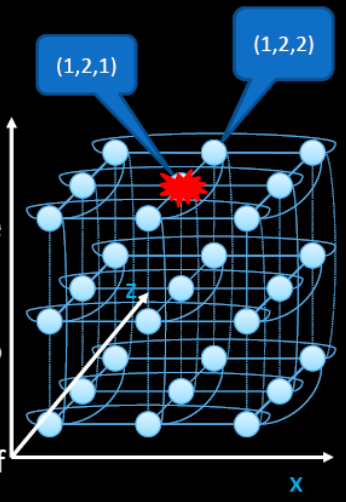
Torus is Target of Leading Technology Players

Microsoft Research Cambridge

CamCube : A novel data center

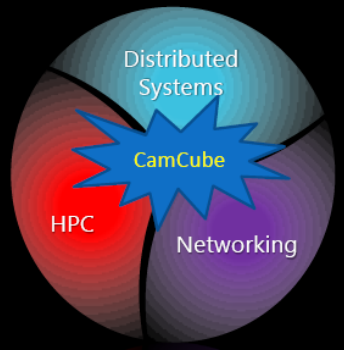
Why use a 3D torus?

- The CamCube API
  - Nodes have **(x,y,z)** coordinate
    - Defines key-s
  - Simple one hop API to
    - send/receive packets to/from 1-hop neighbours
    - Key or address-based
  - Coordinates re-mapped in case of failures
- Enable KBR-like API
  - Merge physical and virtual topologies



What's different?

- High Performance Computing
  - MPI hides the underlying topology
    - CamCube makes topology explicit
    - Services can intercept packet on-path
  - Failure resilience
  - Multiple independent services
- Distributed systems
  - Key-space naturally mapped on the physical topology
  - No need to "reverse-engineer" the network
- Networking
  - No switches or routers (symmetry of role)
  - Cross-layer approach
  - Not using TCP/IP



Major Technology Players Already Invested in Torus

Energy Proportional Datacenter Networks

Google

ABSTRACT

Numerous studies have shown that datacenter computers rarely operate at full utilization, leading to a number of proposals for creating servers that are *energy proportional* with respect to the computation that they are performing. In this paper, we show that as servers themselves become more energy proportional, the datacenter network can become a significant fraction (up to 50%) of cluster power. In this paper we propose several ways to design a high-performance datacenter network whose power consumption is more proportional to the amount of traffic it is moving—that is, we propose *energy proportional datacenter networks*.

We first show that a flattened butterfly topology itself is inherently more power efficient than the other commonly proposed topology for high-performance datacenter networks. We then exploit the characteristics of modern plesiochronous links to adjust their power and performance envelopes dynamically. Using a network simulator, driven by both synthetic workloads and production datacenter traces, we characterize and understand design trade-offs, and demonstrate an 85% reduction in power — which approaches the ideal energy-proportionality of the network.

Our results also demonstrate two challenges for the designers of future network switches: 1) We show that there is a significant power advantage to having independent control of each unidirectional channel comprising a network link, since many traffic patterns show very asymmetric use, and 2) system designers should work to optimize the high-speed channel designs to be more energy efficient by choosing optimal data rate and equalization technology. Given these assumptions, we demonstrate that energy proportional datacenter communication is indeed possible.

Categories and Subject Descriptors

C.2.1 [Network Architecture and Design]: Network communications; Topology; B.4.3 [Interconnections]: Fiber optics; Topology

General Terms

Performance, Design

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ISCA'10, June 19–23, 2010, Saint-Malo, France.

Copyright 2010 ACM 978-1-4503-0053-7/10/06...$10.00.

Keywords

Low-power networking, Datacenter networks, Interconnection networks

1. INTRODUCTION

The cost of power and its associated delivery and cooling are becoming significant factors in the total expenditures of large-scale datacenters. Barroso and Hölzle recently showed a mismatch between common server workload profiles and server energy efficiency [3]. In particular, they show that a typical Google cluster spends most of its time within the 10-50% CPU utilization range, but that servers are inefficient at these levels. They therefore make the call for *energy proportional computing systems* that ideally consume almost no power when idle and gradually consume more power as the activity level increases.

Servers and their processors are the obvious targets to improve energy proportionality because they are today's primary power consumers. Today's typical multi-tiered datacenter network consumes little power, relative to servers, because of its high degree of *oversubscription*. As an example of over-subscription, machines connected to the same rack switch (i.e., the first tier) have significantly more bandwidth to each other than to machines in other racks. The level of bandwidth over-subscription is typically an order of magnitude or more for each subsequent tier.

Figure 1: Comparison of server and network power

338

"Our goal is to provide a network that supports energy proportional communication. That is, the amount of energy consumed is proportional to the traffic intensity (offered load) in the network."

"We propose the flattened butterfly (FBFLY) topology as a cornerstone for energy-proportional communication in large scale clusters with 10,000 servers or more."

"A flattened butterfly is a multi-dimensional direct network, in many ways like a torus."

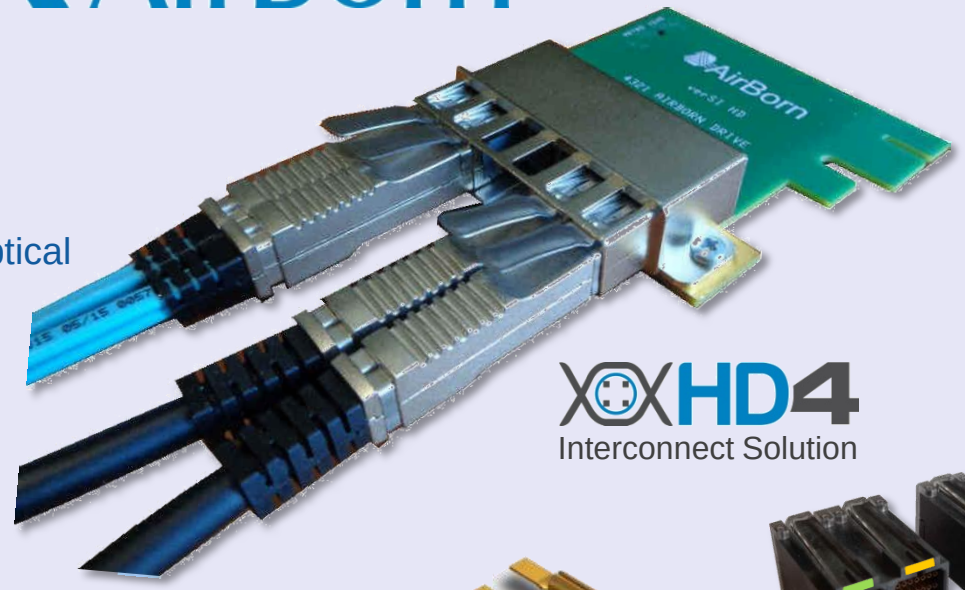
"A datacenter network based on the flattened butterfly topology results in a more power-efficient network and therefore lowers operational expenditures."

"The topology can take advantage of packaging locality in the sense that nodes which are in close physical proximity can be cabled with inexpensive copper cables."

Excerpts from Google White Paper "Energy Proportional Datacenter Networks", 2011

Commercial Products Enabling Torus Networks

AirBorn



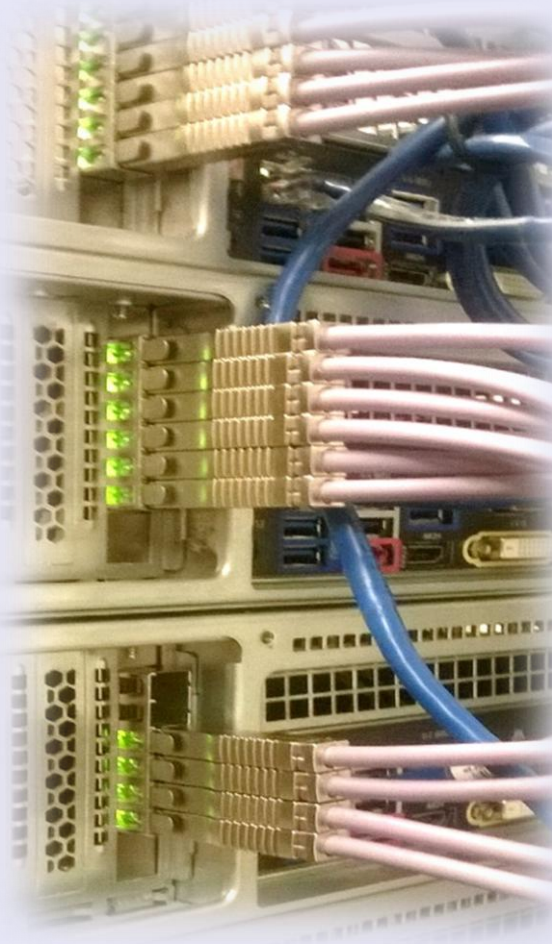
Optical

XCXHD4
Interconnect Solution

Copper

High Reliability
4 Points of Contact
Spring Contact Pins

LED Light Guide
for Easy Link Diagnosis

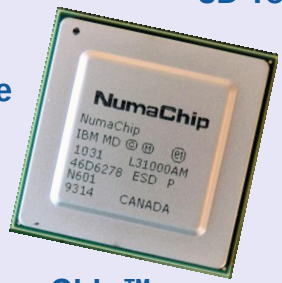


Commercial Products Implementing Torus Networks

NUMASCALE
BIGGER DATA ANALYTICS

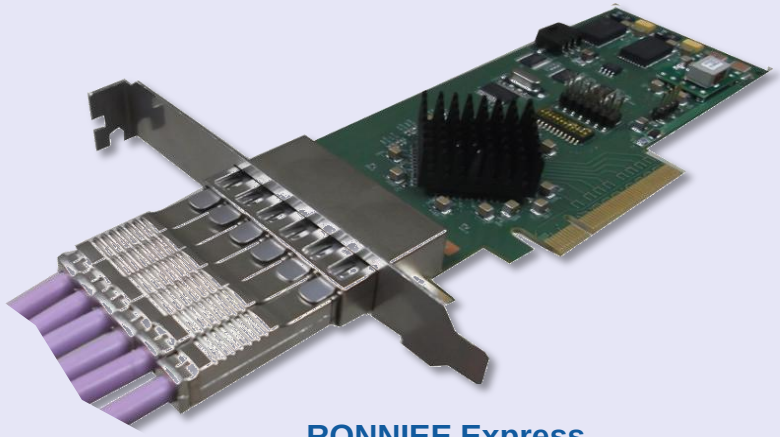
3D Torus
Shared Memory Interconnect
In-Memory Analytics Appliance

NumaConnect™
3D Torus Adapter



NumaChip™
Scalable, Cache Coherent, Shared Memory
3D Torus Interconnect Controller Chip

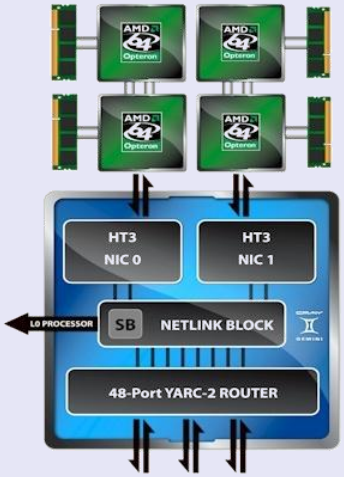
a³cube



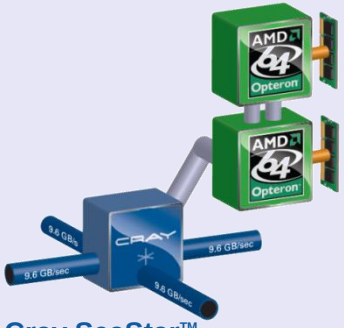
RONNIEE Express
3D Torus
Interconnect Fabric

Torus Networks as Backbone of Supercomputers

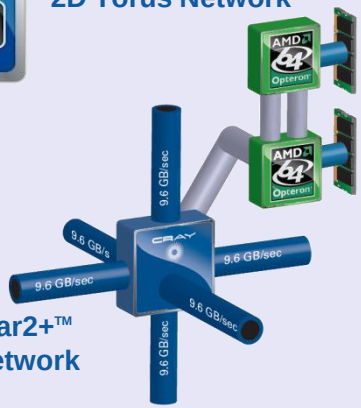
CRAY



Cray Gemini™
3D Torus Network



Cray SeaStar™
2D Torus Network

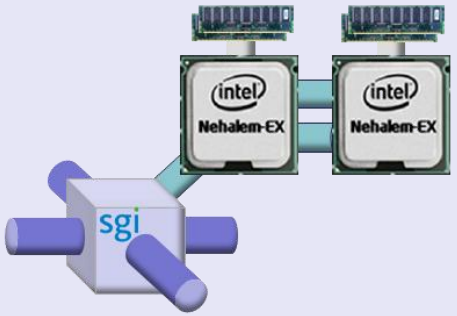


Cray SeaStar2+™
3D Torus Network

IBM

Blue Gene/P
Peta-Scale Supercomputer
73,728 Nodes
Scalable to 221,184 Nodes

sgi



SGI NUMALink®
2D Torus Network
(2009)

Torus Topology Benefits Score Card

Drastic CapEx Reduction

No External Switches
Simplified Cabling

Drastic OpEx Reduction

No External Switching Power Consumption

Significant Latency Performance Improvement

No External Switching Latency

Always Up Operation - Mission-Critical

No External Switches to Fail
No Cable Failure Forcing Equipment Down Time
No Down Time for System Expansion

Unlimited, Highly Granular Scalability