Checklist for Selecting and Deploying Scalable Clusters with InfiniBand Fabrics

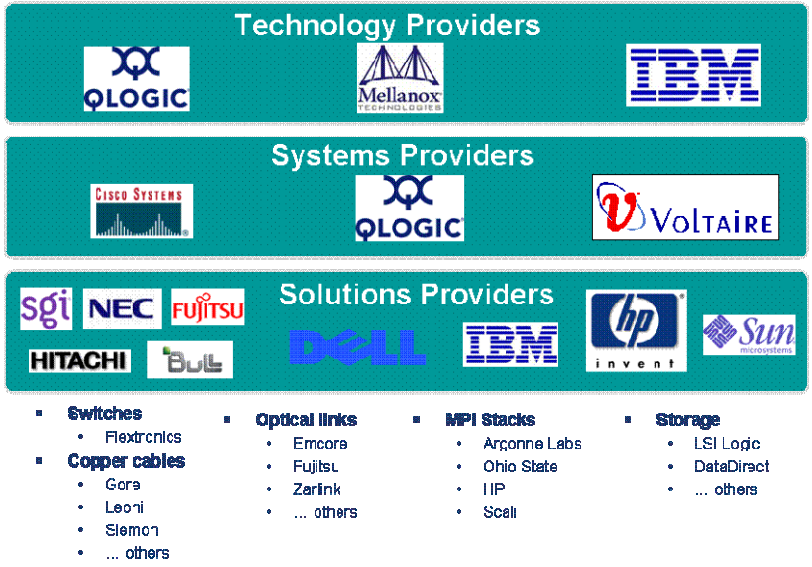
Lloyd Dickman, CTO

InfiniBand Products
Host Solutions Group
QLogic Corporation

November 13, 2007 @ SC07, Exhibitor Forum

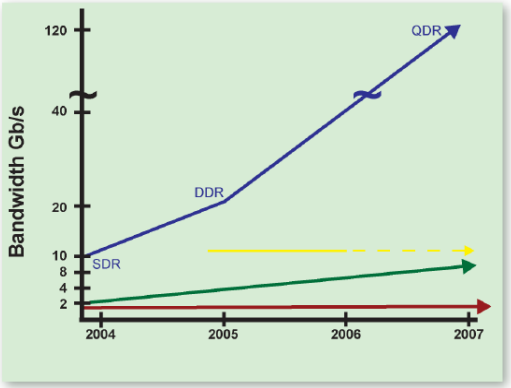
InfiniBand is the Fabric of Choice for HPC

- **Large and growing standards-based eco-system**
- **Aggressive technology roadmap**
- **Low-latency, high messaging rate implementations in market**
- **Evidenced by increasing traction in Top500**



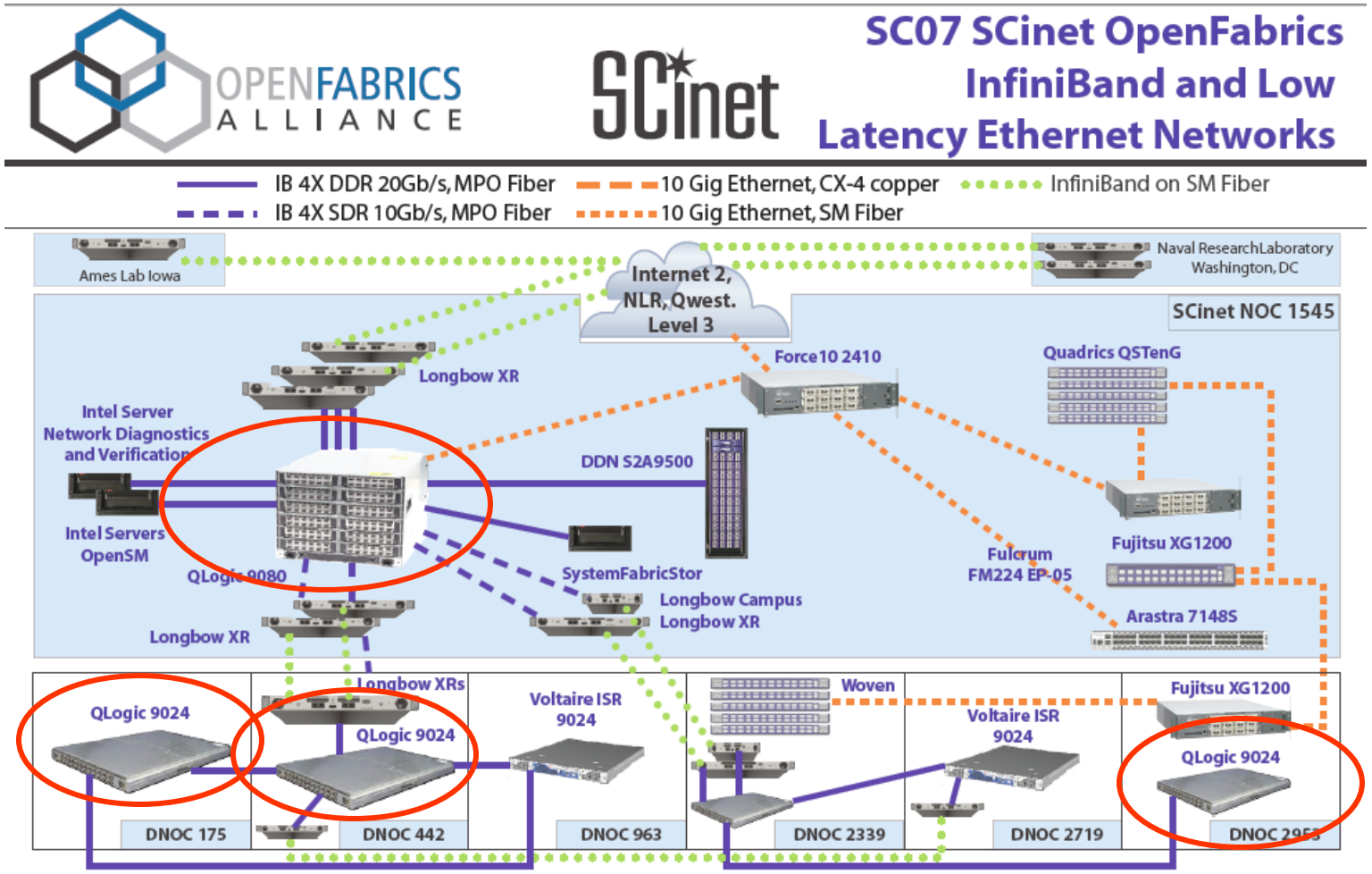
InfiniBand Roadmap

InfiniBand's roadmap outpaces all proprietary and standard based I/O technologies in both pure performance and price/performance



Number of IB Lanes	Per Lane Bandwidth		
	SDR 2.5Gb/s	DDR 5Gb/s	QDR 10Gb/s
4X	10Gb/s	20Gb/s	40Gb/s
8X	20Gb/s	40Gb/s	80Gb/s
12X	30Gb/s	60Gb/s	120Gb/s

InfiniBand Interoperability



InfiniBand Exhibitors

AMD – 757
Ames Lab – 181
HP – 1417
IBM – 1215
Intel – 820
Mellanox – 127
Nat Cntr Data Mining - 2623
OpenFabrics – 2660
QLogic – 261
Sun Microsystems – 514
Violin/SFW – 158
Voltaire – 1137
Yotta Yotta – 703
SC Education Prog – L1

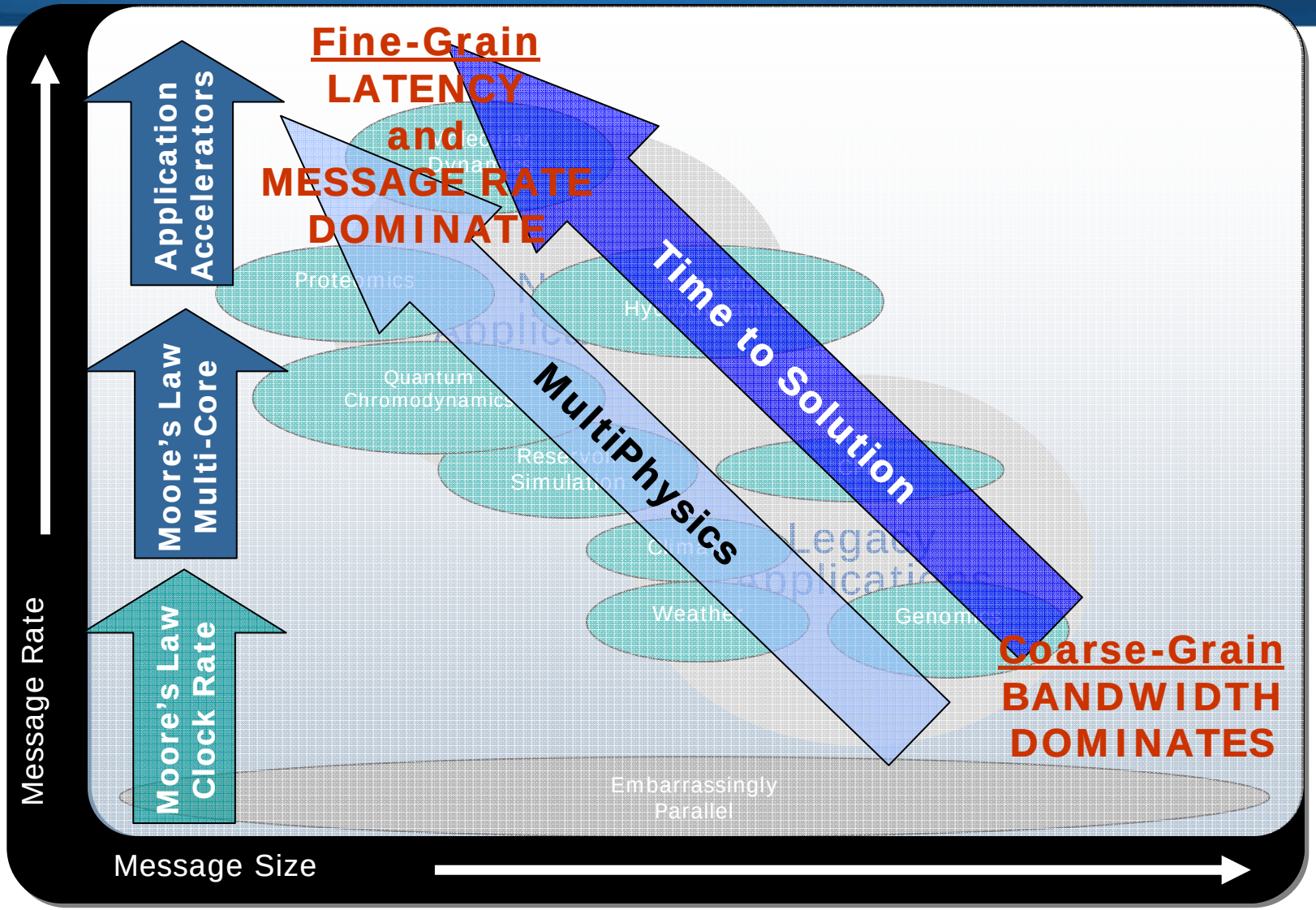
iWARP Exhibitors

Arastra – 936
Chelsio – 2337
Fujitsu – 1429
Fulcrum – 2161
NetEffect – 2753
OpenFabrics – 2660
Woven – 2733

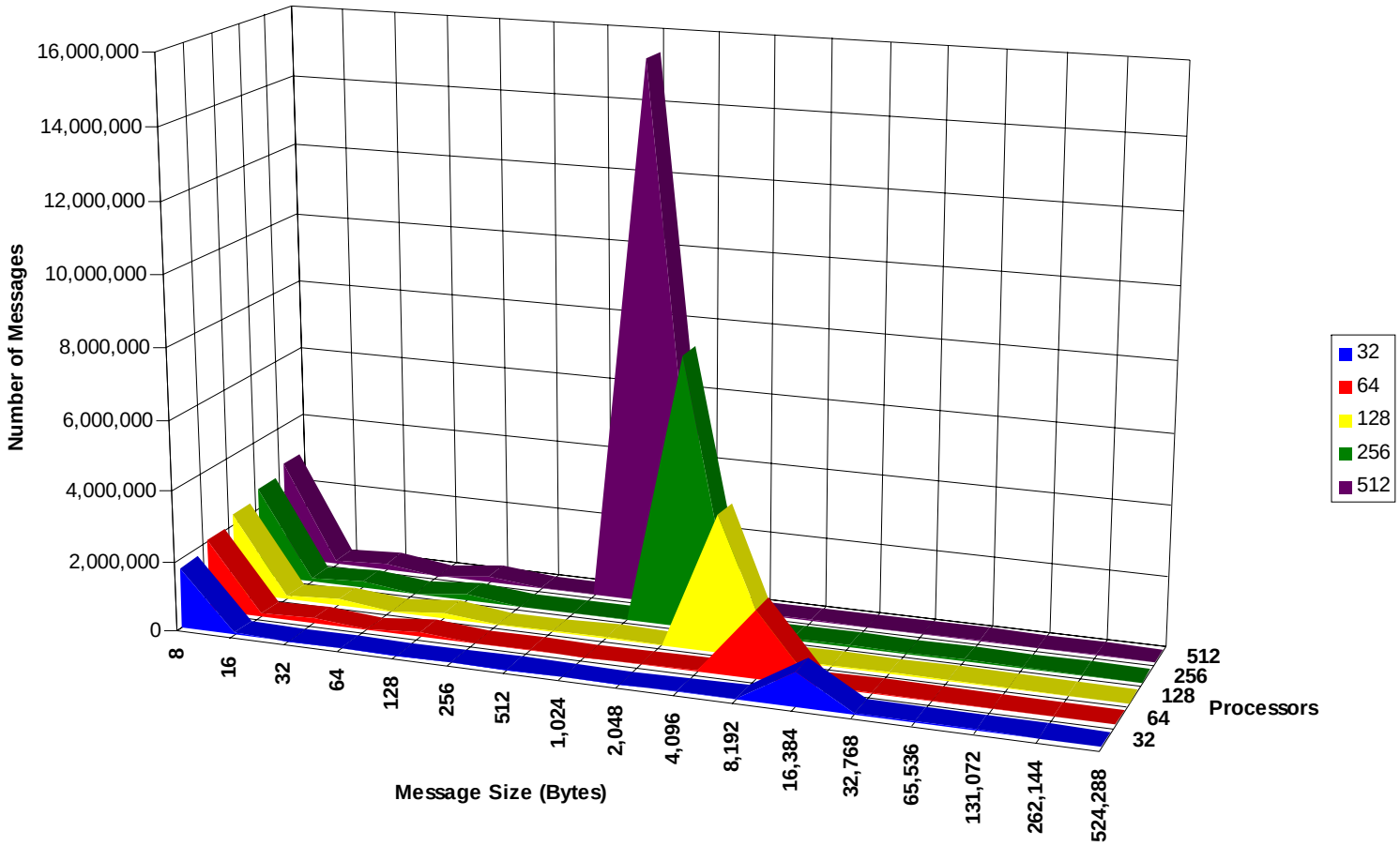
Interdependent Tradeoffs

- **Decisions**
  - ✓ Host Architecture
  - ✓ Fabric Topology
  - ✓ Fabric Convergence / Unification
  - ✓ RAS, Power / Cooling
  - ✓ Software – host applications, ULPs and fabric

For HPC, Application Messaging Profiles Matter!



Application Messaging Patterns Change with Scale



Source: PathScale measurements, averages per node, DL_POLY_2 AWE Medium case, 19dec2005

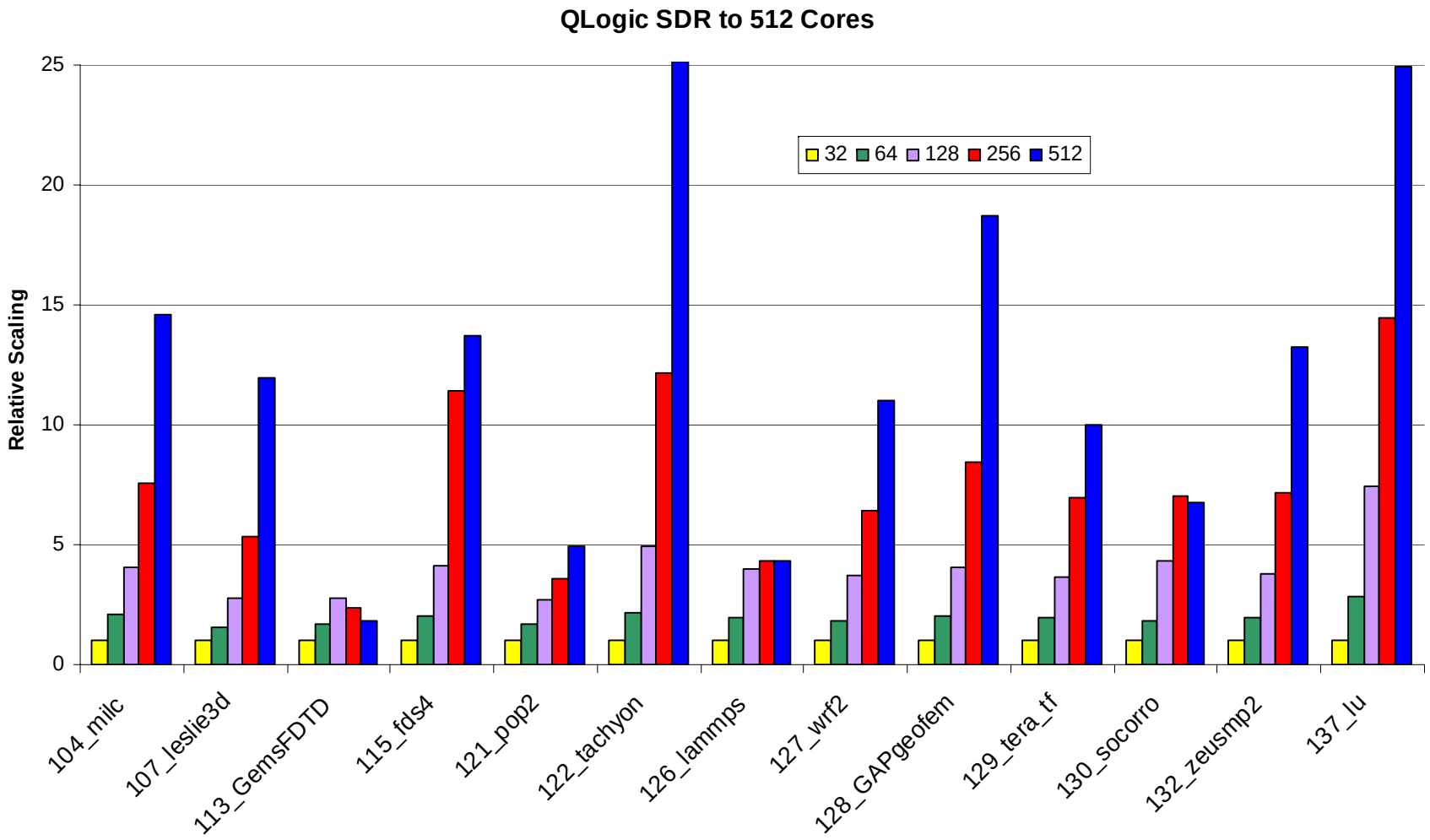
QLogic InfiniBand Provides Superior Application Scaling

	Comparison	Benchmark	QLE7140 SDR	QLE7240 DDR	QLE7280 DDR	vs Competitive
Point-to-Point	**Point-to-Point Bandwidth**	**OSU Bandwidth** (peak)	**0.9 GB/s**	**1.5 GB/s**	**1.9 GB/s**	**+24%**
	Point-to-Point Latency	**OSU Latency** (0 Byte)	**1.6 µs**	**1.5 µs**	**1.4 µs**	**Tie**
	Point-to-Point Message Rate (non-coalesced)	**OSU Message Rate** @ 4 processes/node	**9 M/s**	**12 M/s**	**13 M/s**	**2.6X**
Scaling	**Scalable Latency**	**HPCC Random Ring Latency** @ 32 cores	**1.5 µs**	**1.4 µs**	**1.3 µs**	**1.8X**
	Scalable Message Rate (non-coalesced)	**HPCC MPI Random Access** @ 32 cores	**.12 GUPs**	**.12 GUPs**	**.13 GUPs**	**7.6X**

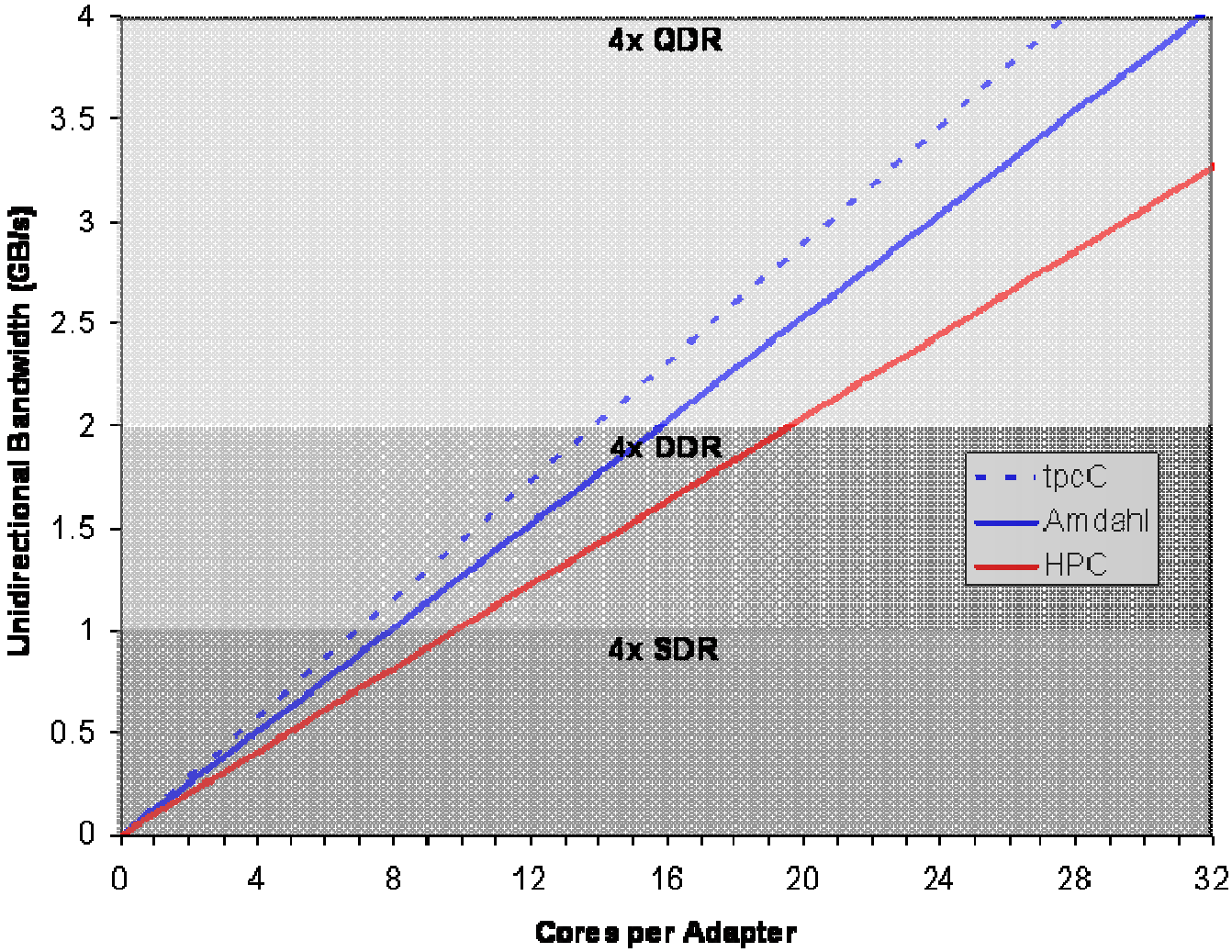
Point-to-point microbenchmarks are no longer sufficient predictors of performance at scale.

Scaling results are for 8 nodes / 32 cores.
Latency results include a single switch crossing.
Results use native MPIs: InfiniPath MPI for QLogic

SPEC MPI2007 Results Confirm Scalability

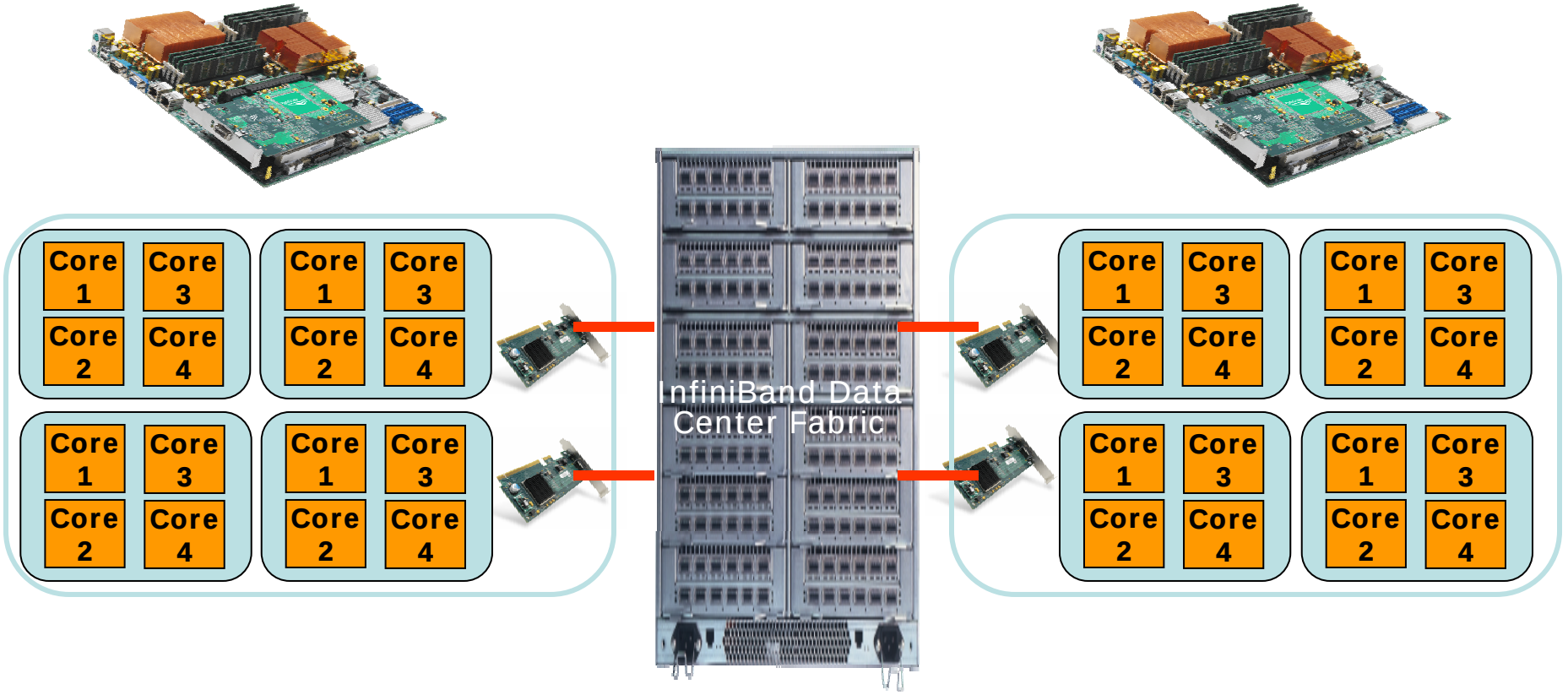


Bandwidth Must Account for ALL Traffic

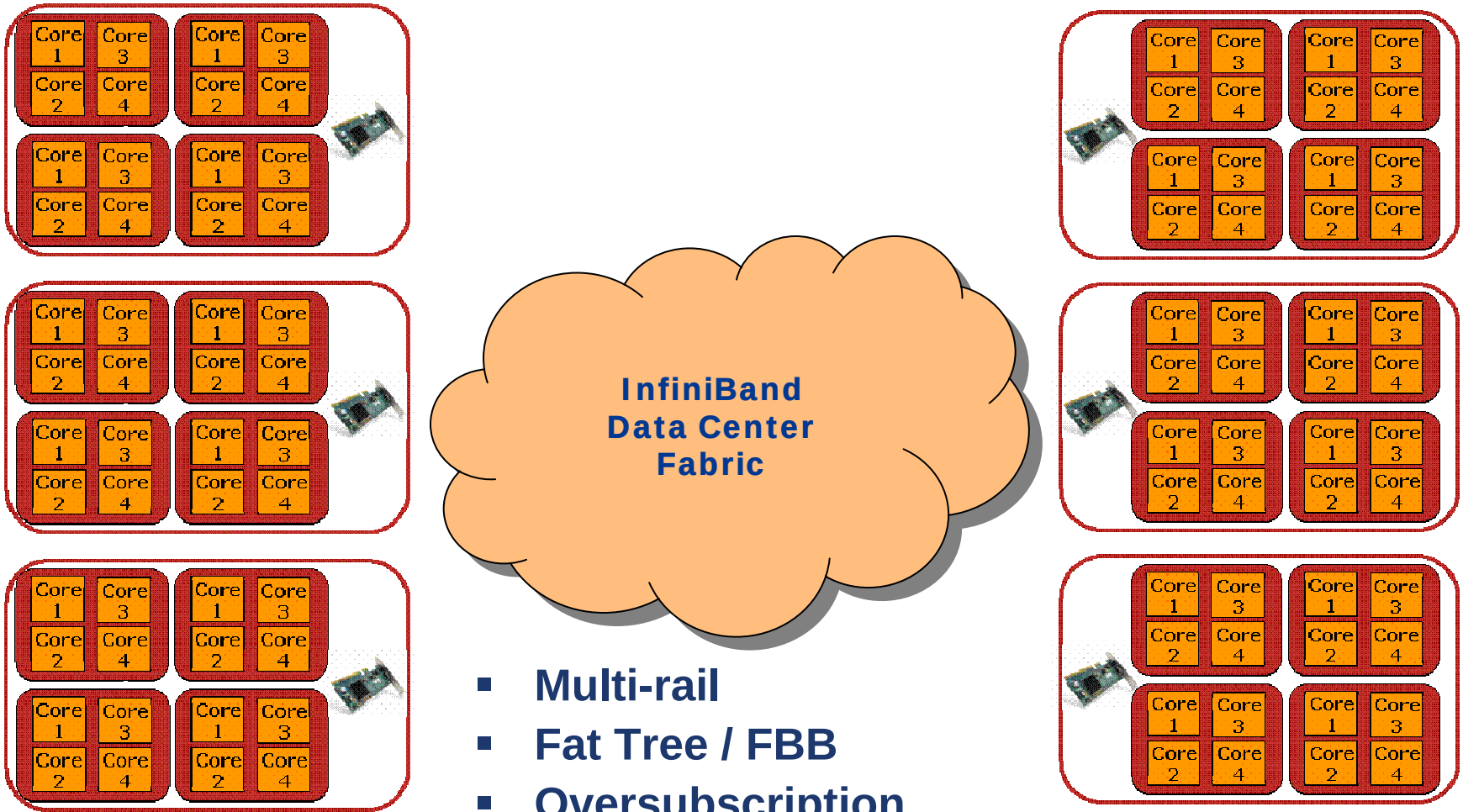


50% Link Utilization

Host-to-Fabric Rails



Fabric Topology Considerations



- Multi-rail
- Fat Tree / FBB
- Oversubscription
- Asymmetric Topology

QLogic InfiniBand Fabric Products

Core Fabric Switches

Multi-Protocol Fabric Director Modules

INFINIBAND™ Trade Association — 4X SDR, 4X DDR

IEEE 802

FCIA

14U — 288 Ports

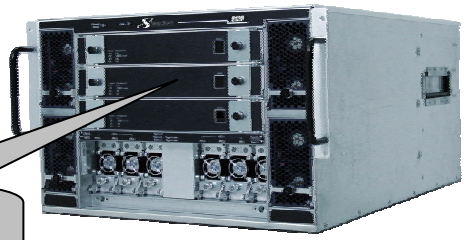
7U — 144 Ports

7U — 96 Ports

4U — 48 Ports

24 Ports

19" rack



SDR and DDR Spines

Edge Fabric Switch

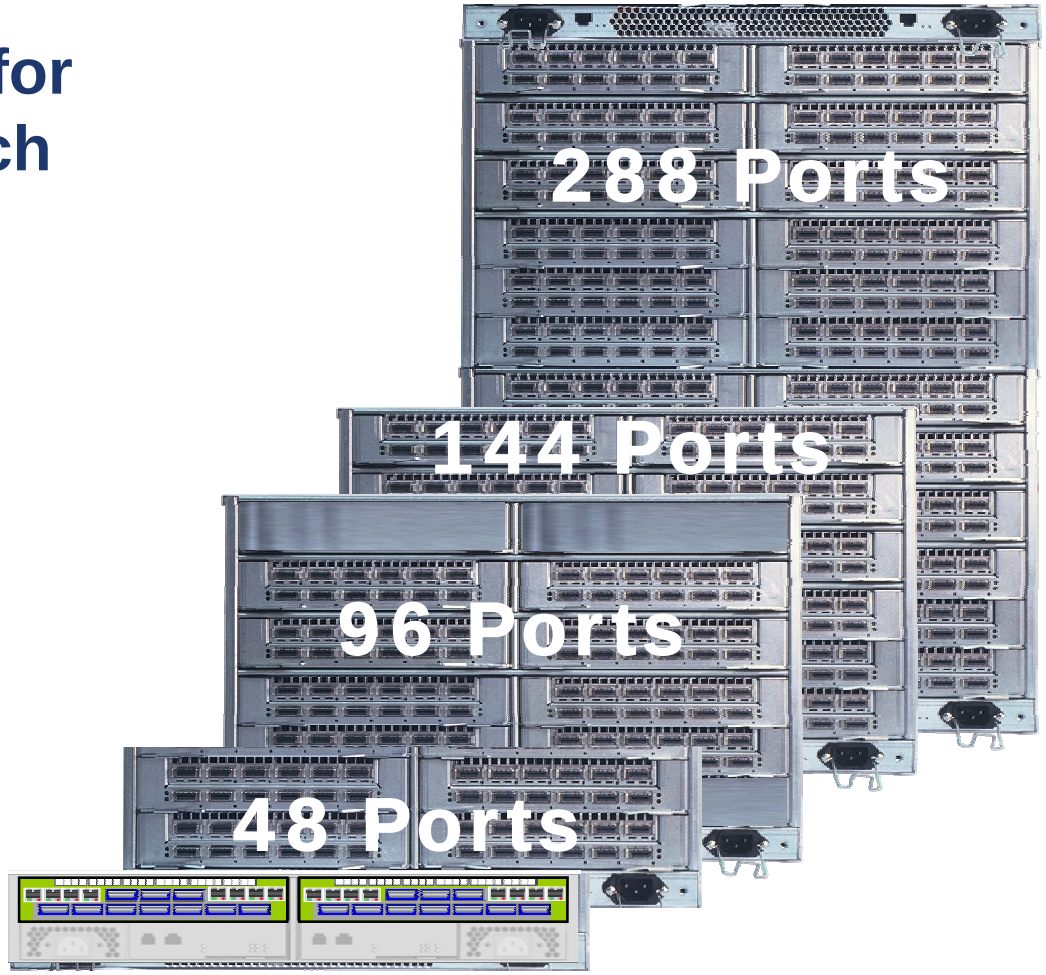
1U — 2U

Production quality DDR since April 2006,

Over 50,000 ports to date with extensive OEM qualifications

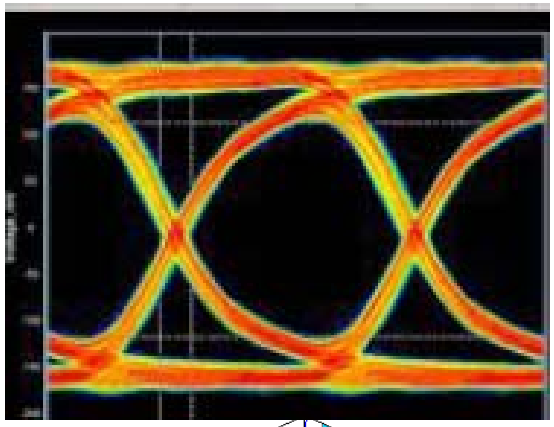
Matching Fabric to Node Count

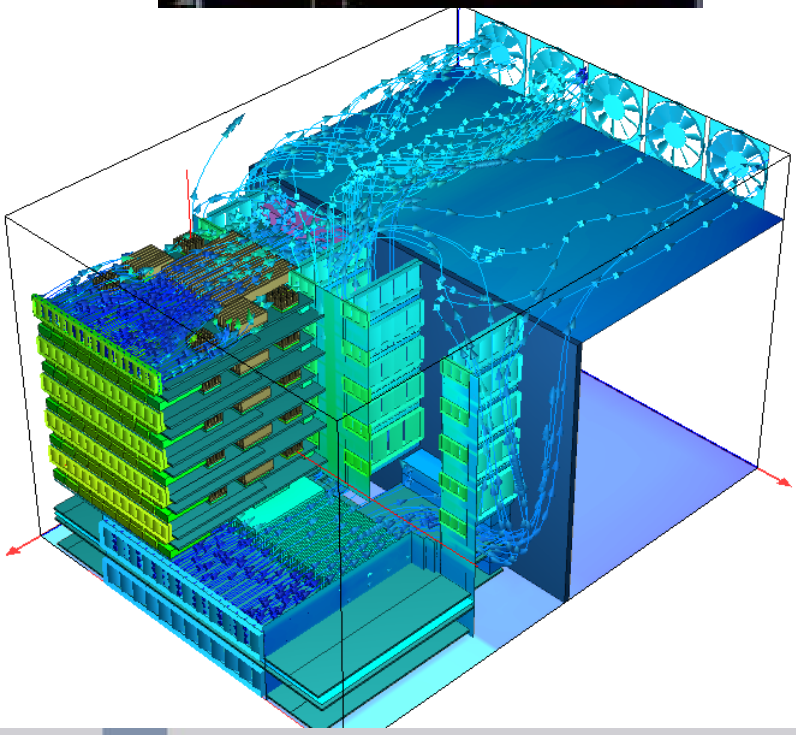
- **Optimal Fabric Size for today's 24-port switch chips**
  - 1 tier – 24 ports
  - 2 tier – 288 ports
  - 3 tier – 3,456 ports



9020 Switch / FC / 10 GbE Gateway

Advanced Design Capabilities

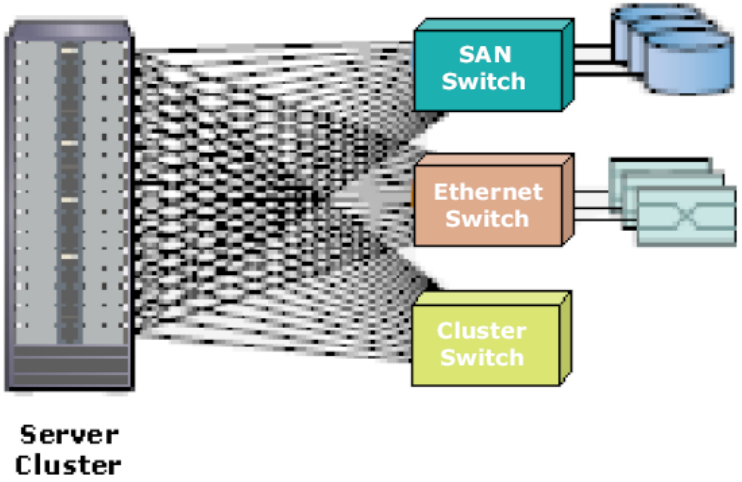


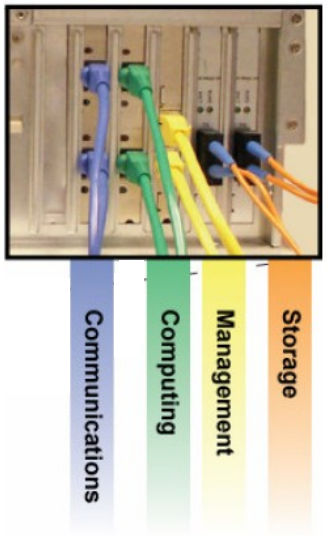


- **Signal Integrity**
  - Internal and external bit lanes capable of 10Gb/sec
  - Significant signal integrity challenge
  - Leveraging established QLogic center for signal integrity analysis at headquarters
- **Thermal Analysis**
  - Chip power nearly triple previous generation
  - Leveraging QLogic CFD capabilities

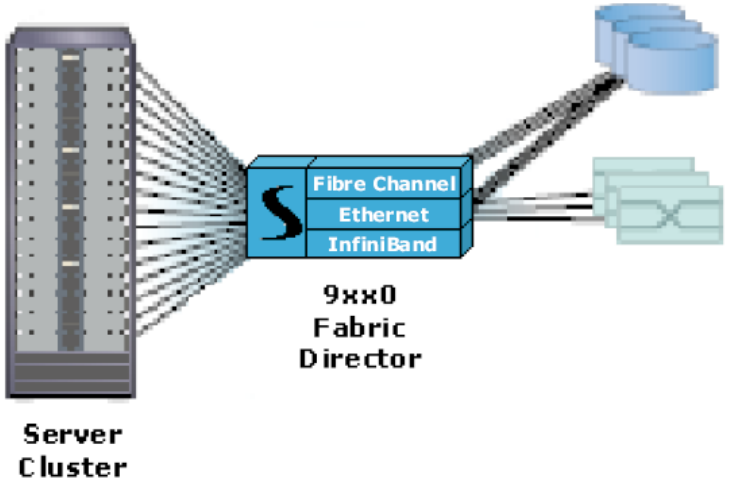
Unified InfiniBand Fabric Reduces Cost

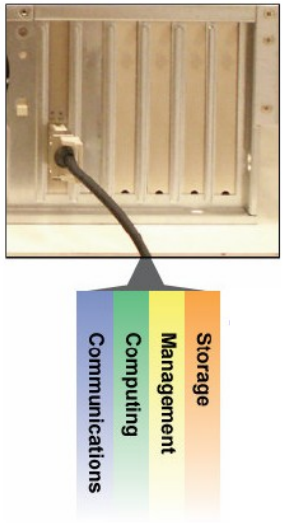
Typical Cluster Interconnect





Unified Fabric Interconnect



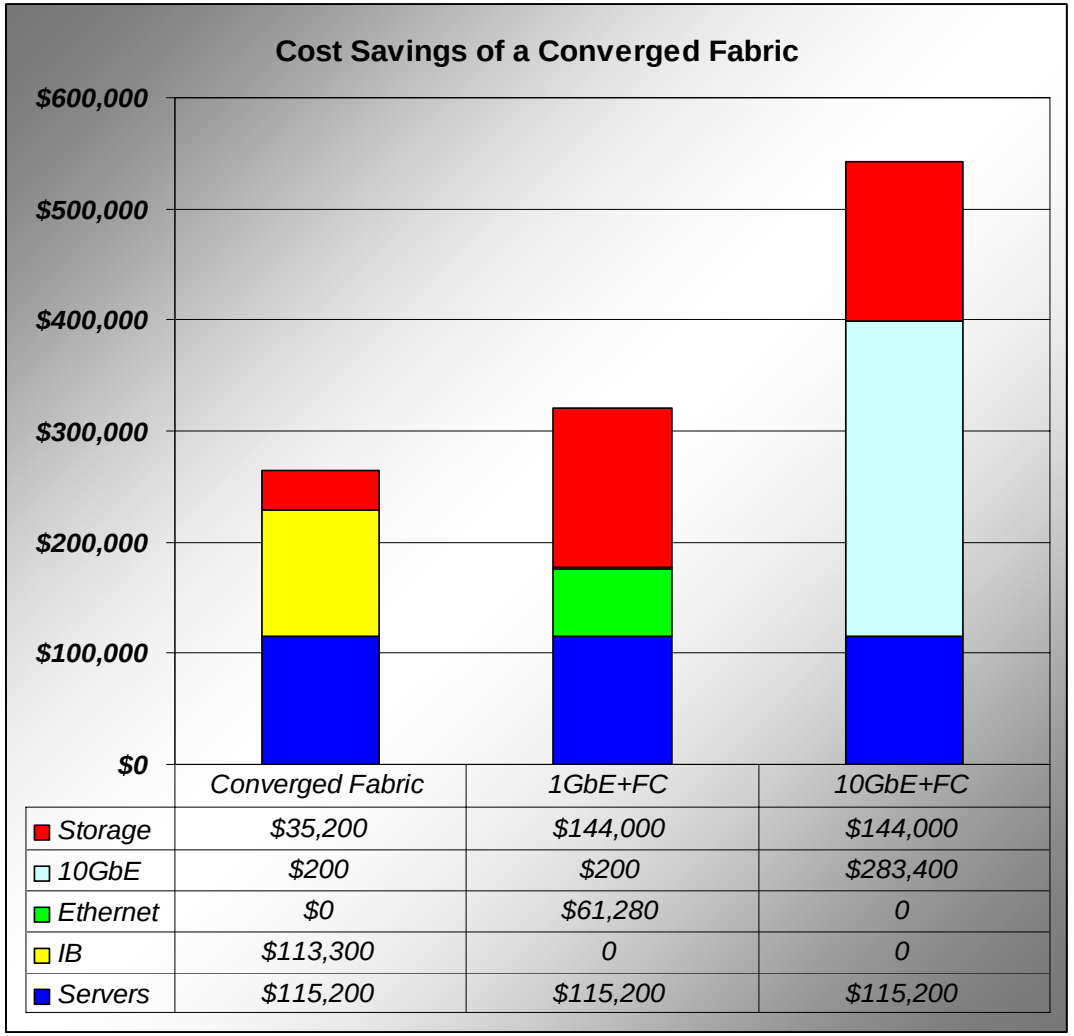


"One Wire"

Unified Fabric Deployments Increasing

Cost and Power Advantages of Converged Fabric

- **CF cost and power savings**
  - FC HBAs and switches
  - Ethernet NICs and switches
  - Cables
- **Additional cost savings**
  - Simpler management: single network
- **Comparisons for a 64-node cluster**
  - CF advantage increases for larger clusters
  - 30-38% cost savings on a 400/800-node FSC cluster
- **Power savings from CF**
  - ~30% vs. GigE



	Converged Fabric	1GbE+FC	10GbE+FC
Storage	$35,200	$144,000	$144,000
10GbE	$200	$200	$283,400
Ethernet	$0	$61,280	0
IB	$113,300	0	0
Servers	$115,200	$115,200	$115,200

Source: HP BoF, Converged Fabrics, SC07

Connectivity to ALL Storage

Direct Connect
InfiniBand
Storage

iSER

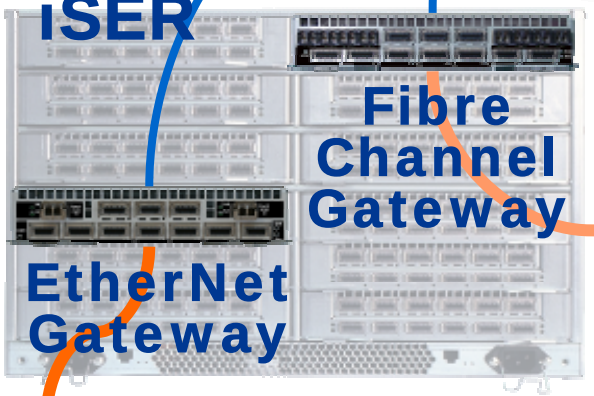
SRP

iSCSI

iSER

Fibre
Channel
Gateway

EtherNet
Gateway



iSCSI

IEEE 802

FCIA

Connectivity to ALL Sockets Applications and Ethernet Networks

Intra-Cluster Sockets Communications

SDP

TCP/IP

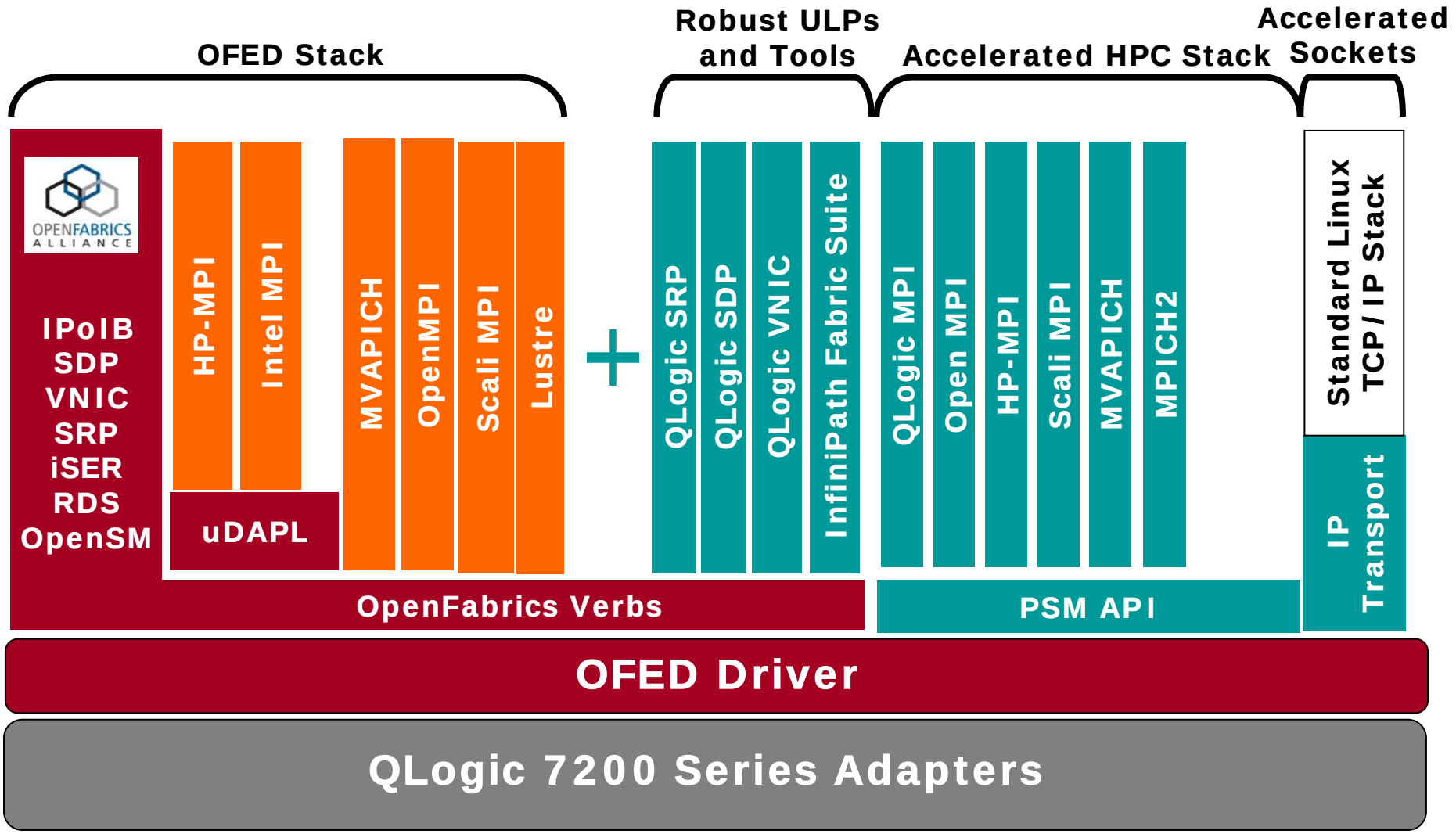
IPoIB-UD
IPoIB-CM

VNIC

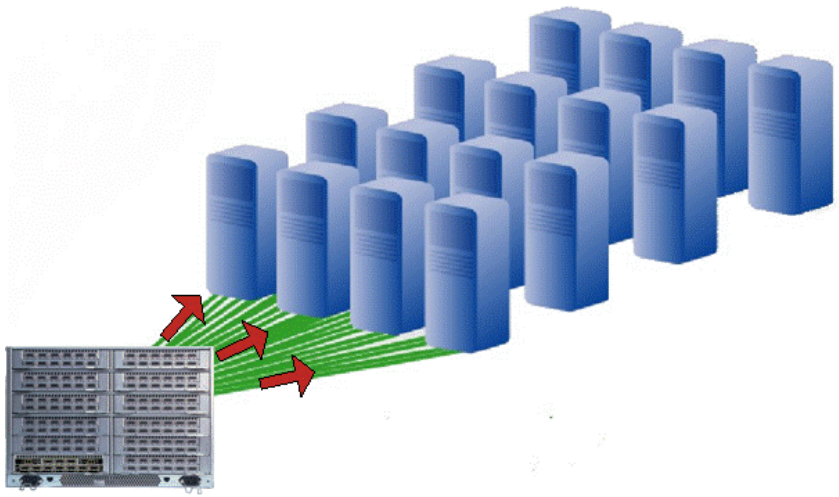
EtherNet Emulation

IEEE 802

OFED+ Provides OpenFabrics Benefits Plus Optional Value Added Capabilities



Large Fabrics Challenges



- **Scale of InfiniBand deployments growing rapidly**
  - 1000s of Nodes, soon 10,000s of Nodes
- **Fabric is the cluster**
  - 1 wire vision being deployed now
  - Requires high degree of fabric stability
- **Fabric Diagnosis and Resiliency**
  - Many moving parts at large scale
- **Demands of End Nodes on SA increasing**

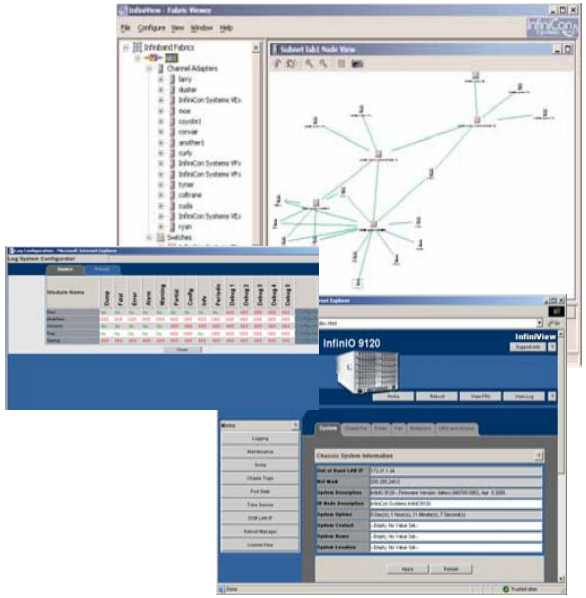
InfiniBand Fabric Suite

```
Fast Fabric IB Host Setup Menu
Host List: /etc/sysconfig/iba/hosts

0) Edit Config and Select/Edit Hosts Files    [ Skip  ]
1) Verify Hosts via Ethernet ping             [ Skip  ]
2) Verify rsh/rcp Configured                  [ Skip  ]
3) Setup Password-less ssh/scp                [Perform]
4) Copy /etc/hosts to all hosts               [ Skip  ]
5) Show uname -a for all hosts                [ Skip  ]
6) Configure IPoIB IP Address                 [ Skip  ]
7) Reboot Hosts                               [ Skip  ]
8) Refresh ssh Known Hosts                    [ Skip  ]
9) Run a command on all hosts                 [ Skip  ]
a) Copy a file to all hosts                   [ Skip  ]
b) View ibtest result files                   [ Skip  ]

P) Perform the selected actions
N) Select None

X) Return to Previous Menu (or ESC)
```



Lowering InfiniBand OpEx

- **Automates Cluster Install, Verification and Administration**
  - For both Hosts and Switches
- **Rich set of tools and capabilities for ongoing Fabric Administration**
  - Easy to use TUI and CLI
  - Easy integration into site specific tools
- **Rapidly Pinpoints fabric errors**
  - Extensive verification and diagnostic capabilities
- **Industry Leading Performance and Scalability**
  - 1024 node fabric initialization in <20 seconds
  - Fabric Change in < 12 seconds
  - Trap Based event notification & subscription supported
  - Load balances LID(s) across ISL(s)
- **Highly Resilient**
  - No single point of failure
  - Automatic Fabric Reconfiguration
- **Complete IBTA Support**
  - Optional SA queries (Trace Route, Link records, etc)
- **Flexible Deployment Options**
  - Embedded Fabric Management
  - Host Fabric Management

Please Visit Us to Learn More About Complete InfiniBand – Booth #261

QLOGIC

Cluster Administration using InfiniBand Fabric Suite

Advanced Technology Demonstrations

HPCtrack Partnership Program

9020 InfiniBand Switch and Multi-Protocol Gateway

Fluent Application running multiple data sets on a converged fabric with Fibre Channel storage via a gateway

InfiniBand DDR Adapter for highly scaling applications

QLOGIC

Thank You