Cisco HPC and Infiniband solutions

ICFA DDW06

Dré Van Brussel

Sr Mgr Business Development
Data Centre & HPC Technologies
Cisco Emerging Markets - CEE

October 2006

Data Center challenges : budget vs capability

Business and Technology Scenario

To prevent the data center from consuming the entire IT budget,

increased **manageability** and **utilization**

through **standardization** and **automation**

are essential.

Take Aways

HPC Maps into Cisco's Core Area of Strength

Shift to server clusters; Increased requirement for multiple networks (IPC+Mngt), Storage convergence, Enterprise class management and security

SFS and Catalyst *combined* provide most complete HPC clustering solution

SFS extends Cisco into performance sensitive environments; Enables high performance storage over IPC network; Catalyst for broad market and management of SFS clusters

Cisco is leading manufacturer of InfiniBand clustering equipment

Hundreds of clusters running across automotive, aerospace, banking, media, oil & gas; Hands on experience building World's largest standard server (and InfiniBand) cluster (Sandia Thunderbird; 4400 SFS attached servers)

Cisco Server Fabric Switching provides most complete InfiniBand HPC solution

Family of four SFS 7000 switches supporting clusters from tens to thousands of servers; host adapters&drivers; HPC management software and diagnostics

Only Network Manufacturer with Global Channels, Expertise, Support for HPC

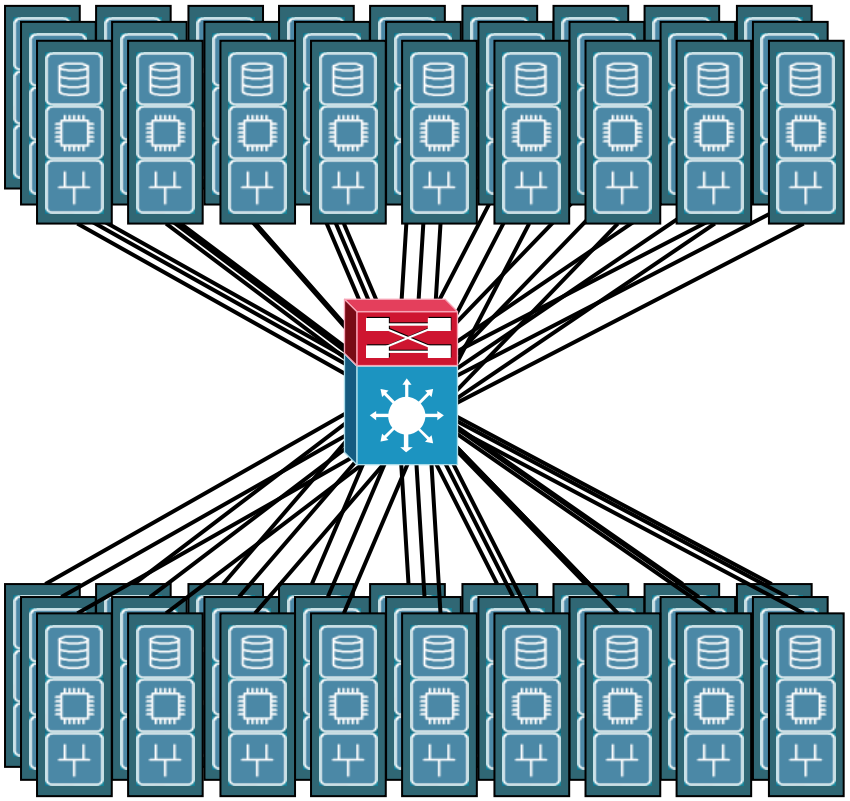
Joint Sourcing and Engagements with Server Vendors; Internal Cluster lab; Best Practicing Guides; 2000 tech support professionals; >300,000 customer issues resolve every month

Take Aways
With Ongoing Scalability Testing

- Cisco has nine month lead time with large IB clusters
  - Sandia Thunderbird 4400 node cluster built in June'05
  - One of three >1000 node IB clusters Cisco built in '05
- Cisco is only vendor with access to large IB cluster today
  - Daily remote and direct access to 4400 node cluster
  - Enables continuous improvement of HPC Subnet Manager
  - Scalability testing with LUSTRE and IB attached storage
- Cisco has ~1000 InfiniBand nodes inside it's own test labs
  - Cisco's developers in US, EMEA and Asia have access around the clock
- Cisco is directly funding the open source community to increase stability of large scale cluster
  - Two grants already made in 2006 for MVAPICH development
- Cisco is already developing unified management across large scale InfiniBand and Ethernet connected clusters

Network Computing is Supercomputing

- Supercomputer performance at a fraction of the cost
- Clusters of Industry standard servers with Scalable, Standards-based Network interconnect
  - [Gigabit] Ethernet
  - Infiniband
- Demands on the Network :
  - Storage volumes measured in Peta-Bytes of Data
  - Real-time Visualization in the 100s of Mbps
  - CPU efficiency can demand network latency of <10µs latency



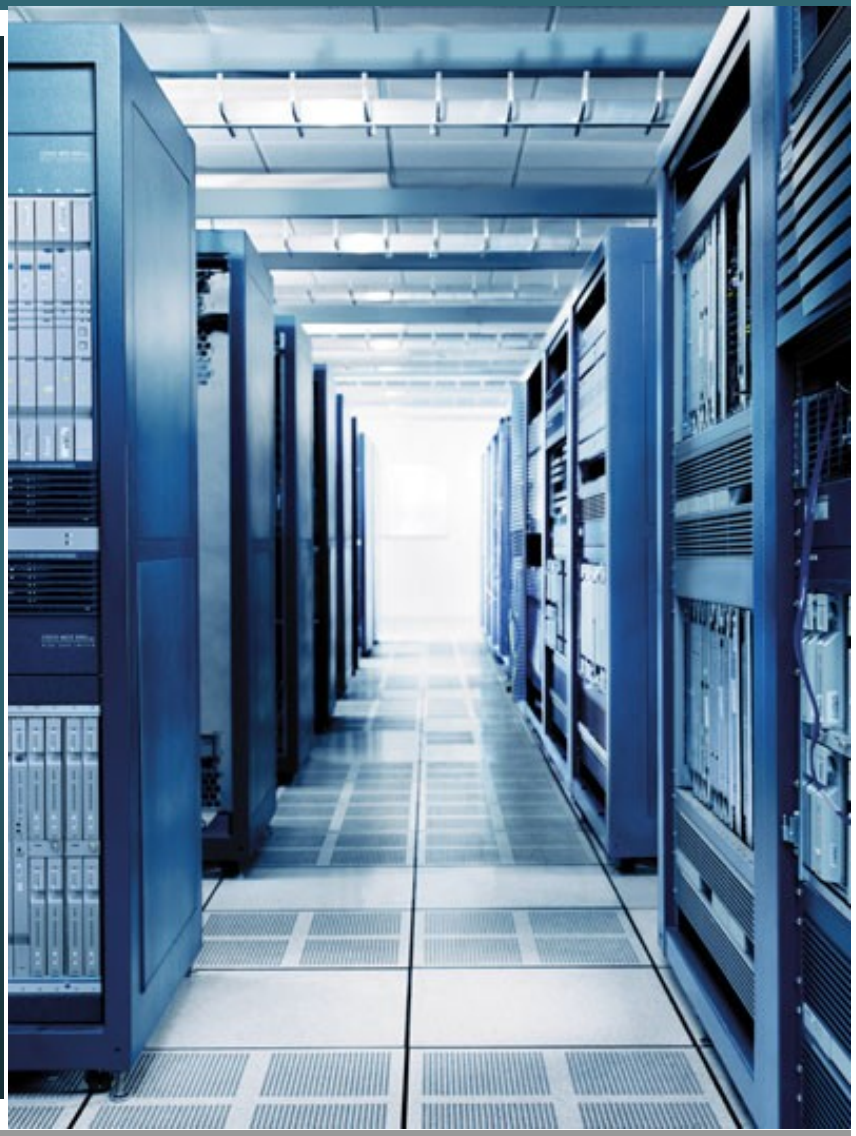
Networked Supercomputer

Why Infiniband?

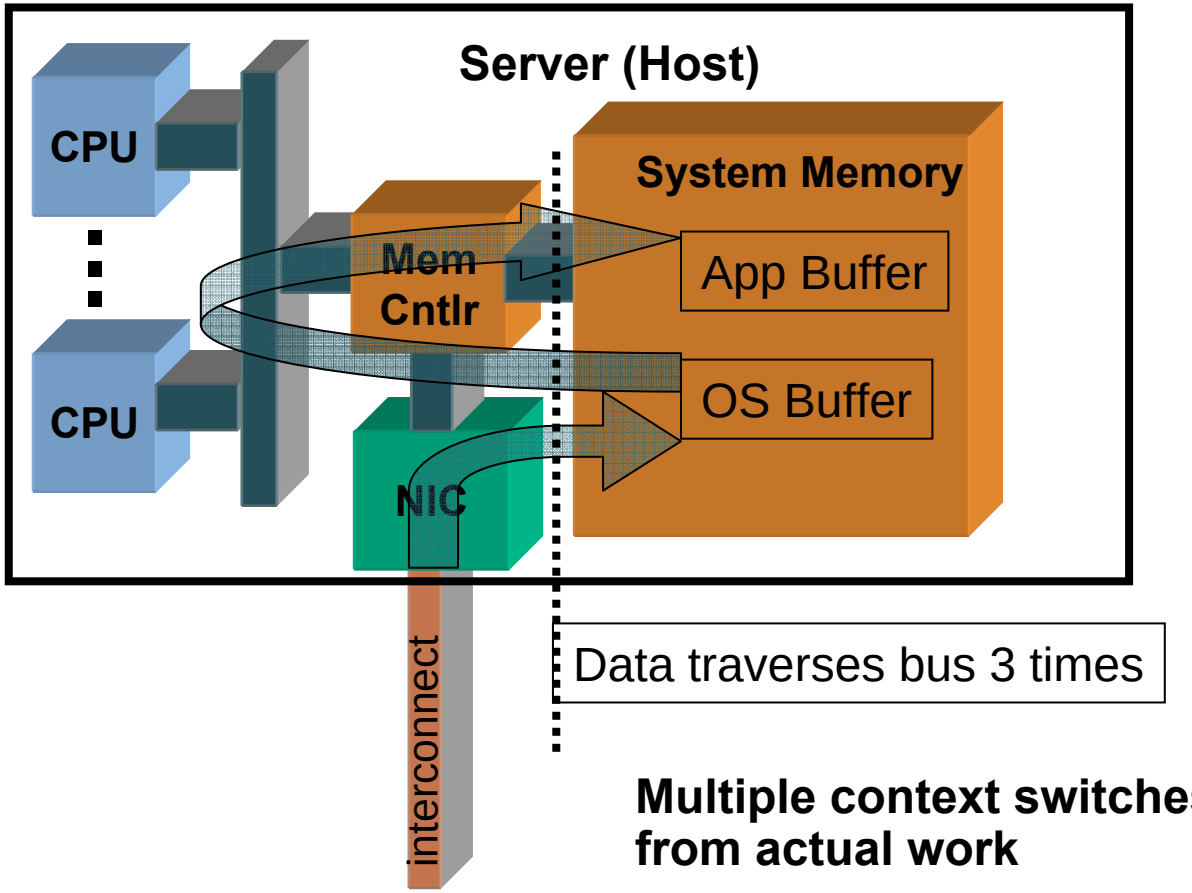
Technology Benefits	• Standards-based high speed performance capabilities • Economics for server consolidation and virtualization • Appeals to decision makers who are looking for server I/O optimized technology
New Markets	• Extends Cisco's solution for HPC and grid computing space to performance-sensitive applications • New connectivity option appeals to server infrastructure technologist
Customer Demand	• Broad deployment across enterprise, research and geography (auto, aerospace, retail, financial) in addition to traditional university and lab deployment • Enterprise customers are in pilot and production deployment phases

Infiniband Overview

- Standards-based interconnect
  - http://www.infinibandta.org
- Channelized, connection-based interconnect, optimized for high performance computing
- Supports server and storage attachments
- Bandwidth Capabilities (SDR/DDR/QDR)
  - 1x—2.5/5 Gbps: 2/4 Gbps actual data rate (base rate for InfiniBand)
  - 4x—10/20 Gbps: 8/16 Gbps actual data rate
  - 12x—30/60 Gbps: 24/28 Gbps actual data rate
- Built-in RDMA as core capability for inter-CPU communication



Standard Ethernet NIC Architecture

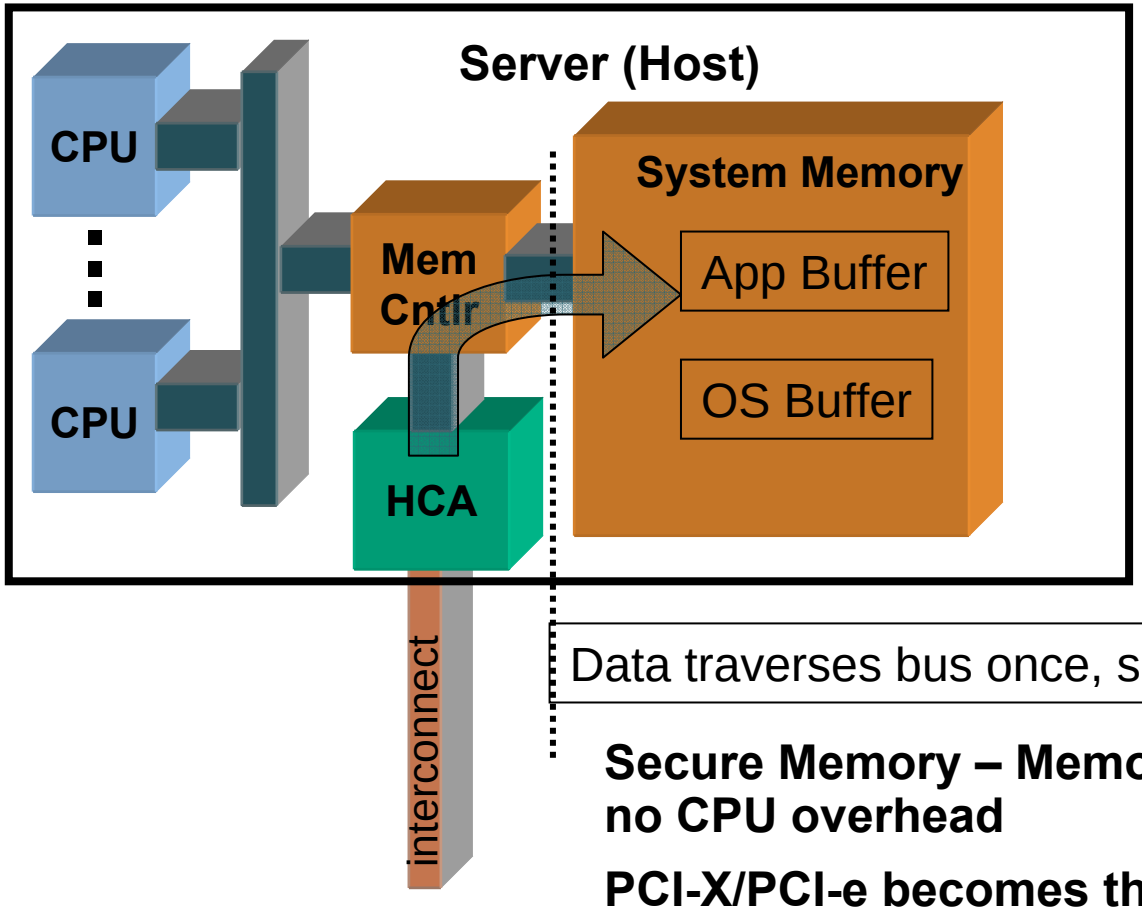


Multiple context switches robs CPU cycles from actual work

Memory bandwidth and per packet interrupts limit max throughput

OS manages end-to-end communications path

With RDMA and OS Bypass



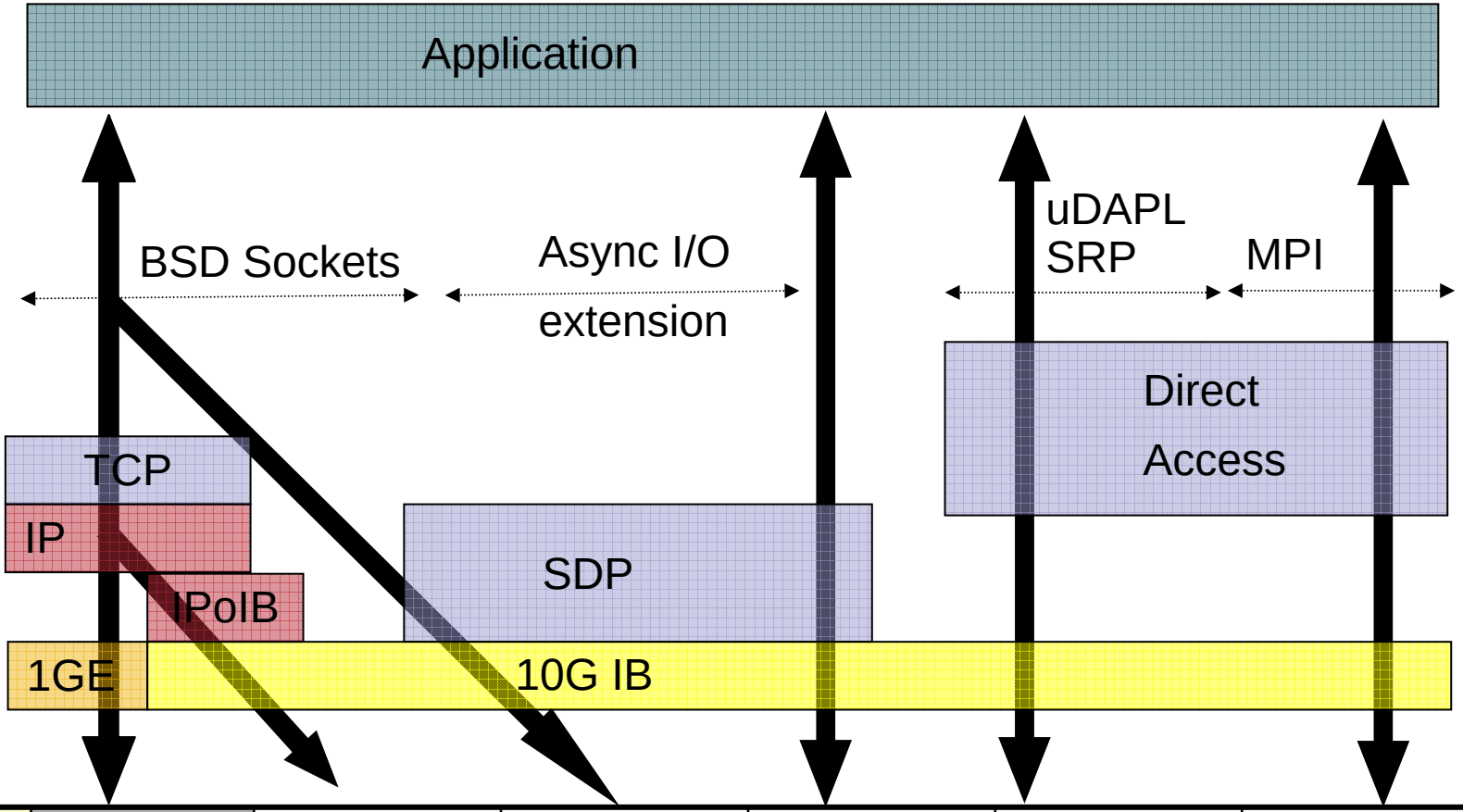
Data traverses bus once, saving CPU and memory cycles

Secure Memory – Memory transfers with no CPU overhead

PCI-X/PCI-e becomes the bottleneck for network data transmission

HCA manages remote data transmission

InfiniBand Performance
Measured Results



Throughput	**1-3 Gb/s**	**4.1Gb/s**	**4.5 Gb/s**	**7.9 Gb/s**	**8 Gb/s**	**8 Gb/s**
Latency	**40-60 usec**	**30 usec**	**18 usec**	**18 usec**	**8 usec**	**5 usec**

Ethernet vs Infiniband in HPC Clusters: typical Performance Results

	procs	nodes	time (hours)	speedup
	1 proc (estimated)		*close to 2 days*	
	8	8	6:45:00	1
	16	8	3:54:10	1.7 / 2
	16	16	3:33:05	1.9 / 2
Ethernet	32	16	2:12:06	3.1 / 4
	32	32	2:06:50	3.2 / 4
	64	32	1:30:07	4.5 / 8
	64	64	1:25:22	4.7 / 8
	96	96	1:45:20	3.8 / 12
	128	128	2:03:41	3.2 / 16

	procs	nodes	time (hours)	speedup	IB/Eth. Ratio
	8	8	6:32:34	1	1,03
	16	8	3:41:16	1.8 / 2	1,06
	16	16	3:20:40	2.0 / 2	1,06
	32	16	1:52:41	3.5 / 4	1,17
Infiniband	32	32	1:40:47	3.9 / 4	1,26
	64	32	1:01:39	6.4 / 8	1,46
	64	64	0:55:50	7.3 / 8	1,53
	96	48	0:47:31	8.3 / 12	
	96	96	0:42:13	9.3 / 12	2,50
	128	128	0:33:02	11.9 / 16	3,74

From days to minutes .

IB outperforms Ethernet from 3 to 374%

18

Introducing the Cisco Server Fabric Switching Products

Server Fabric Switch

InfiniBand

SFS 7000p

(24) 4XIB

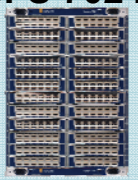
SFS 7008p

(96) 4XIB

SFS 7012

(144) 4XIB

SFS 7024



(288) 4XIB

Multifabric

SFS 3001

(12) 4XIB + 1 Gw

SFS 3012

(24) 4XIB + 12 Gws

(2) 2G FC GW

(6) GE GW

Blade Server

IBM BladeCenter

- HCA (2) 1XIB PCI-X
- Embedded switch (14) 1XIB (Internal) + (1) 4XIB and (1) 12XIB (External)

Dell 1855

- HCA (2) 4XIB PCI-ex
- Passthru Module (10) 4XIB

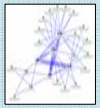
HCA

- (2) 4XIB PCI-X
- (2) 4XIB PCI-E
- (1) 4XIB PCI-E

- Remote Boot
- Linux Host Driver
- Windows Host Driver

Subnet Mgt

- High Performance Subnet Manager Software

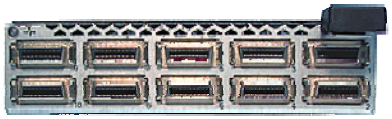


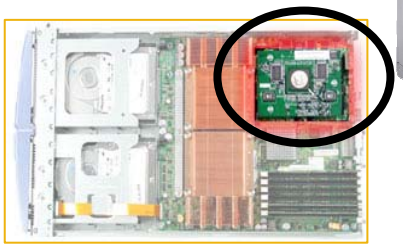
- Embedded Subnet Manager

DDR : Cisco SFS 7000D Infiniband Switches

- Cisco SFS 7000D DDR and SDR switches delivery low-latency, non-blocking fabric
  - SFS 7000D – 24 ports
  - SFS 7012D – 144 ports
  - SFS 7024D – 288 ports
- High availability and serviceability
- DDR delivers 20Gbps bandwidth to x86/Opteron PCI-Express systems
- Lower end-to-end *systems* latency
- Future-proof investment
  - backwards compatibility using DDR-to-SDR switching
- Ideal Technology for demanding high-performance applications

Blade Chassis Offerings – some examples





IBM BladeCenter / BC-H

- Integrated Switch Modules for BC1 & BC-H
- 1x / 4x Host Channel Adapter daughter-cards for blades
- Only on IBM GPL – manufactured by Cisco

Dell Blade Chassis

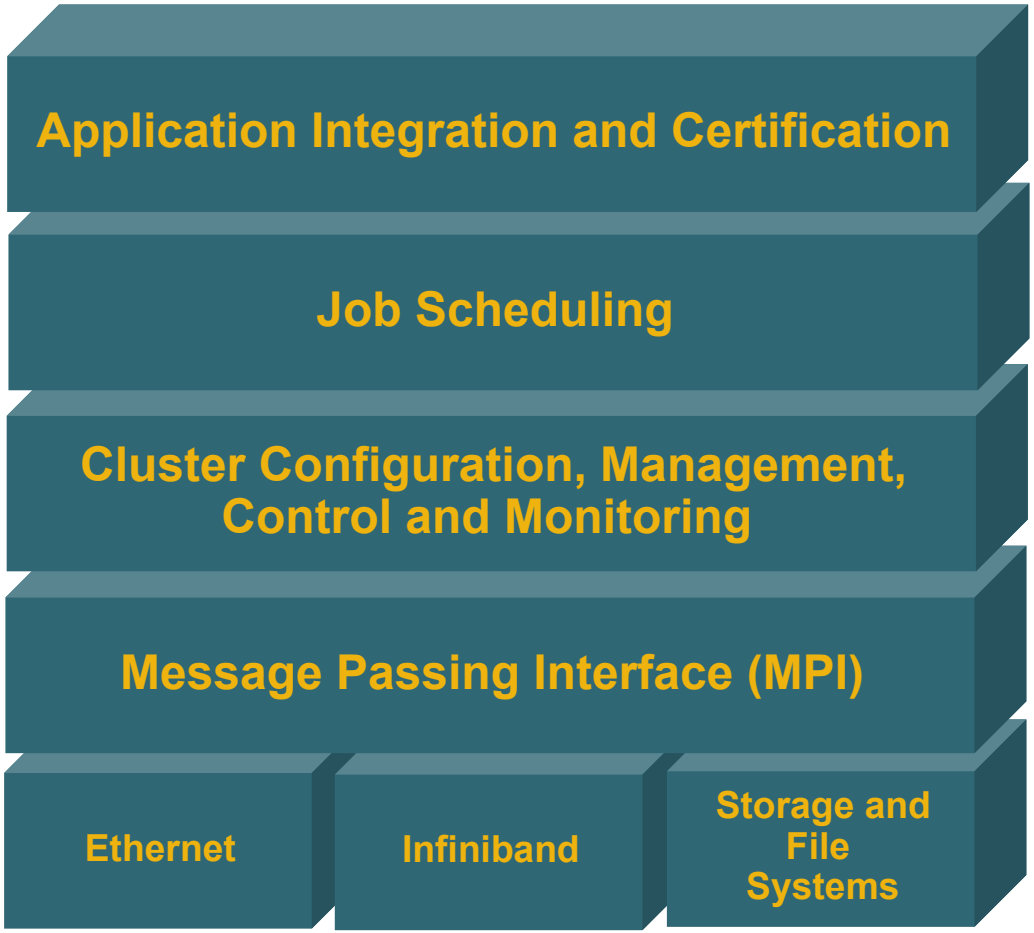
- Pass-through module for Dell Blade Chassis
- PCI-Express InfiniBand HCA daughter-card
- Only available on Dell GPL

Addressing HECS - InRow Cooling for High Density Racks

- 16.7 kW / module
- 50 kW N + 1 with 2 rack positions
- User replaceable fans
- Chilled Water

For medium to large data centers including high density environments.

Problem: Fragmented Solutions for High Performance Computing



Do-it-yourself testing, integration and certification with multiple fabrics, MPI and protocol stacks

No application awareness of network capabilities

Hand-scripted management per network (IPC vs. I/O vs. storage)

Multiple MPI stacks – no standards to build off of, more options to certify

Multiple Fabric Options – no uniform integration of capabilities

Cisco's Multifabric Strategy

Only vendor to offer complete multifabric solution.

Multi-fabric Management

- Bring Ethernet Ease-of-Use to InfiniBand
- Single management tool and APIs for large diverse fabrics

Multi-fabric Protocols

- Unify MPI across fabrics with Open MPI.
- Unify RDMA across fabrics with Open Fabrics.

Multi-Fabric Cisco Certification

- GigE, InfiniBand, 10GigE
- HPC Storage Certification

Multi-Fabric I/O

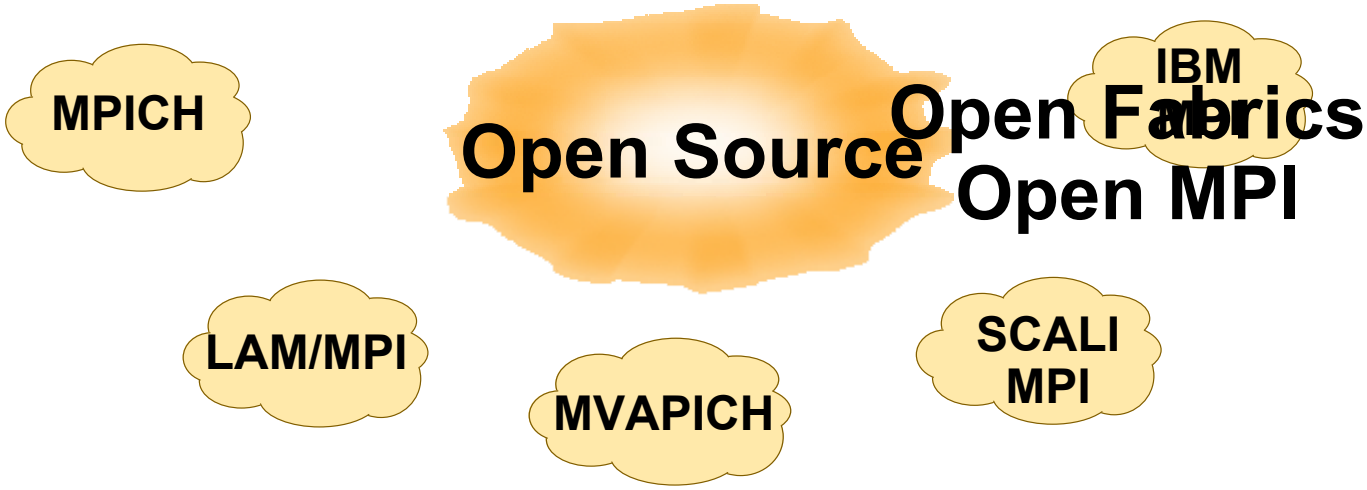
- Unified Fabric over IB with I/O gateways.
- IB-attached storage for HPC

Cisco Solution – Driving the Evolution o High Performance Computing

From	To
No application validation	Standards enable rapid testing and certification
No link between network and schedulers	Open APIs for bandwidth monitoring and feedback
Too many MPI Choices	Consolidated MPI
No management – hand-scripted and managed	Uniform management over IPC, management, storage network
Fabric-specific implementations	Fabric Agnostic

+ World-class Service and Support for High Performance Computing

Driving Open Standards
Open Fabrics, Open RDMA and Open MPI



Problem: **Proprietary protocol stacks** multiply options for application development and complicate certifications

Solution: Cisco moves strategic focus to **Open Fabrics and Open MPI**

Benefits: Delivers **one stack, fully interoperable** between vendors, **accelerating** application development for the ISV and deployment and certification for the end-user

Verticals in HPC

High Performance Computing

Universities and Labs

Education – universities, labs, foundations

Government – research, military, science

Enterprise

Petroleum - Oil and Gas Exploration

Manufacturing – Automotive, Aerospace

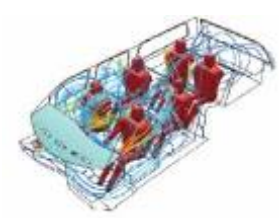
Bio-sciences

Financial – Data mining and market modeling

Enterprise HPC Application Areas and Verticals

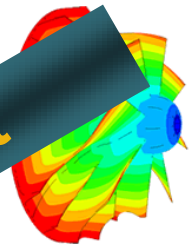
Computational Fluid Dynamics

- Fluent
- Star-CD
- Exa PowerFlow
- Vectis



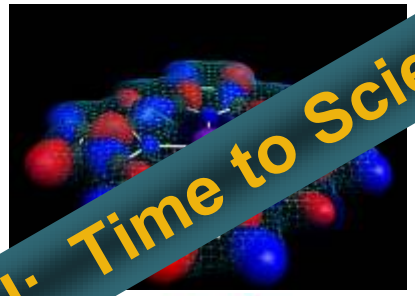
Finite Element Analysis

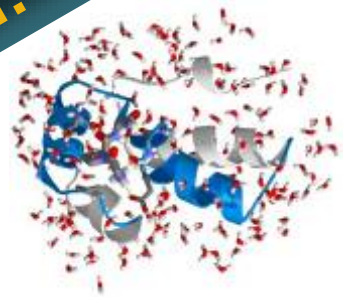
- Abaqus
- Ansys
- LS-Dyna (3-D FEA)
- RADIOSS
- NASTRAN
- PAM-Cra



Computational Chemistry Material Sciences:

- Accelrys
- ADF
- GAMESS
- CHARMm
- Namd
- NW
- ACS



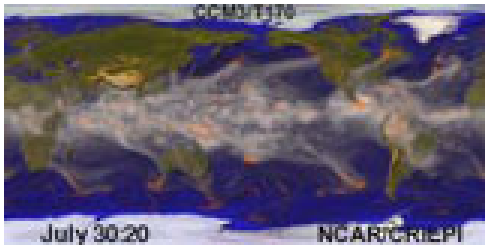


Bi… matics

- BLAST
- FASTA
- ClustaW
- EMBOSS

Climate & Weather Simulation

- MM5
- WRF
- CCM3



Goal: Time to Science; Time to Market

Performances Results

	procs	nodes	time (hours)	speedup
	1 proc (estimated)		*close to 2 days*	
	8	8	6:45:00	1
	16	8	3:54:10	1.7 / 2
	16	16	3:33:05	1.9 / 2
Ethernet	32	16	2:12:06	3.1 / 4
	32	32	2:06:50	3.2 / 4
	64	32	1:30:07	4.5 / 8
	64	64	1:25:22	4.7 / 8
	96	96	1:45:20	3.8 / 12
	128	128	2:03:41	3.2 / 16

	procs	nodes	time (hours)	speedup	IB/Eth. Ratio
	8	8	6:32:34	1	1,03
	16	8	3:41:16	1.8 / 2	1,06
	16	16	3:20:40	2.0 / 2	1,06
	32	16	1:52:41	3.5 / 4	1,17
Infiniband	32	32	1:40:47	3.9 / 4	1,26
	64	32	1:01:39	6.4 / 8	1,46
	64	64	0:55:50	7.3 / 8	1,53
	96	48	0:47:31	8.3 / 12	
	96	96	0:42:13	9.3 / 12	2,50
	128	128	0:33:02	11.9 / 16	3,74

From days to minutes .

IB outperforms Ethernet from 3 to 374%

18

Case Study: Leading Research Facility
High Performance Computing Cluster - NCSA

Application:

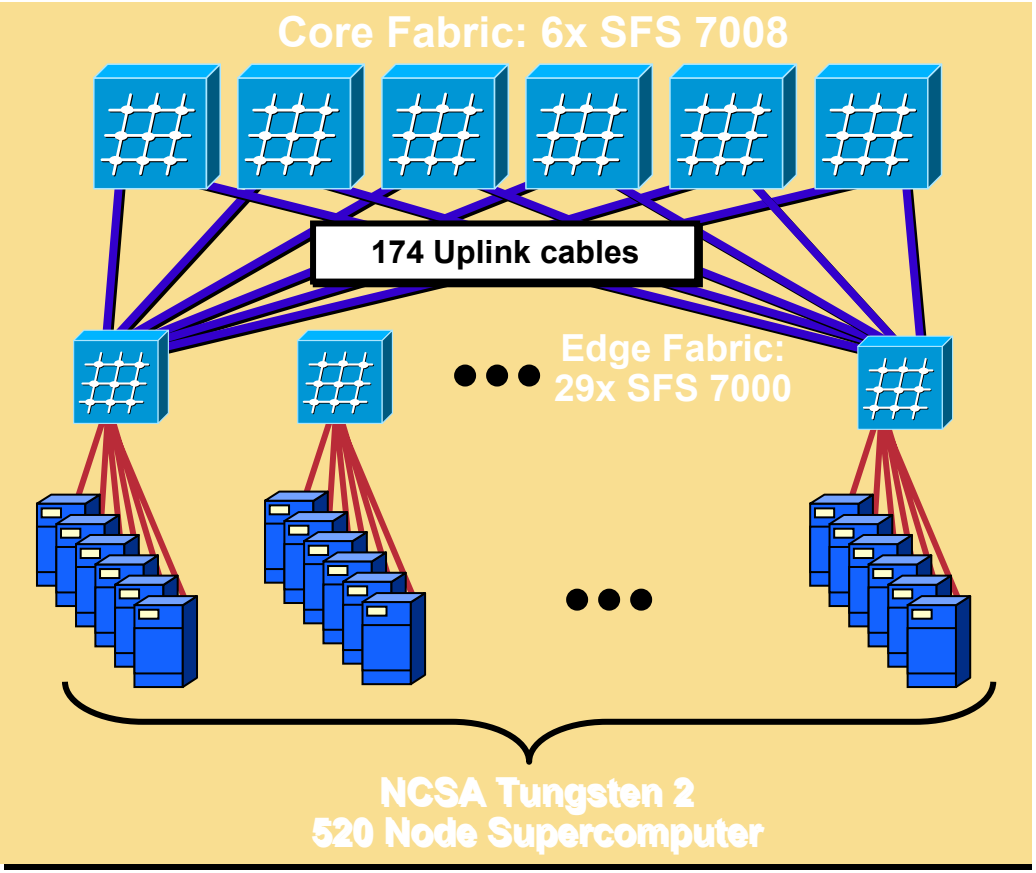
- High Performance Computing Cluster
- Compute time outsourced to Commercial Enterprises (major oil & gas)

Environment:

- 520 Dell Servers
- 3:1 Blocking ratio
- 6x SFS 7008
- 29x SFS 7000

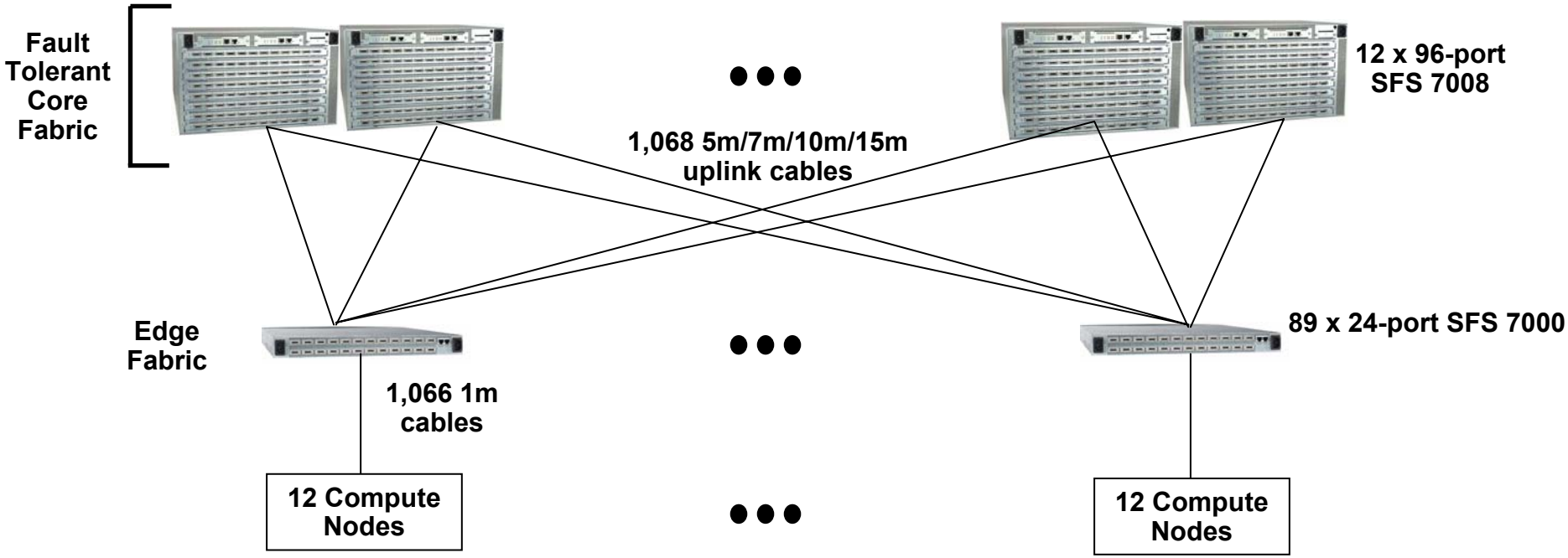
Benefits:

- Compelling Price/ Performance
- Measured MPI latency 5.2µs



Case Study : Bio-Informatics Cluster - 1,068 Node Supercomputer

1,068-node fully Non-Blocking Fault Tolerant IB Cluster



Key decision factors:

Cisco benchmarked and tuned customer MPI application

Best operational experience with large clusters – best references

“Rapid Service” architecture proved 2-min vs. 2-day MTTR.

Case Study : Sandia National Labs – 4096 Nodes Cluster

Application:

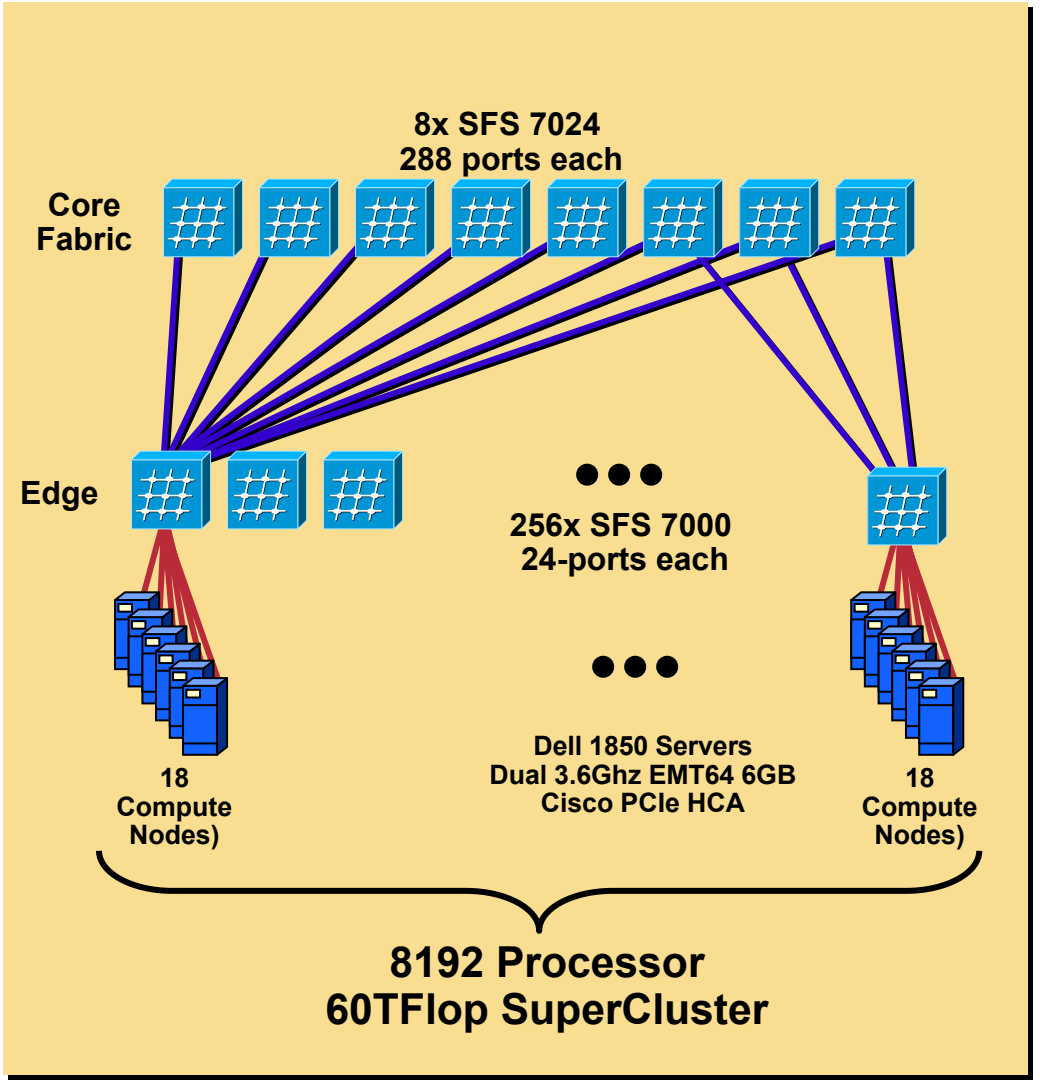
- High Performance SuperComputing Cluster

Environment:

- 4096 Dell Servers
- 50% Blocking Ratio
- 8 x SFS 7048
- 256 x SFS 7000

Benefits:

- Compelling Price/Performance
- Largest IB Cluster ever built - 8,192 Processor, 60TFlop Cluster
- Expected to be 3rd Largest Supercomputer in the world



Additional Information

www.cisco.com/go/hpc

www.cisco.com/go/serverswitching

www.cisco.com/go/datacenter