An Introduction to Infiniband*

/sharan kalwani/

sharan.kalwani@ieee.org

sharan.kalwani@computer.org

Chair, IEEE Southeastern Michigan Section

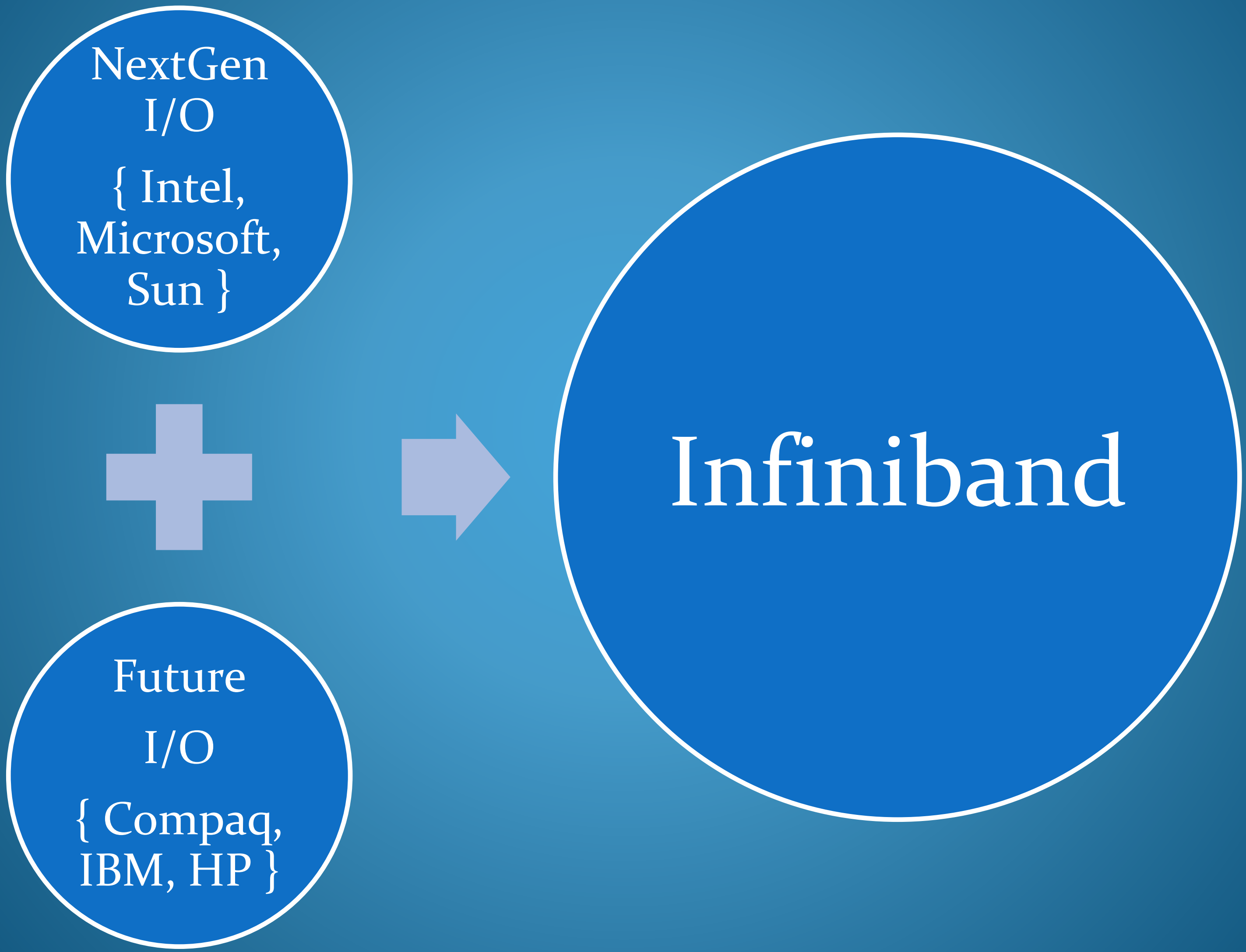
https://www.linkedin.com/in/sharankalwani

** Original Title: "Everything you wanted know to know about Infiniband: but did not know who to ask…."*

History

- During the late 90's, growth of CPU outstripped popular I/O
- Most choices were those of "shared bus" architectures
- Next Generation I/O (or NGIO) specifications was proposed by
  - Intel, Microsoft and SUN (Now called Oracle)
- Future I/O specifications was proposed by
  - Compaq (now HPE), IBM and HP (now HPE)
  - Compaq used a lot of technical detail from Tandem's ServerNet design
- A merger came about, so as to avoid confusion
- For a short time, it was known as *System I/O*
- The name **Infiniband (IB)** was chosen to represent:
  - **INFINI**TE
  - **BAND**WIDTH
- The major driver for Infiniband is now their Trade Association
- Web site – http://www.infinibandta.org

History



NextGen I/O

{ Intel, Microsoft, Sun }

Future

I/O

{ Compaq, IBM, HP }

Infiniband

Why was it needed?

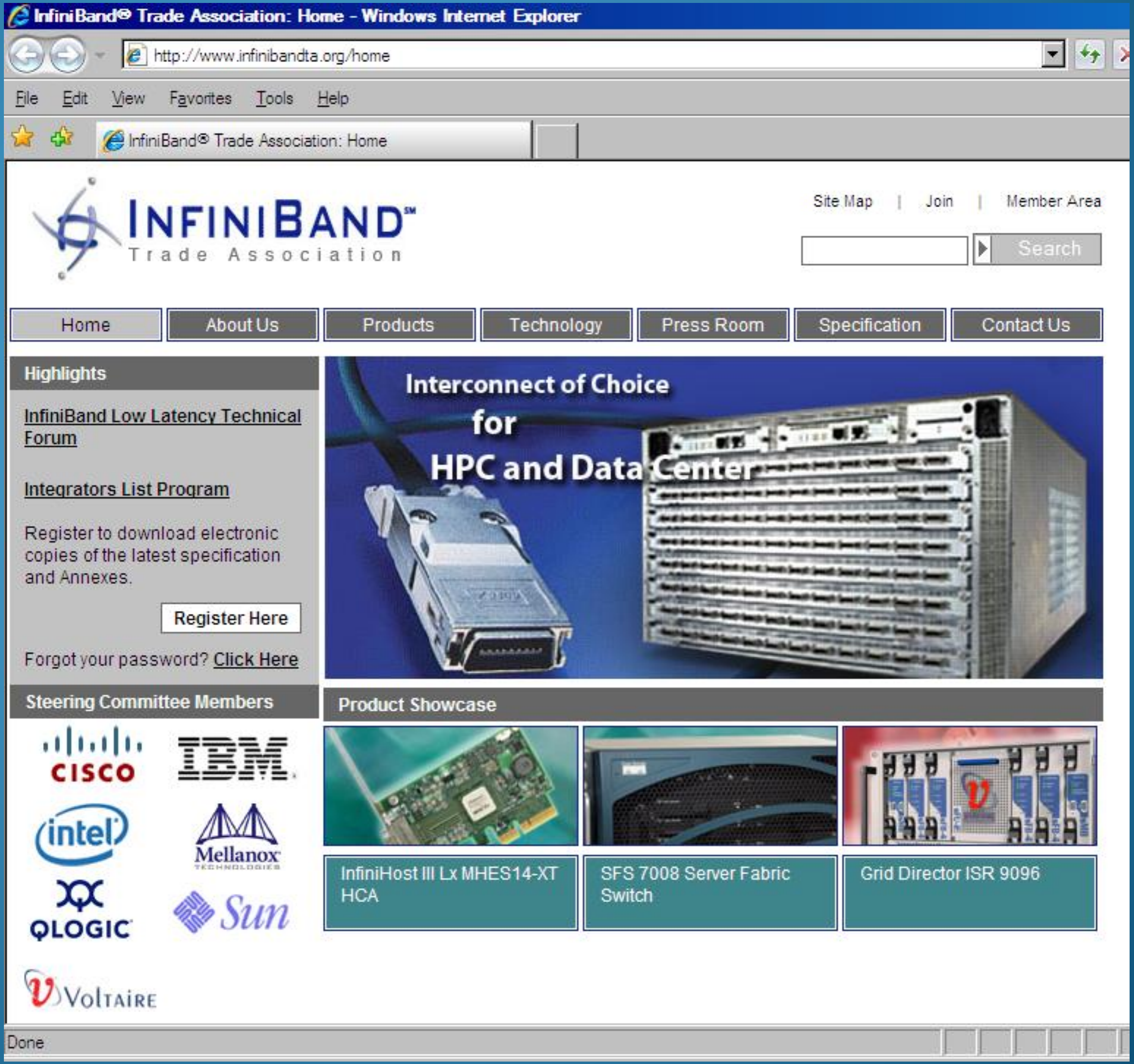
- CPU, memory, screen, hard disks, LAN and SAN interface
- All use a systems bus for communications
- As these elements became faster,
- The systems bus and overhead associated with data movement or I/O between devices became a **limiting** factor in performance.
- To address this problem (*I/O in particular*)
  - InfiniBand was developed as a standards-based protocol
  - It *offloads data movement* from CPU to dedicated hardware,
  - Allowing more CPU to be dedicated to application processing.
  - InfiniBand, *by leveraging* networking technologies & principles
  - Provided scalable, high-bandwidth transport

Infiniband

- IB Architecture Leverages two principles:
  - Switching and Routing –
  - Provides transport layer for upper-layer protocols
  - Supports flow control and quality of service (QoS)
  - Provides ordered, guaranteed packet delivery fabric.
- An IB fabric may comprise a number of IB subnets
- Subnets interconnected using IB routers, and
- Contain IB devices, switches, etc.
- Each point-to-point connection is a link, and may be
- Copper, optical, or even a printed circuit board!

Standards Body

- http://www.infinibandta.org (2009)



InfiniBand® Trade Association: Home - Windows Internet Explorer

http://www.infinibandta.org/home

File Edit View Favorites Tools Help
InfiniBand® Trade Association: Home
INFINIBAND℠ Trade Association
Site Map | Join | Member Area
Search
Home
About Us
Products
Technology
Press Room
Specification
Contact Us
Interconnect of Choice for HPC and Data Center
Done

Highlights

InfiniBand Low Latency Technical Forum

Integrators List Program

Register to download electronic copies of the latest specification and Annexes.

Register Here

Forgot your password? Click Here

Steering Committee Members

CISCO
IBM
intel
Mellanox TECHNOLOGIES
QLOGIC
Sun
Voltaire

Product Showcase

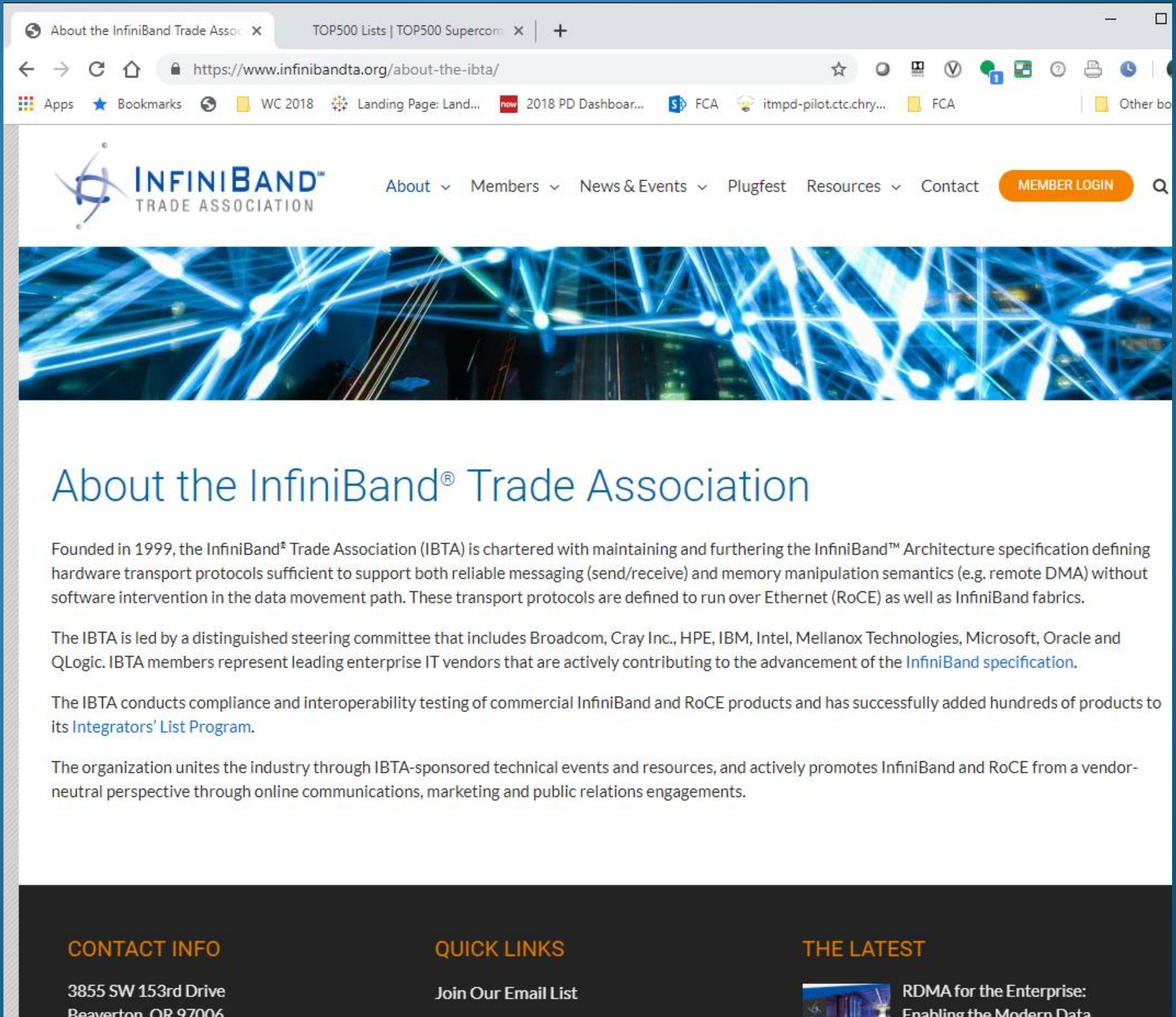
InfiniHost III Lx MHES14-XT HCA

SFS 7008 Server Fabric Switch

Grid Director ISR 9096

Standards Body

- ❖http://www.infinibandta.org (2019)



About the InfiniBand® Trade Association

Founded in 1999, the InfiniBand® Trade Association (IBTA) is chartered with maintaining and furthering the InfiniBand™ Architecture specification defining hardware transport protocols sufficient to support both reliable messaging (send/receive) and memory manipulation semantics (e.g. remote DMA) without software intervention in the data movement path. These transport protocols are defined to run over Ethernet (RoCE) as well as InfiniBand fabrics.

The IBTA is led by a distinguished steering committee that includes Broadcom, Cray Inc., HPE, IBM, Intel, Mellanox Technologies, Microsoft, Oracle and QLogic. IBTA members represent leading enterprise IT vendors that are actively contributing to the advancement of the InfiniBand specification.

The IBTA conducts compliance and interoperability testing of commercial InfiniBand and RoCE products and has successfully added hundreds of products to its Integrators' List Program.

The organization unites the industry through IBTA-sponsored technical events and resources, and actively promotes InfiniBand and RoCE from a vendor-neutral perspective through online communications, marketing and public relations engagements.

CONTACT INFO

3855 SW 153rd Drive
Beaverton, OR 97006

QUICK LINKS

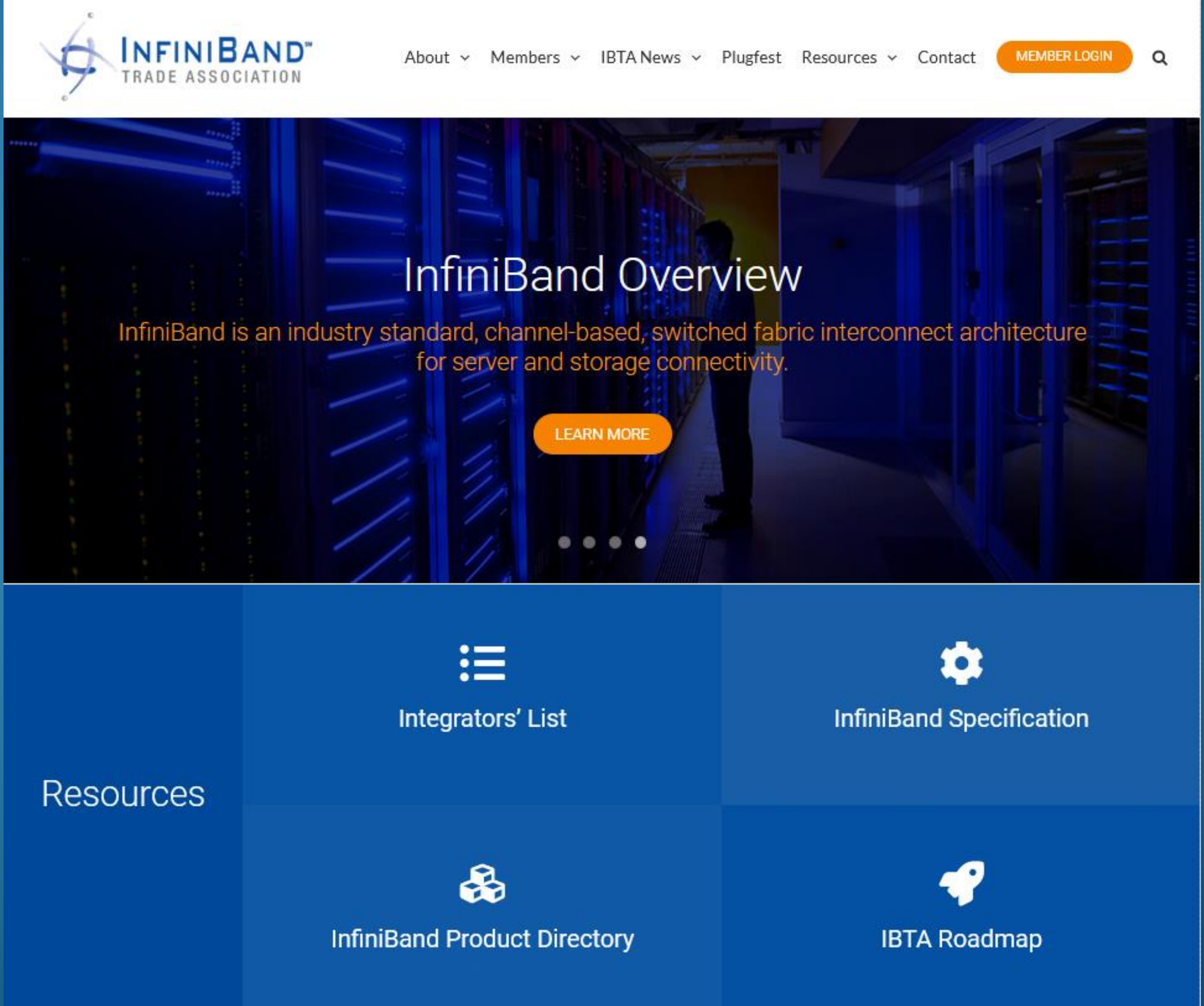
Join Our Email List

THE LATEST

RDMA for the Enterprise:
Enabling the Modern Data

Standards Body

- http://www.infinibandta.org (2023)



InfiniBand Overview

InfiniBand is an industry standard, channel-based, switched fabric interconnect architecture for server and storage connectivity.

IB Market Place (UPDATE!)

- Lots of Server OEMs
- Lots of Switch Suppliers
- Lots of Cable Suppliers
- Lots of Component Suppliers
- Major Silicon Chips made by:
  - Many sources – list available on the IBTA website or see the last slide

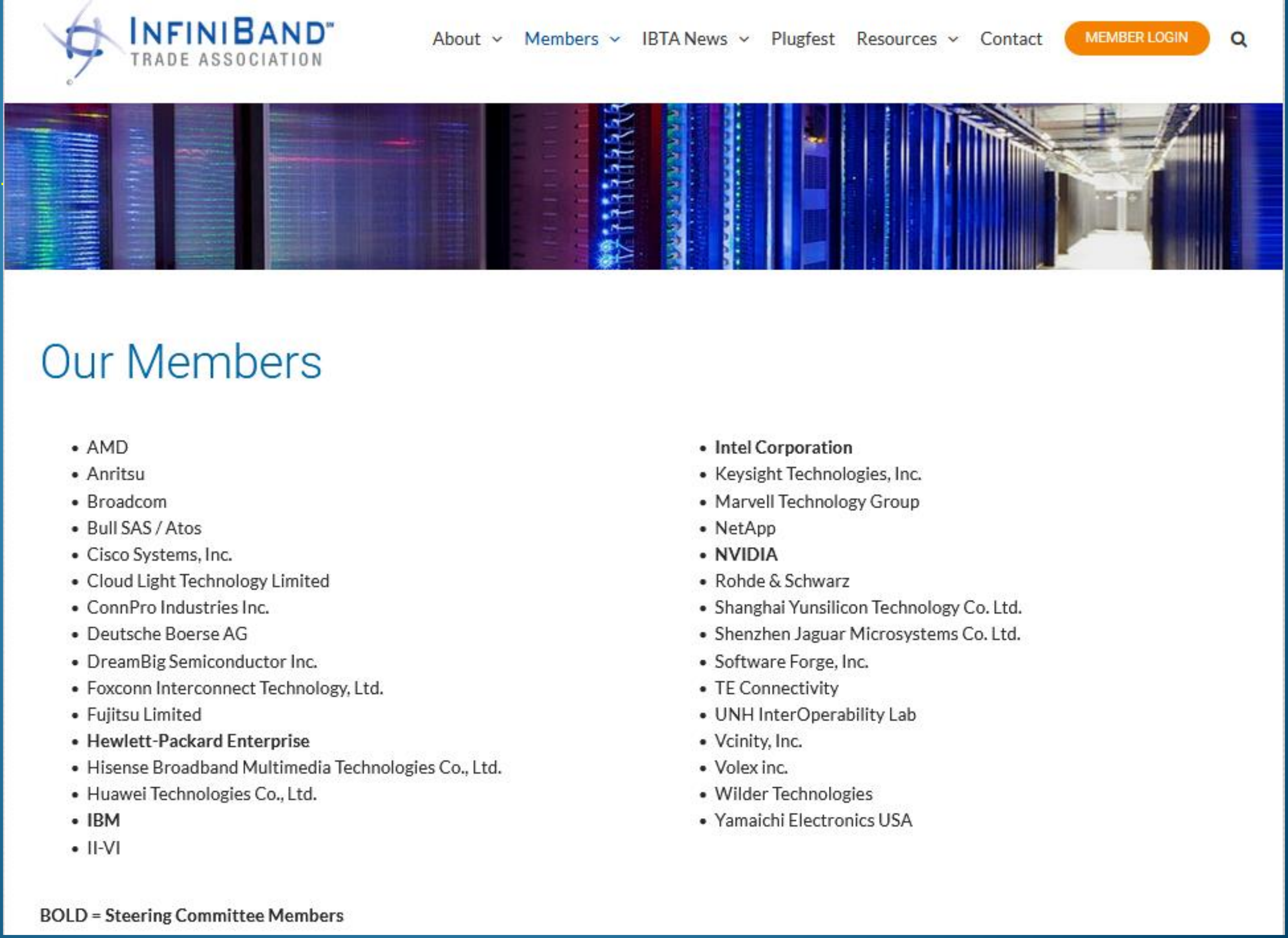
Infiniband Trade Association Members (2009)

- Steering Committee:
  - IBM
  - Intel
  - Cisco
  - Mellanox (now NVIDIA)
  - Qlogic (Intel)
  - SUN (Oracle)
  - Voltaire (Mellanox)
- Sponsor – Hitachi

- General members:
  - Amphenol
  - Brocade
  - Bay
  - Fujitsu
  - Lamprey
  - LSI Logic
  - Luxtera
  - NEC
  - Obsidian
  - Molex
  - WL Gore Associates
  - Xsigo, etc……

Infiniband Trade Association Members (2023)

- Steering Committee:
  - IBM
  - Intel
  - NVIDIA
  - HPE



Our Members

- AMD
- Anritsu
- Broadcom
- Bull SAS / Atos
- Cisco Systems, Inc.
- Cloud Light Technology Limited
- ConnPro Industries Inc.
- Deutsche Boerse AG
- DreamBig Semiconductor Inc.
- Foxconn Interconnect Technology, Ltd.
- Fujitsu Limited
- **Hewlett-Packard Enterprise**
- Hisense Broadband Multimedia Technologies Co., Ltd.
- Huawei Technologies Co., Ltd.
- **IBM**
- II-VI
- **Intel Corporation**
- Keysight Technologies, Inc.
- Marvell Technology Group
- NetApp
- **NVIDIA**
- Rohde & Schwarz
- Shanghai Yunsilicon Technology Co. Ltd.
- Shenzhen Jaguar Microsystems Co. Ltd.
- Software Forge, Inc.
- TE Connectivity
- UNH InterOperability Lab
- Vcinity, Inc.
- Volex inc.
- Wilder Technologies
- Yamaichi Electronics USA

BOLD = **Steering Committee Members**

Infiniband Trade Association Members (2023)

- AMD
- Anritsu
- Broadcom
- Bull SAS / Atos
- Cisco Systems, Inc.
- Cloud Light Technology Limited
- ConnPro Industries Inc.
- Deutsche Boerse AG
- DreamBig Semiconductor Inc.
- Foxconn Interconnect Technology, Ltd.
- Fujitsu Limited
- Hisense Broadband Multimedia Technologies Co., Ltd.
- Huawei Technologies Co., Ltd.
- II-VI

- Keysight Technologies, Inc.
- Marvell Technology Group
- NetApp
- Rohde & Schwarz
- Shanghai Yunsilicon Technology Co. Ltd.
- Shenzhen Jaguar Microsystems Co. Ltd.
- Software Forge, Inc.
- TE Connectivity
- UNH InterOperability Lab
- Vcinity, Inc.
- Volex inc.
- Wilder Technologies
- Yamaichi Electronics USA

- ***TOTAL: 31***
- ***(27+4) members***

What is Infiniband?

- Based on *Switched* Fabric as opposed to *Shared* Fabric (Ethernet)
- Hence - clear opposite ends of the spectrum (pun intended)
- Inspired by Fiber Channel, PCI Express and Serial link designs
- Point to Point link
- Bi-directional! (important)
- Links can be bonded

- Basic standard unit is called **1**X signaling rate or simply **1**x
- Equates to 2.5 Gigabits per second (or Gbps) in each direction

IB Signaling Speed (upto QDR)

- Basic standard unit is set around Single Data Rate (SDR)
- Starts w/ 2.5 Gbps in each direction
- (2009) Supported:
  - Double Data Rate (DDR), later Quad Data Rate (QDR)
- Signaling *achieves 80% efficiency…..*
  - since it uses 8B w/ 10B encoding
- In other words: 10 bits carries 8 bits of data
- Thus net **actual** data transmitted is 2.0 Gigabits/sec
- *Reminder:* Signaling speed is 2.5 Gigabits/sec

IB Data Speed (2009)

- Because links can be bonded or aggregated – they are usually
  - 1X
  - 4x
  - 12X

Effective theoretical throughput in different configurations

	Single (SDR)	Double (DDR)	Quad (QDR)
1X	2 Gbit/s	4 Gbit/s	8 Gbit/s
4X	8 Gbit/s	16 Gbit/s	32 Gbit/s
12X	24 Gbit/s	48 Gbit/s	96 Gbit/s

IB Signaling Speed (upto GDR)

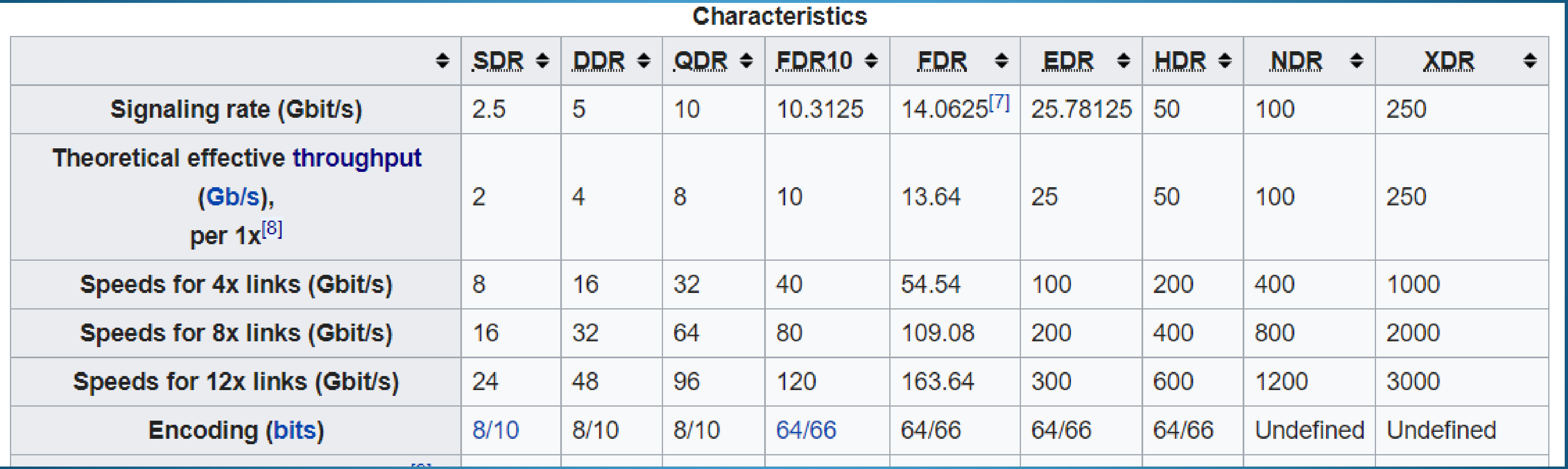
- (2009-2019) Supported:
  - Federated Data Rate (FDR), Enhanced Data Rate (EDR) and later HDR
- (2023) Now supports:
  - Next Data Rate (NDR)
  - eXtended Data Rate (XDR) and now
  - GDR (Greater Data Rate)

IB Signaling Speed (post QDR)

- (2019) Supports FDR, EDR, HDR, NDR and XDR
- FDR - Fourteen Data Rate, (exclude FDR-10)
- EDR - Enhanced Data Rate,
- HDR – High Data Rate,
- *NDR – Next data Rate, (see next table)*
- *XDR – eXtended Data Rate (see next table)*
- *Changed signaling pattern from 8B/10B encoding to 64/66*
- In other words: 66 bits carries 64 bits of data
  - Audience Pop Quiz – new efficiency??
-

IB Data Speed (2019)

- Because links can be bonded or aggregated – they are usually
  - 1X
  - 4x
  - 12X

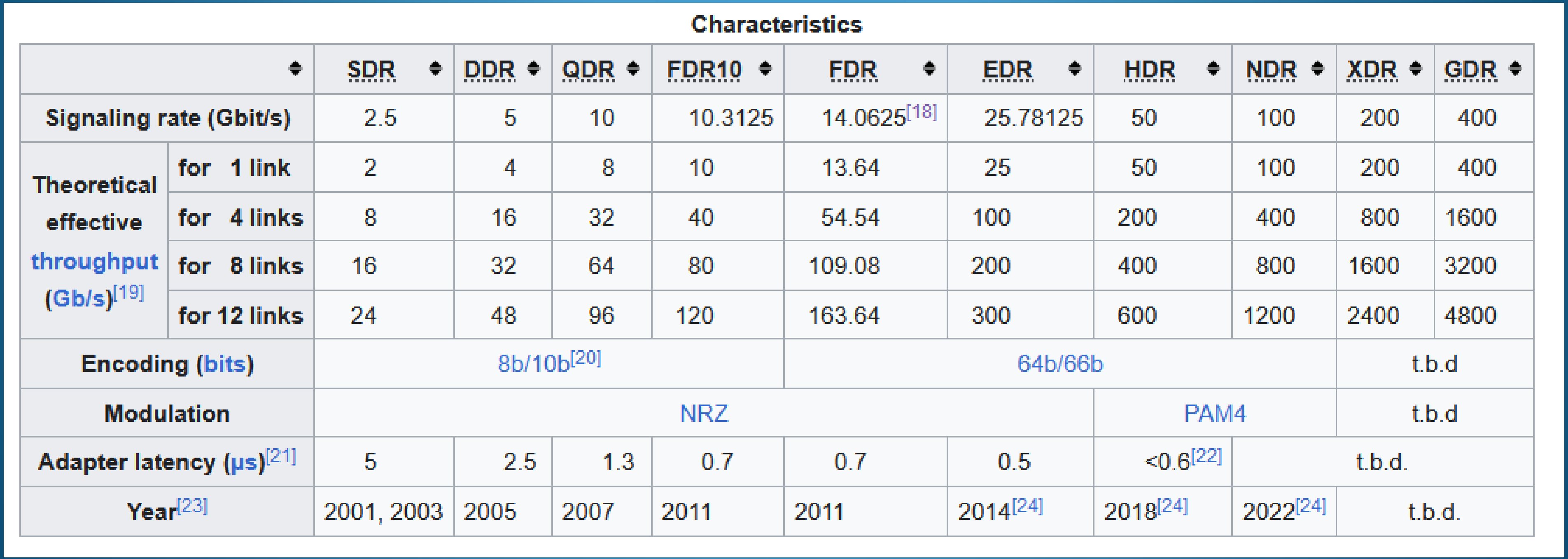


Characteristics

	SDR	DDR	QDR	FDR10	FDR	EDR	HDR	NDR	XDR
Signaling rate (Gbit/s)	2.5	5	10	10.3125	14.0625[7]	25.78125	50	100	250
Theoretical effective throughput (Gb/s), per 1x[8]	2	4	8	10	13.64	25	50	100	250
Speeds for 4x links (Gbit/s)	8	16	32	40	54.54	100	200	400	1000
Speeds for 8x links (Gbit/s)	16	32	64	80	109.08	200	400	800	2000
Speeds for 12x links (Gbit/s)	24	48	96	120	163.64	300	600	1200	3000
Encoding (bits)	8/10	8/10	8/10	64/66	64/66	64/66	64/66	Undefined	Undefined

IB Data Speed (2023)

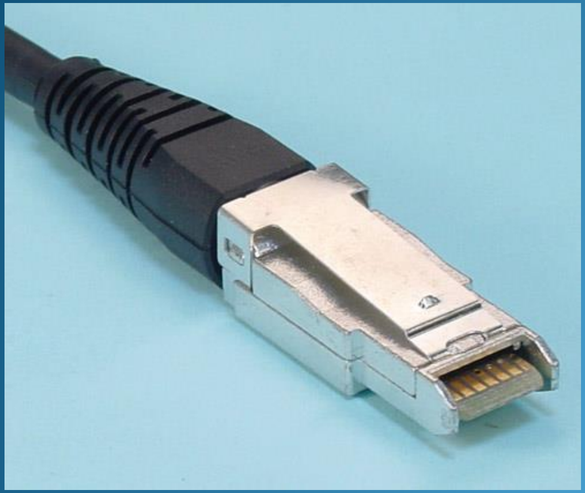
- Because links can be bonded or aggregated – they are usually
  - 1X
  - 4x
  - 12X



Characteristics

		SDR	DDR	QDR	FDR10	FDR	EDR	HDR	NDR	XDR	GDR
Signaling rate (Gbit/s)		2.5	5	10	10.3125	14.0625[18]	25.78125	50	100	200	400
Theoretical effective throughput (Gb/s)[19]	**for 1 link**	2	4	8	10	13.64	25	50	100	200	400
	for 4 links	8	16	32	40	54.54	100	200	400	800	1600
	for 8 links	16	32	64	80	109.08	200	400	800	1600	3200
	for 12 links	24	48	96	120	163.64	300	600	1200	2400	4800
Encoding (bits)		8b/10b[20]				64b/66b				t.b.d	
Modulation		NRZ						PAM4		t.b.d	
Adapter latency (µs)[21]		5	2.5	1.3	0.7	0.7	0.5	<0.6[22]	t.b.d.		
Year[23]		2001, 2003	2005	2007	2011	2011	2014[24]	2018[24]	2022[24]	t.b.d.	

IB Cables/Connectors



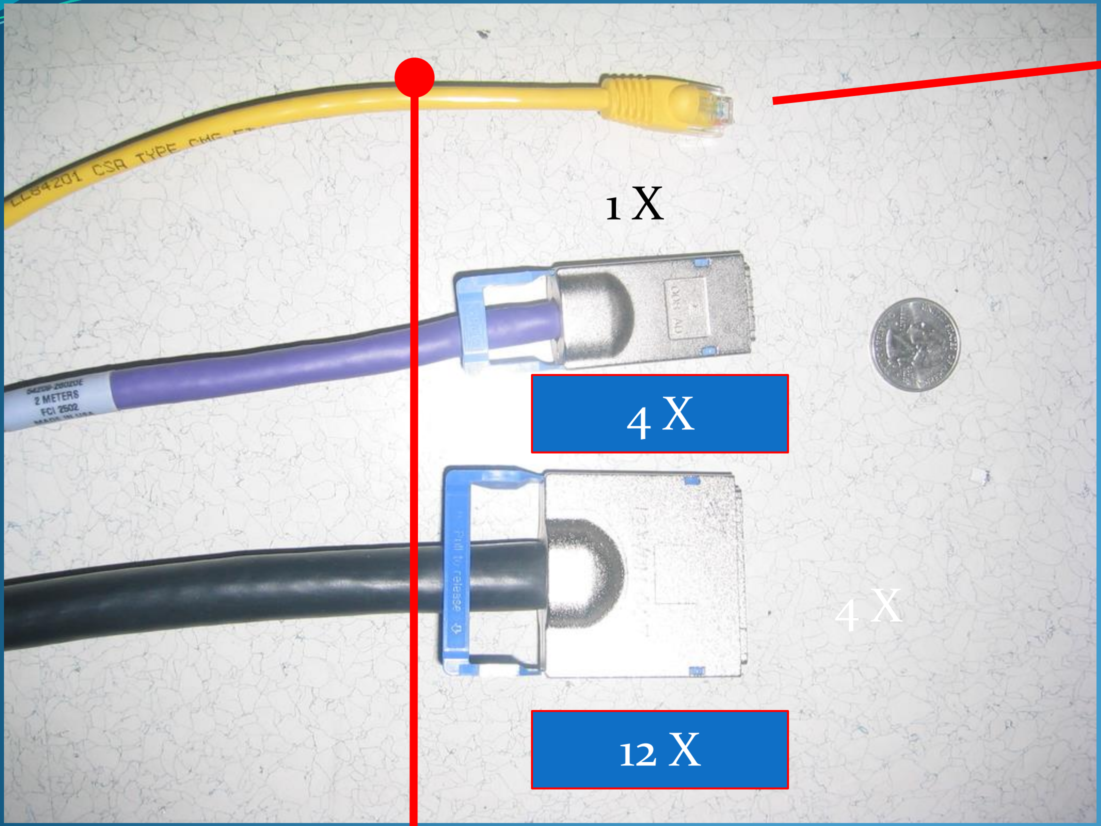
1 X

4 X

12 X

Examples of IB connectors and cables

IB Cables/Connectors



Ethernet RJ45 connector

Relative Sizes

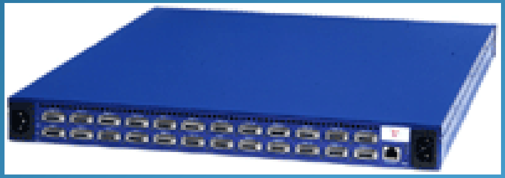
IB Cables/Connectors

- Industry standard Media types
  - Copper: 7 Meter QDR , 3 METER FDR
  - Fiber: 100/300m QDR & FDR
- 64/66 encoding on FDR links
  - Encoding makes it possible to send digital high speed signals enhances performance & bandwidth effectiveness
  - X actual data bits are sent on the line by Y signal bits
  - 64/66 * 56 = 54.6Gbps
- 8/10 bit encoding (DDR and QDR)
  - X/Y line efficiency (example 80% * 40 = 32Gbps)

4X QSFP Fiber 4X QSFP Copper

IB Switches

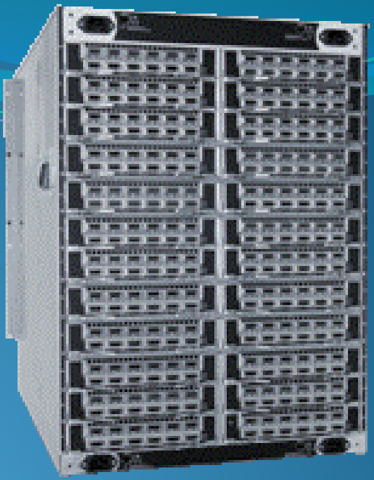
Flextronics Reindeer Switch (24 ports)



Voltaire (24 ports)

Microway FasTree switch (72 ports)

More IB Switches



Cisco 7008p (96 ports)

Voltaire (96 ports)

QLOGIC 9000 series (288 ports)

WORLD’s BIGGEST, BADDEST Infiniband SWITCH

SUN MAGNUM SWITCH
(3456 ports, SDR or DDR,
12X, over 110 Terabits/second,

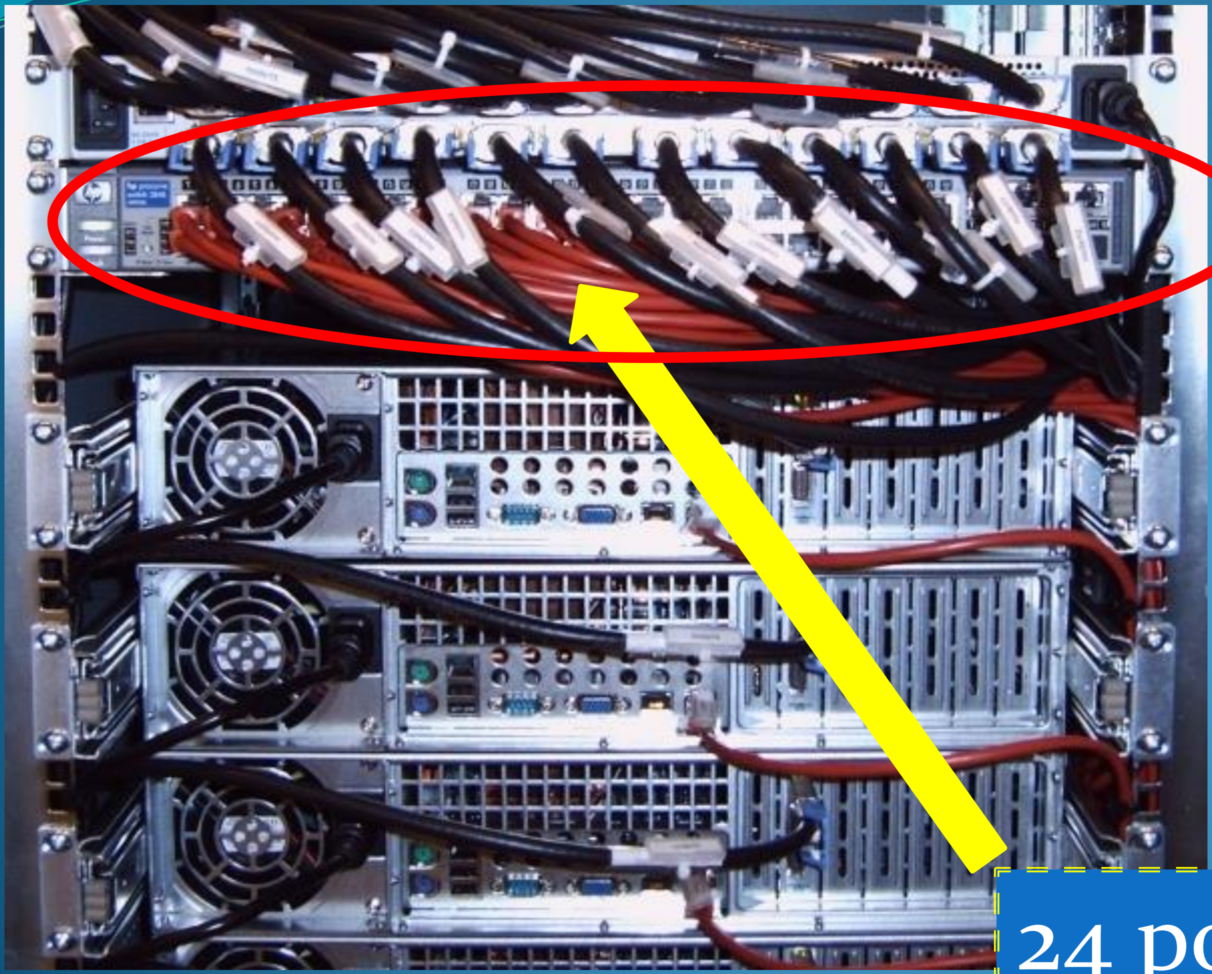
1 u-second latency, non-blocking,

With over 720 IB elements!)

IB ecosystem/family

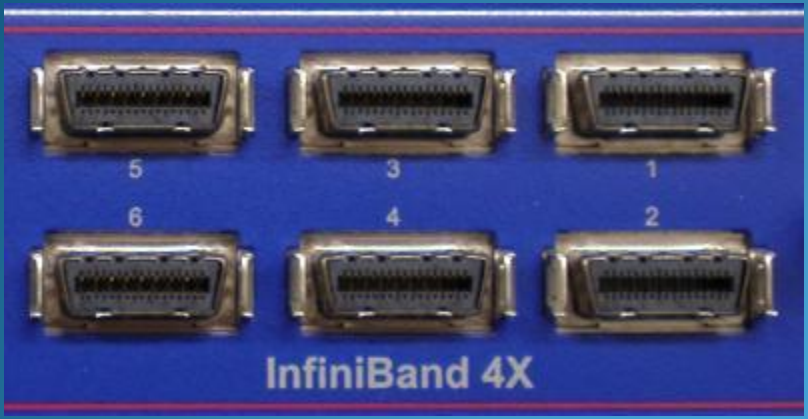
ICs Adapter Cards Switches/Gateways Host/Fabric Software

Behind the curtain……



24 port cabling

Typical IB back panel looks like this...

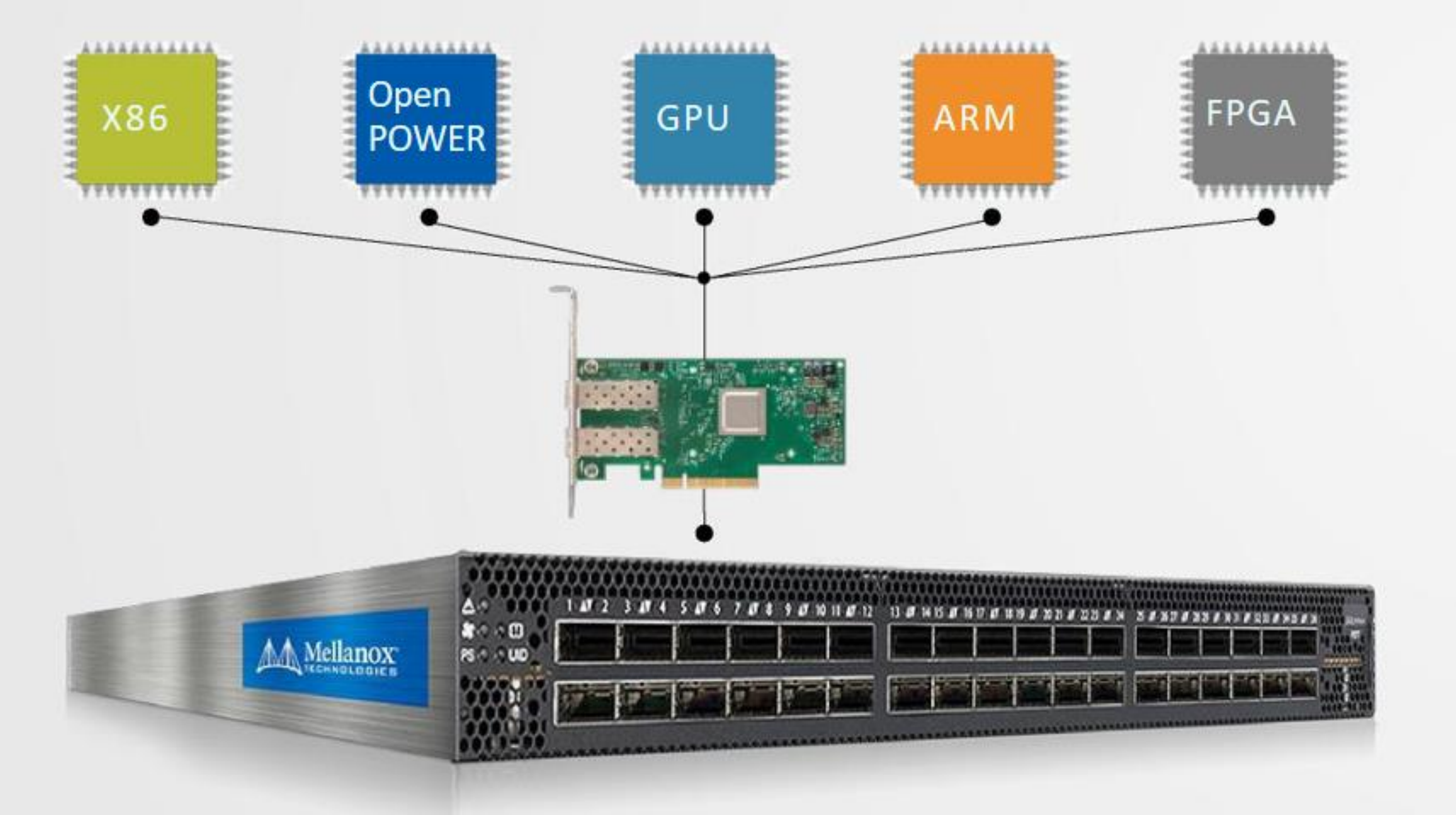


6 ports 4X

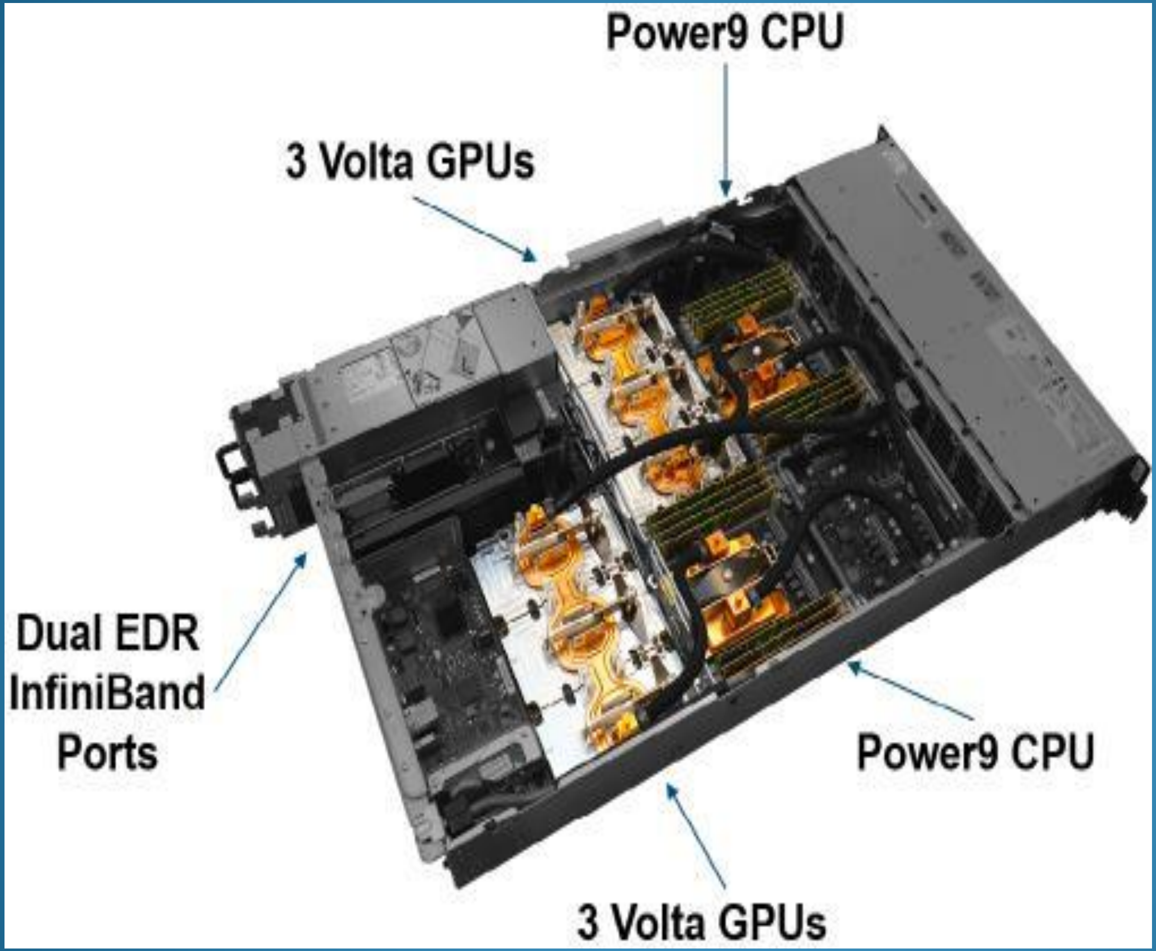
IB Advantages

- Origins were from Server to Server
- Designed from the start to support
  - Quality of Service
  - Failover and
  - Scalability
- The IB architecture specification defines
  - a connection between processor nodes and any I/O nodes
  - such as storage devices.
  - Connection logic is a superset of the Virtual Interface Architecture.
  - VIA is Intel’s contribution to IB
- Host Channel Adaptor (HCA)
- Target Channel Adaptor (TCA)

IB Advantages

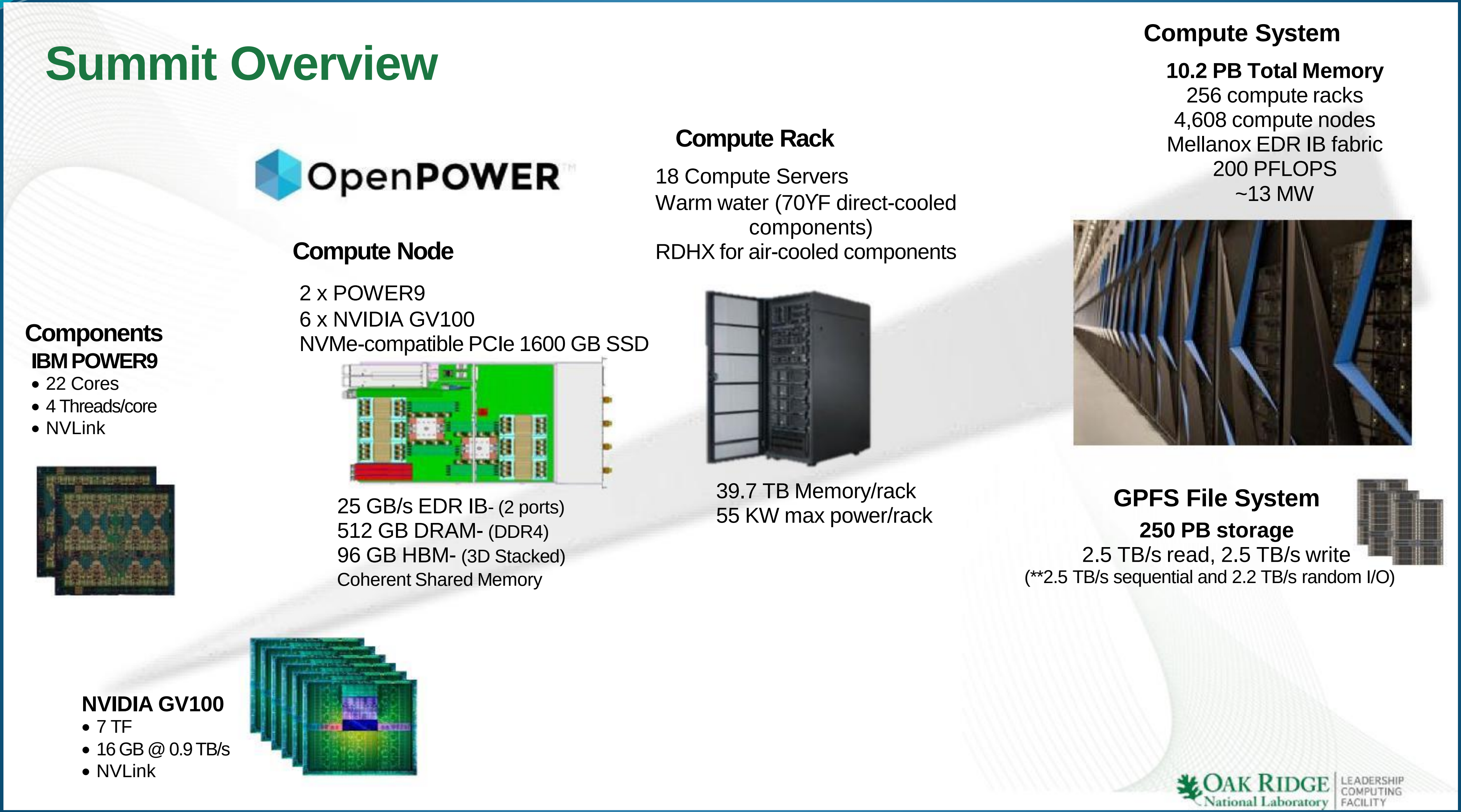


IB Advantages



Summit Server Configuration

IB Advantages



Summit Overview

OpenPOWER

Components

IBM POWER9

- 22 Cores
- 4 Threads/core
- NVLink

NVIDIA GV100

- 7 TF
- 16 GB @ 0.9 TB/s
- NVLink

Compute Node

2 x POWER9
6 x NVIDIA GV100
NVMe-compatible PCIe 1600 GB SSD

25 GB/s EDR IB- (2 ports)
512 GB DRAM- (DDR4)
96 GB HBM- (3D Stacked)
Coherent Shared Memory

Compute Rack

18 Compute Servers
Warm water (70YF direct-cooled components)
RDHX for air-cooled components

39.7 TB Memory/rack
55 KW max power/rack

Compute System

10.2 PB Total Memory
256 compute racks
4,608 compute nodes
Mellanox EDR IB fabric
200 PFLOPS
~13 MW

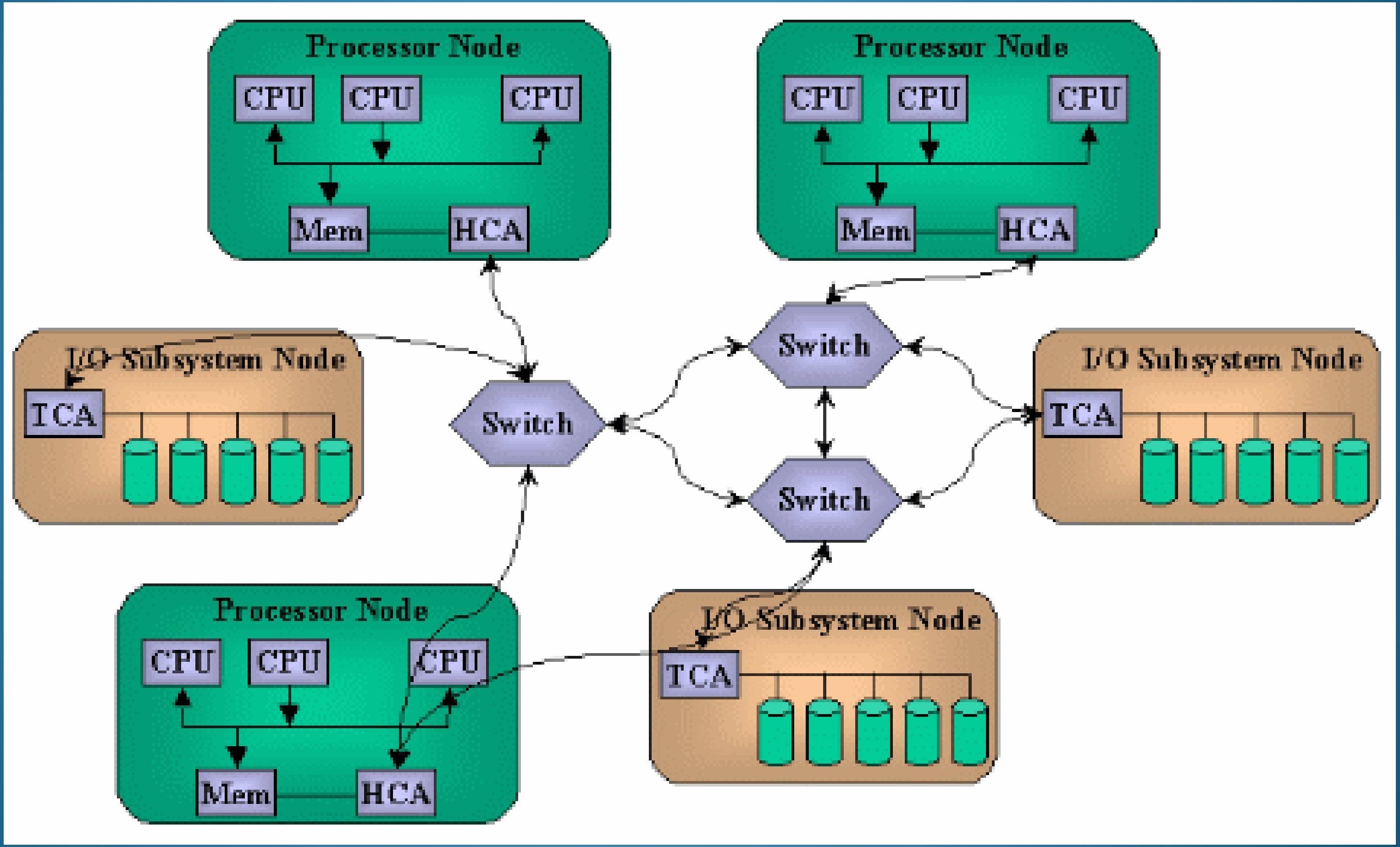
GPFS File System

250 PB storage
2.5 TB/s read, 2.5 TB/s write
(**2.5 TB/s sequential and 2.2 TB/s random I/O)

OAK RIDGE National Laboratory | LEADERSHIP COMPUTING FACILITY

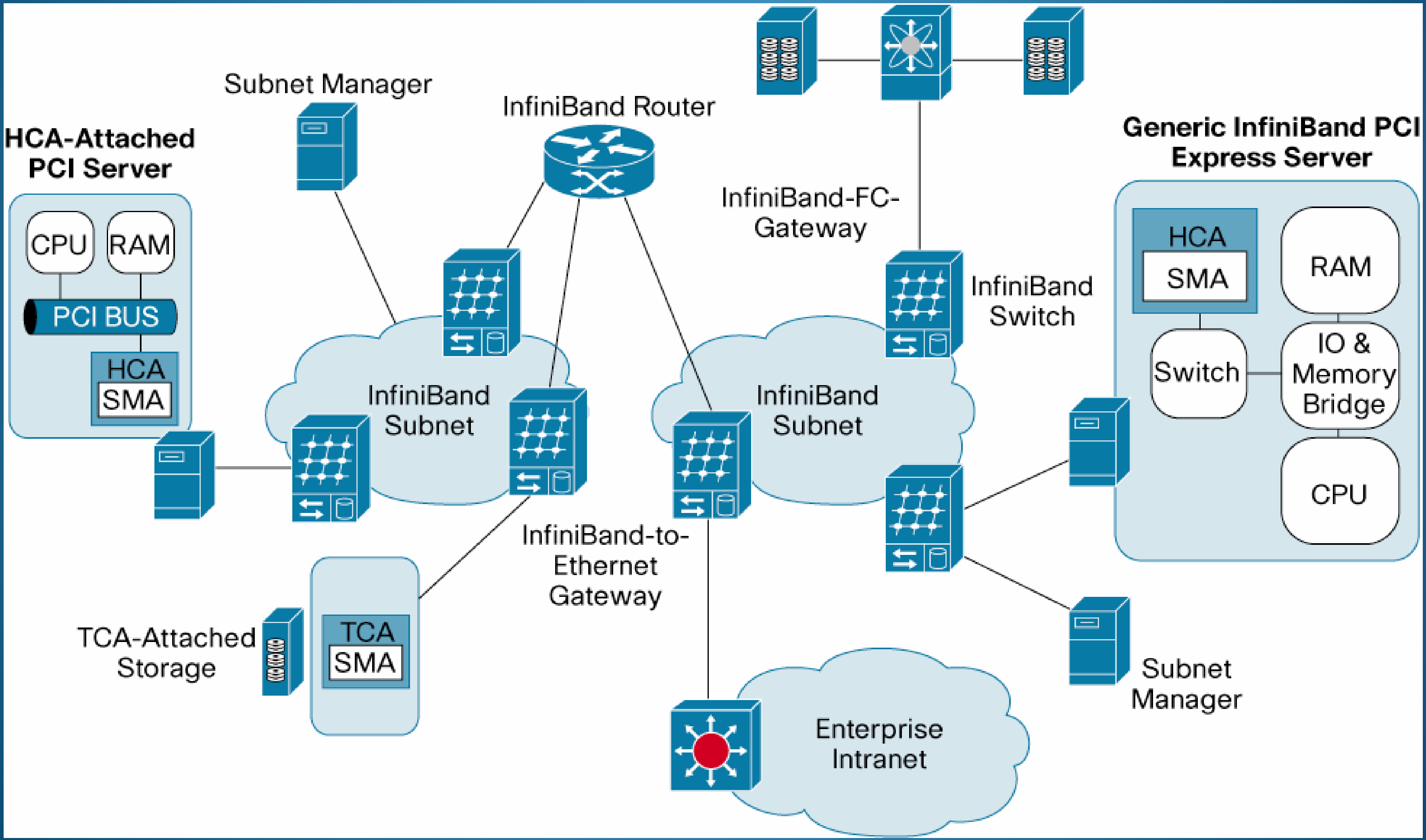
IB logical picture (example A)

- What does it looks like?



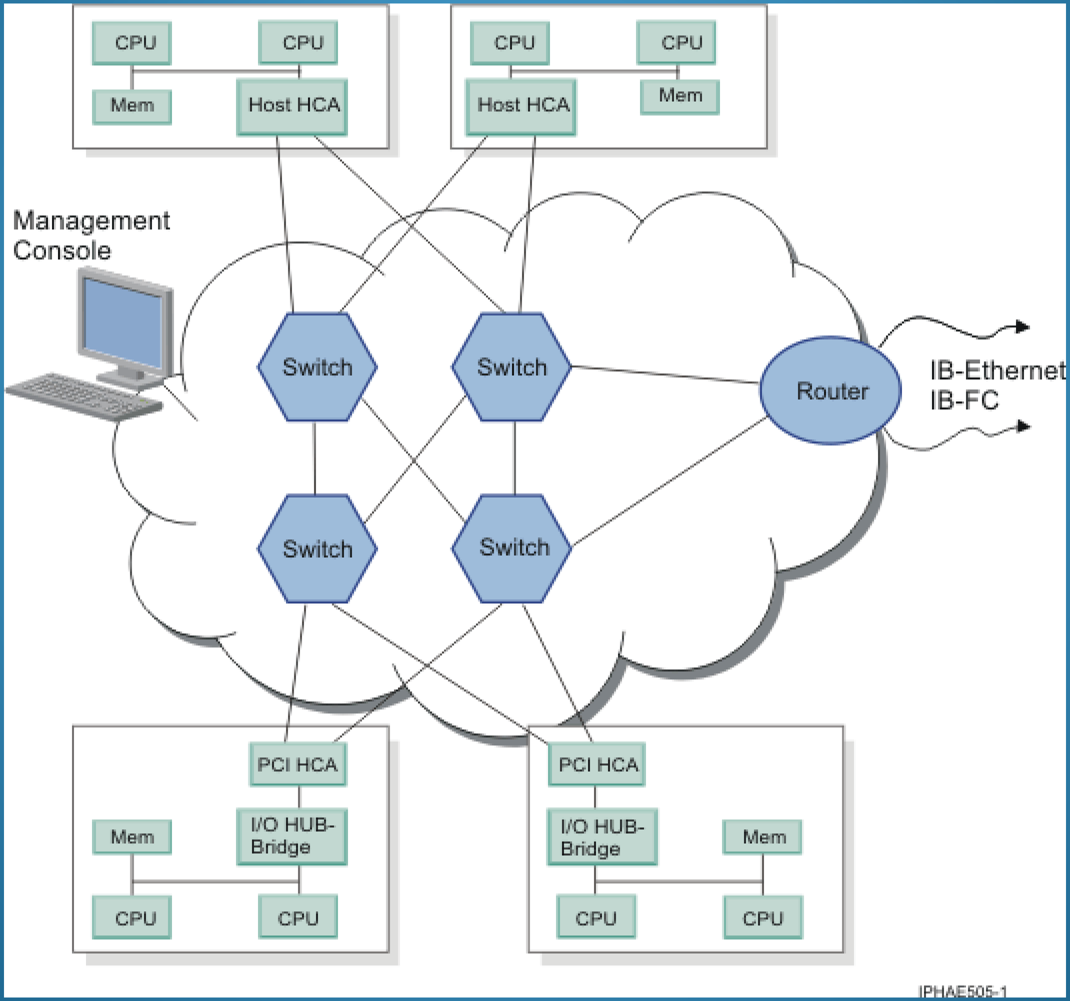
IB logical picture (example B)

- What does it looks like?



IB logical picture (example C)

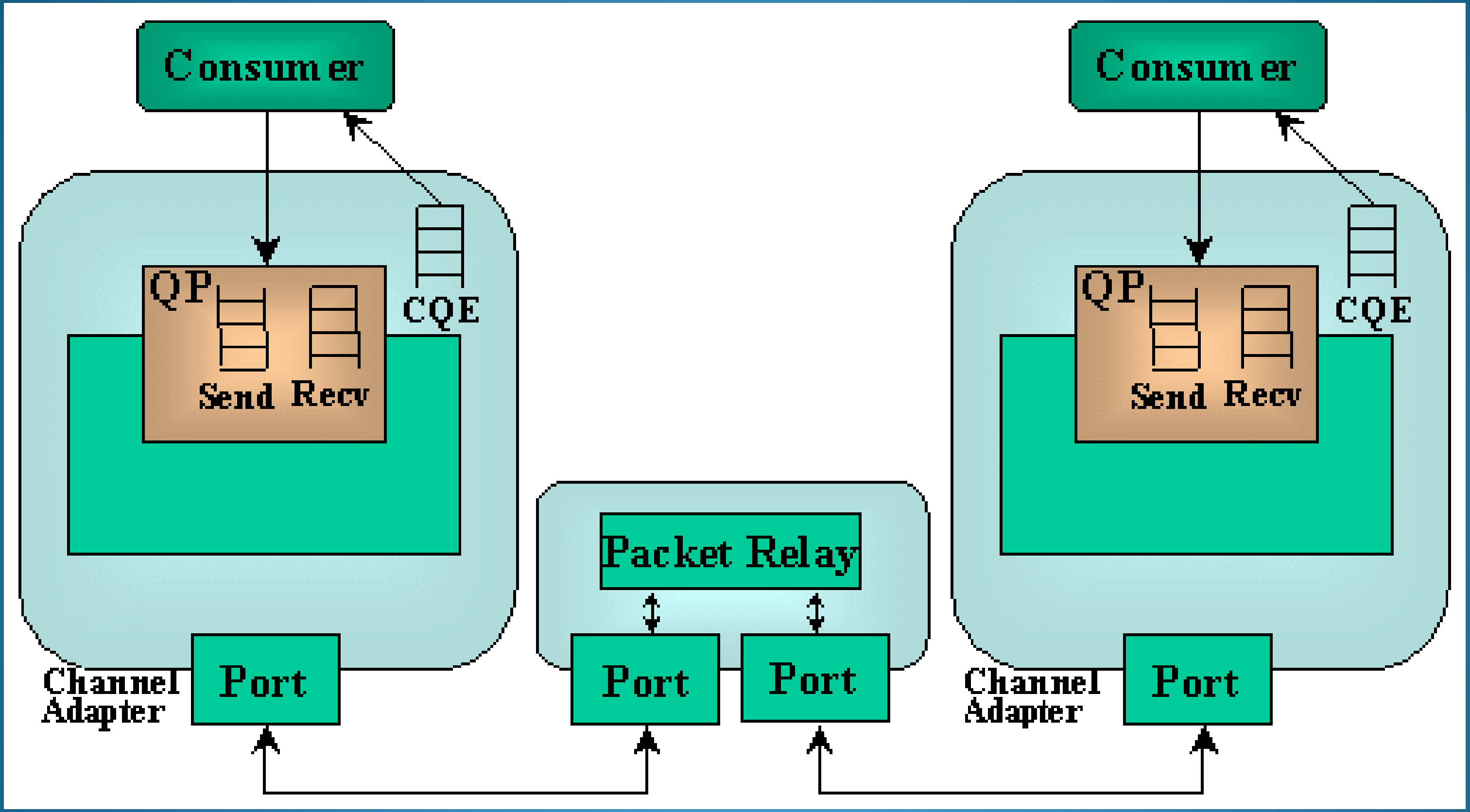
- Another view:



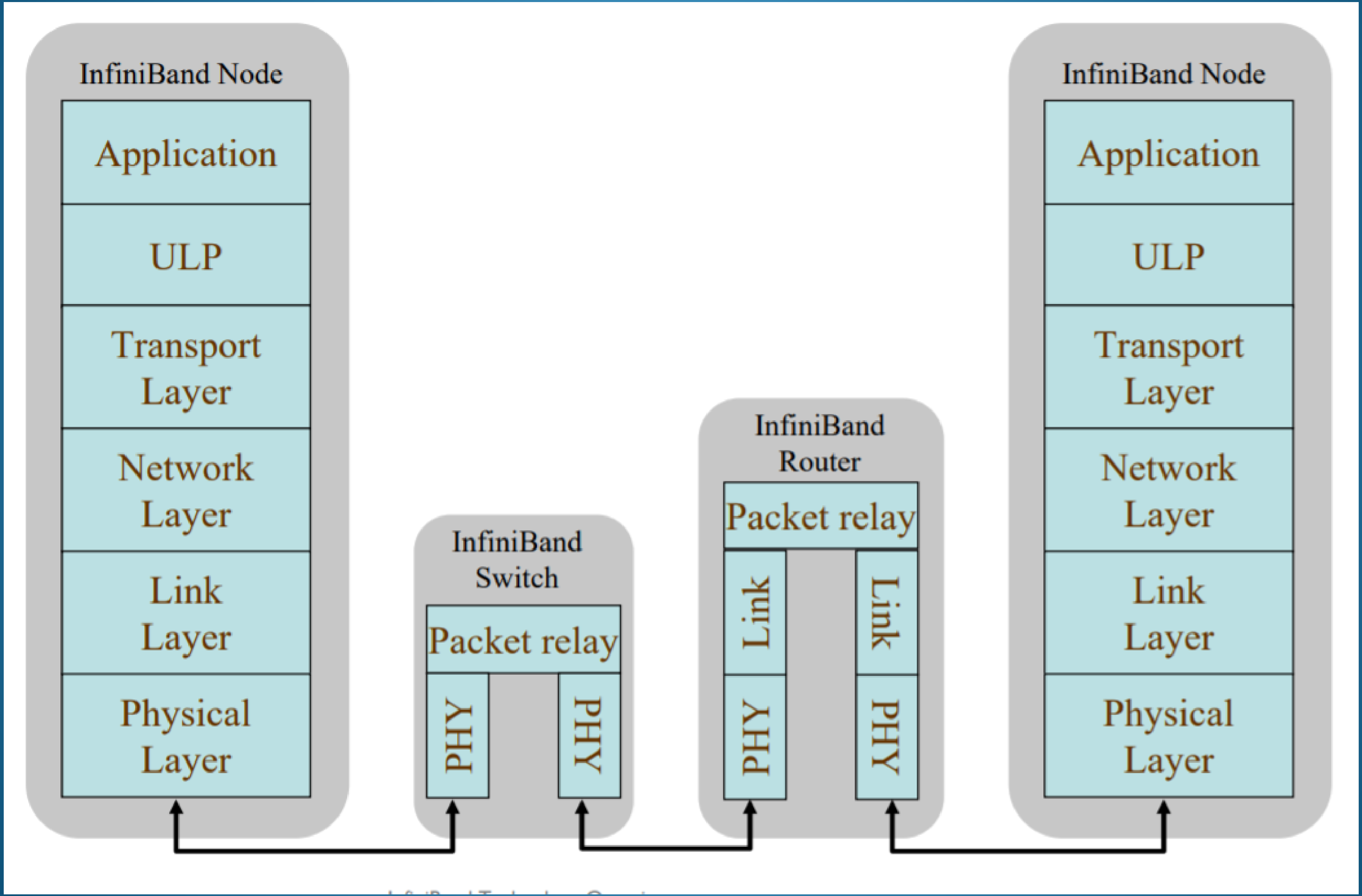
IB advantages

- Having multiple paths available means:
- In getting the data from one node to another,
- IB is able to achieve transfer rates at the full channel capacity
- Avoiding congestion issues that arise in a shared bus architecture.
- Having alternative paths results in increased reliability
- Scalability is also built in since overhead is low

IB Design (simple)

- 

IB Design (simple)



IB Features

- Designed for Remote Data Memory Access (RDMA) and
- Supports Sockets Direct Protocol (SDP)
- In addition it also supports:
  - IP over IB
  - SCSI over IB
  - FC over IB
  - *ad infinitum…..*
-

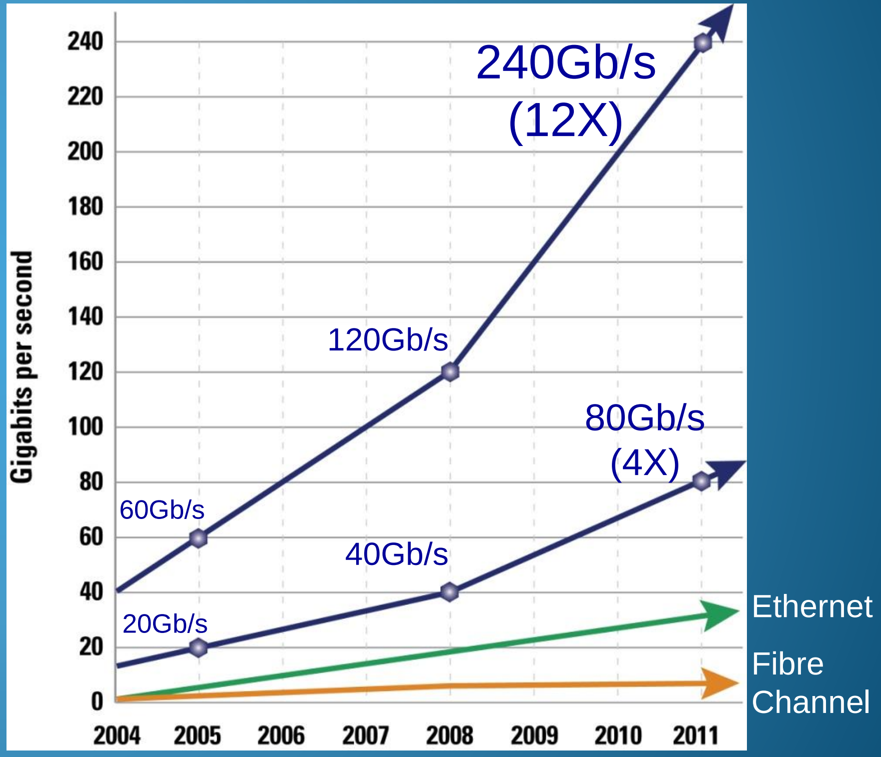
IB Application Advantages

- Due to all of the preceding advantages -
- *One also gets:*
  - *Very good* Latency and
  - Bandwidth

IB Application Advantages (2009)

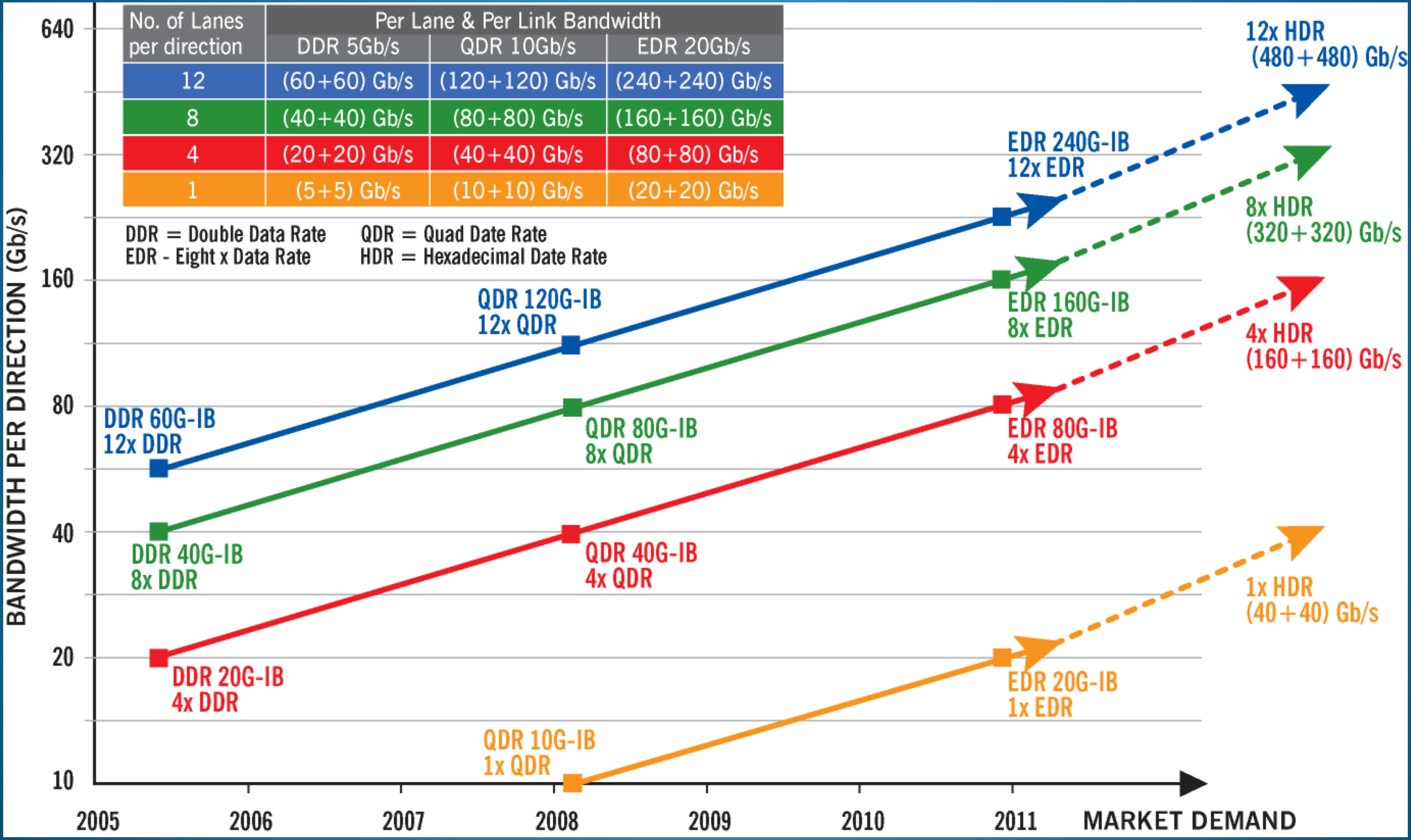
- Published Industry Standard
  - Hardware, software, cabling, management
- Price and Performance
  - 40Gb/s node-to-node
  - 120Gb/s switch-to-switch
  - 1 µs application latency
- Reliable w/congestion management
- Efficient
  - RDMA and Transport Offload
  - CPU focuses on application processing
- Scalable for Large Scale
- End-to-end quality of service
- I/O consolidation including storage

The InfiniBand Performance Gap is Increasing



InfiniBand Delivers the Lowest Latency

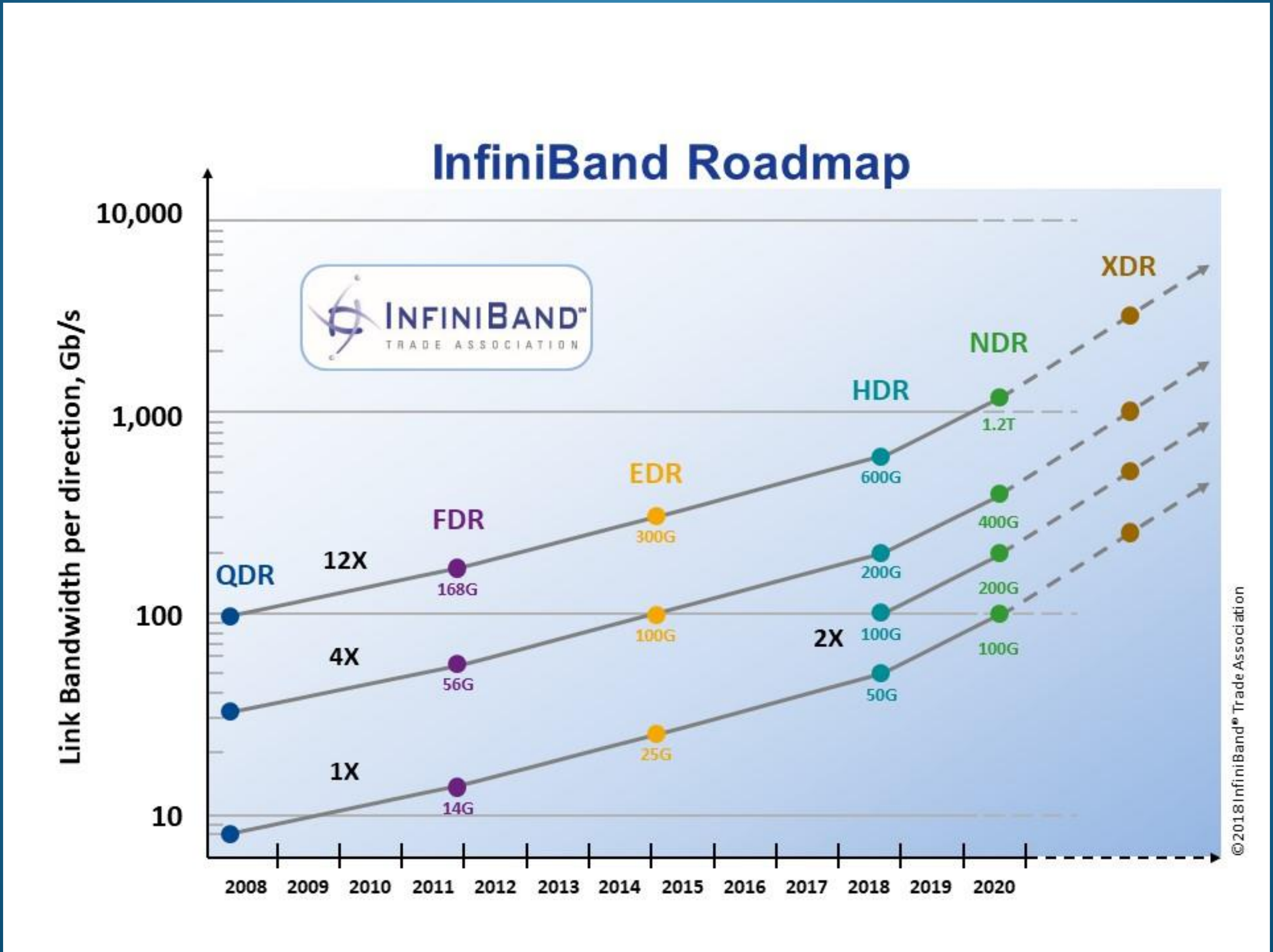
IB Application Advantages (2009)



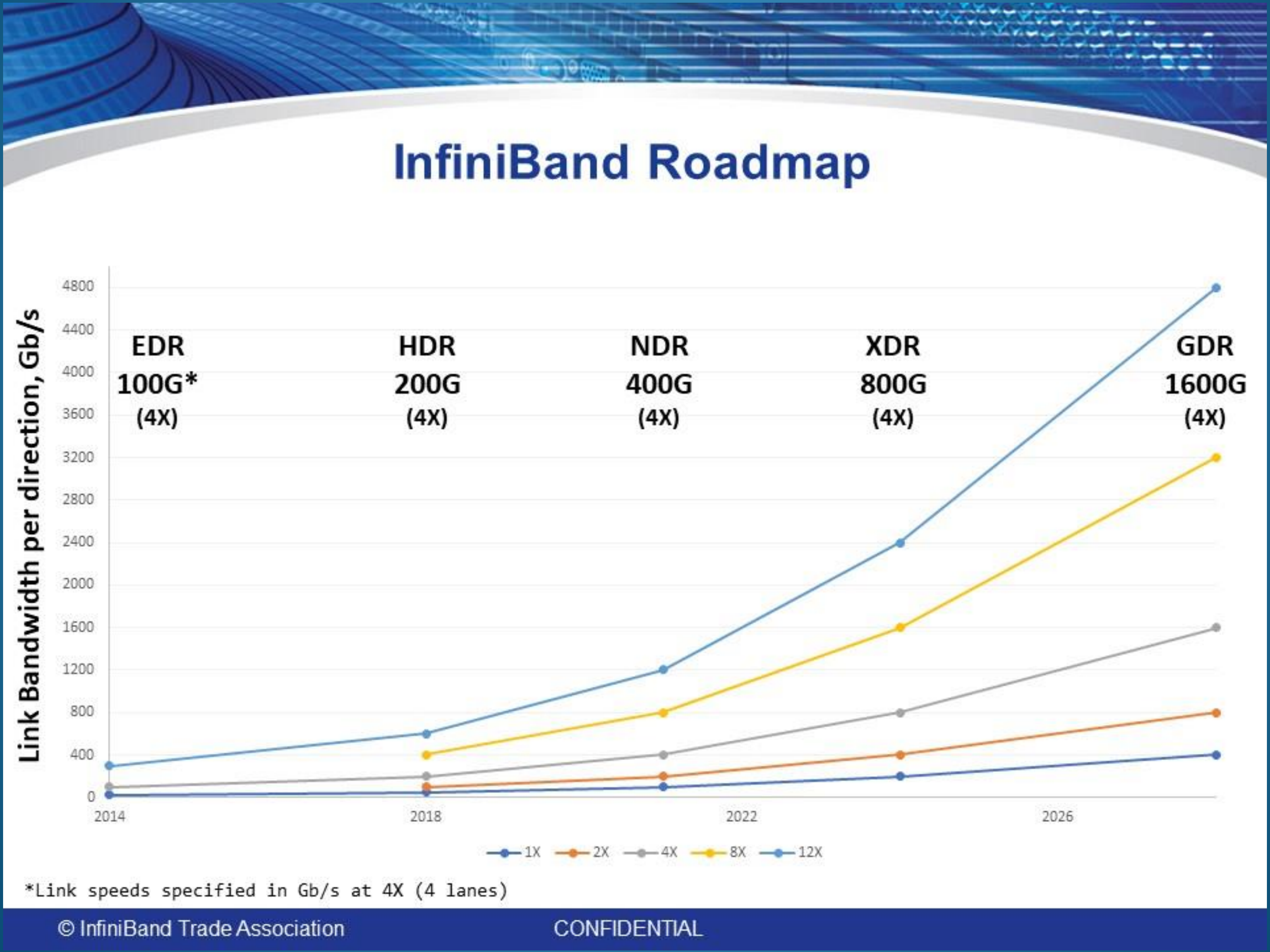
No. of Lanes per direction	Per Lane & Per Link Bandwidth		
	DDR 5Gb/s	QDR 10Gb/s	EDR 20Gb/s
12	(60+60) Gb/s	(120+120) Gb/s	(240+240) Gb/s
8	(40+40) Gb/s	(80+80) Gb/s	(160+160) Gb/s
4	(20+20) Gb/s	(40+40) Gb/s	(80+80) Gb/s
1	(5+5) Gb/s	(10+10) Gb/s	(20+20) Gb/s

DDR = Double Data Rate QDR = Quad Date Rate
EDR - Eight x Data Rate HDR = Hexadecimal Date Rate

IB Application Advantages (2019)



IB Application Advantages (2023)



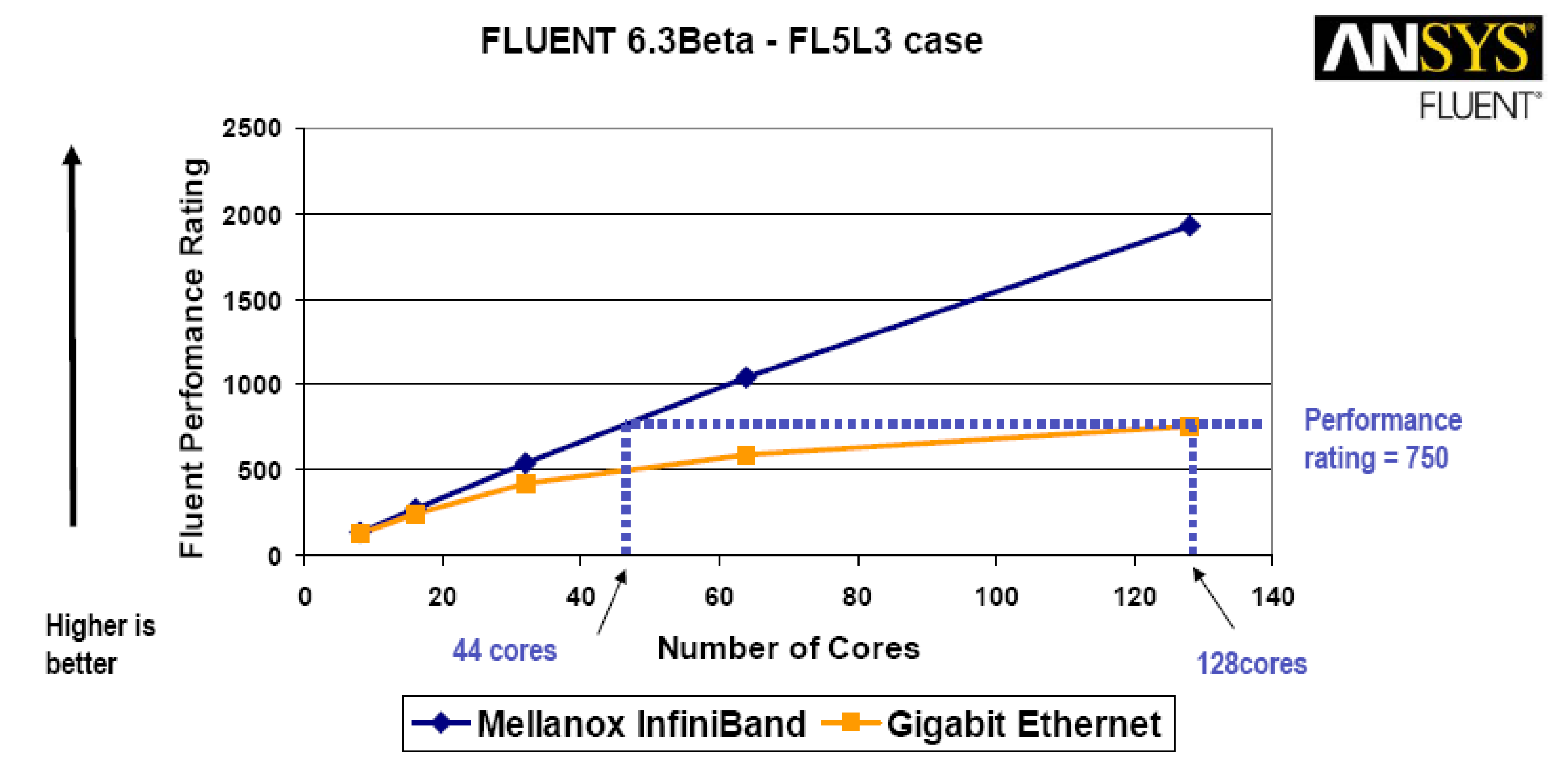
InfiniBand Roadmap

*Link speeds specified in Gb/s at 4X (4 lanes)

© InfiniBand Trade Association

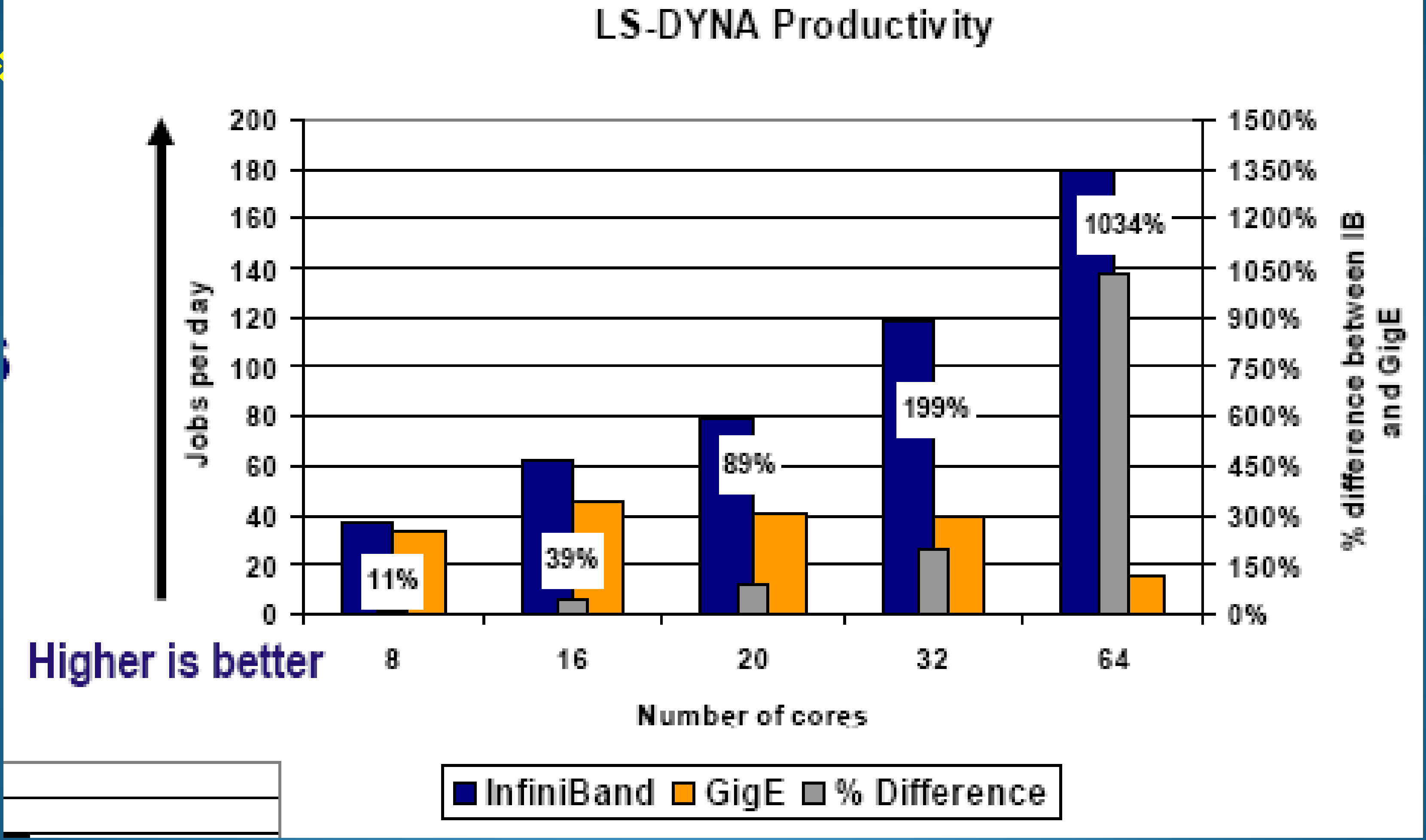
CONFIDENTIAL

IB Performance in HPC



SDR – Single Data Rate

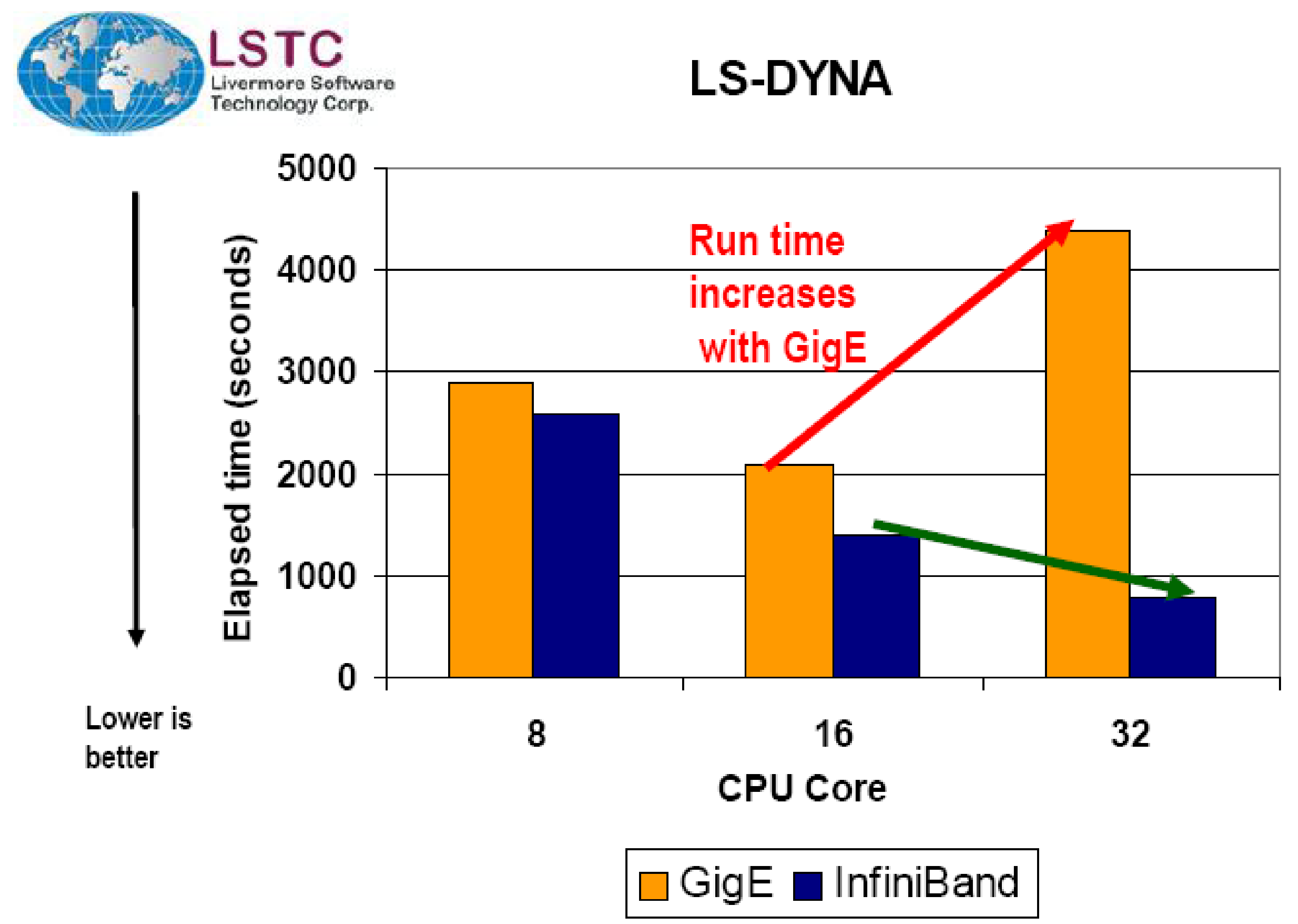
IB Performance in HPC



SDR – Single Data Rate

IB Performance in HPC

❖

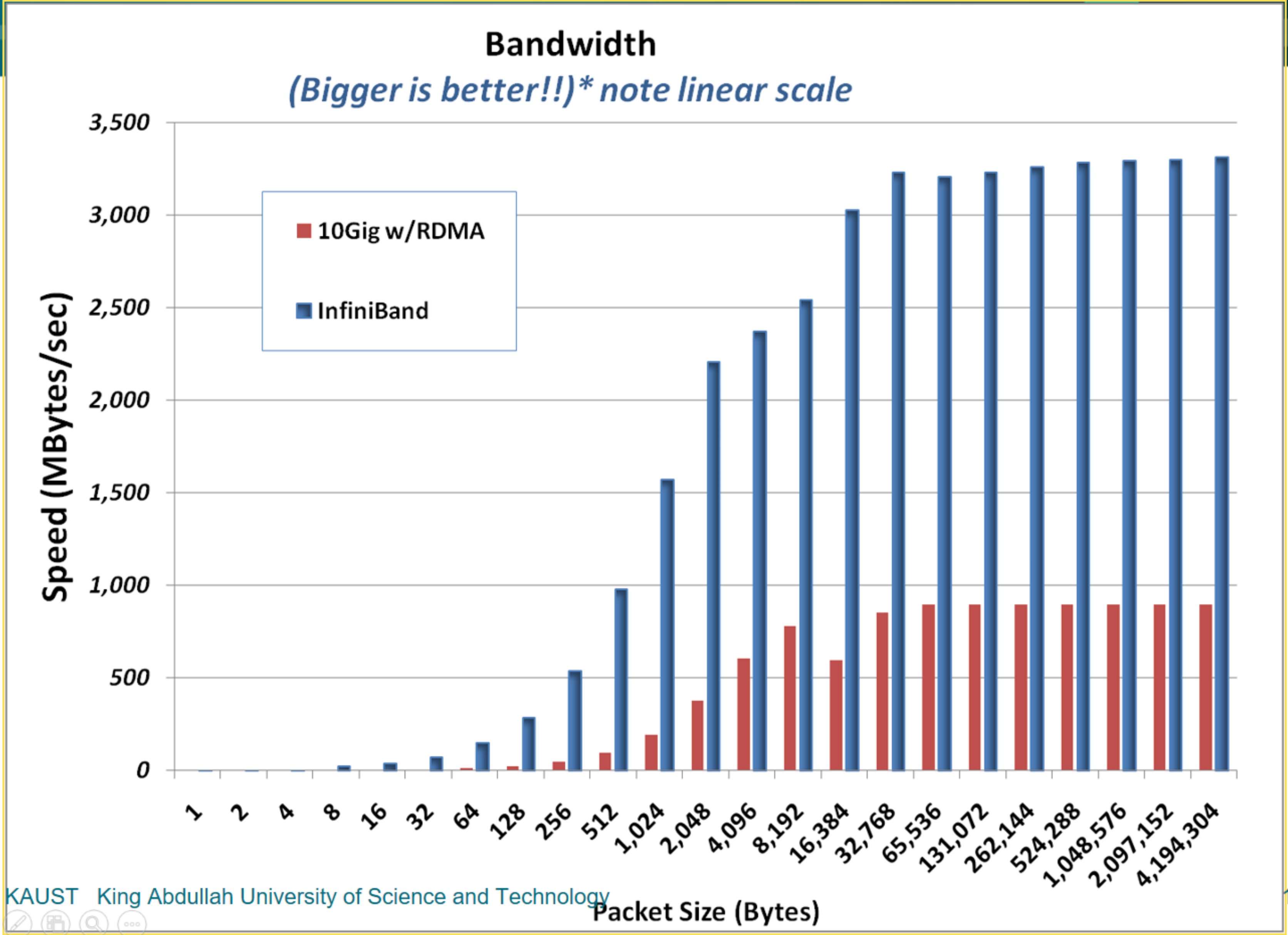


SDR – Single Data Rate

IB Performance in HPC*

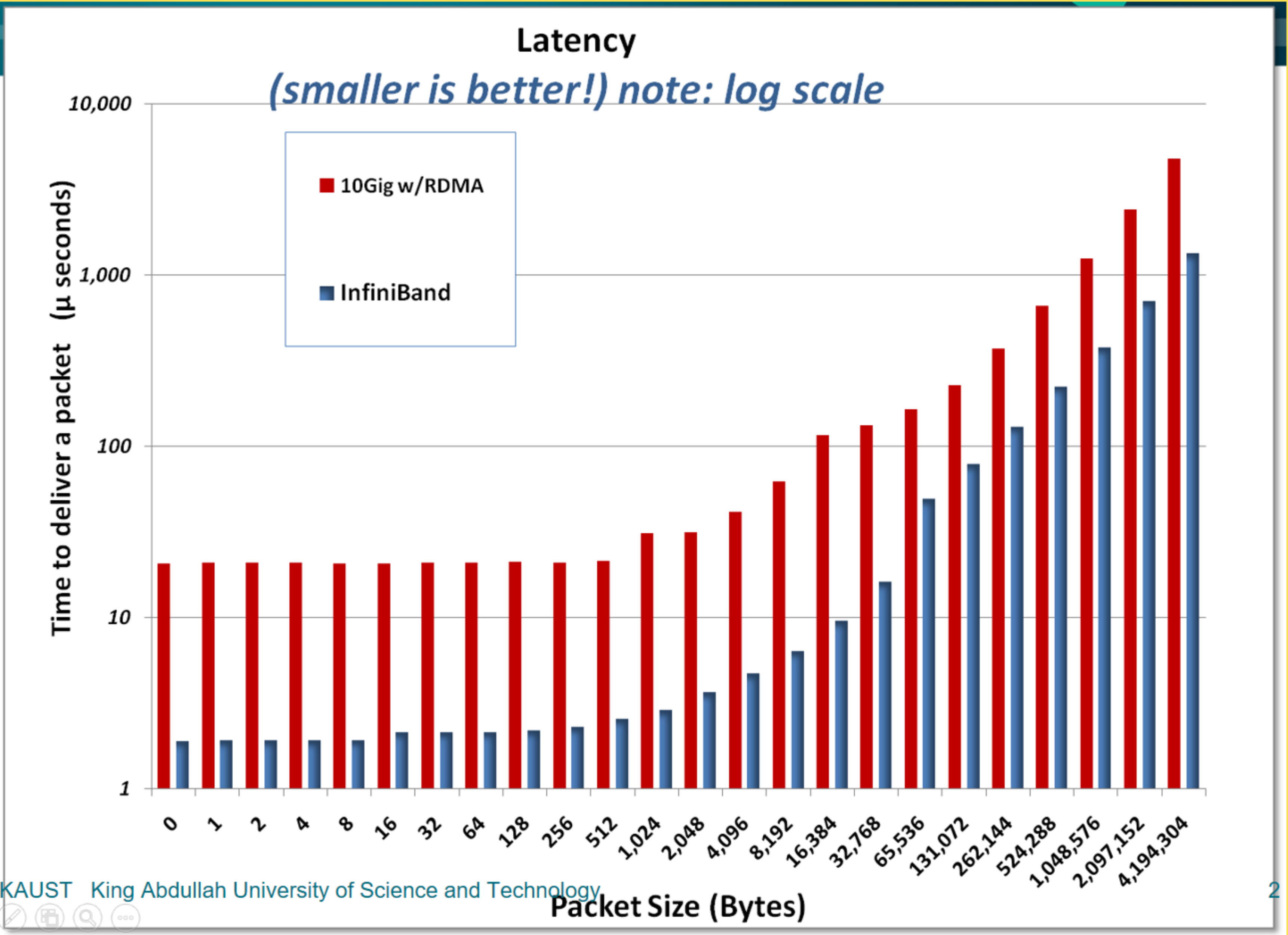
- Reflect my own benchmarking data
- Based on next gen IB roadmap in 2009
- Done in 2010

Positive IB experience

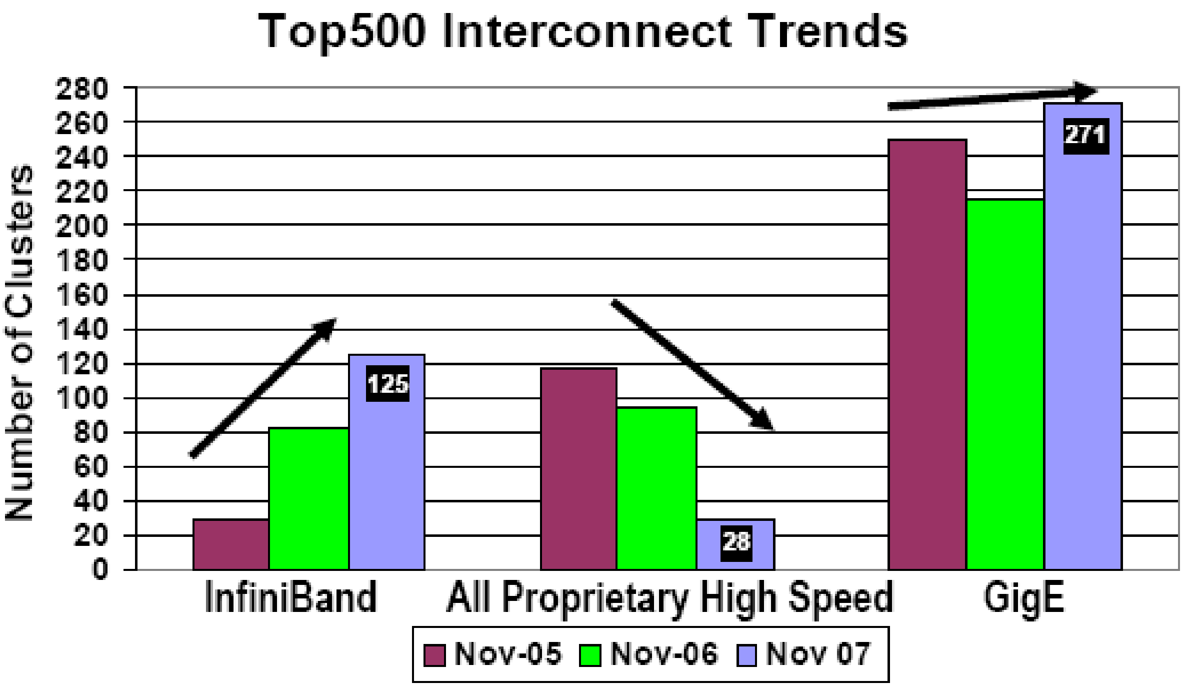


Positive IB experience

King Abdullah University of Science and Technology KAUST



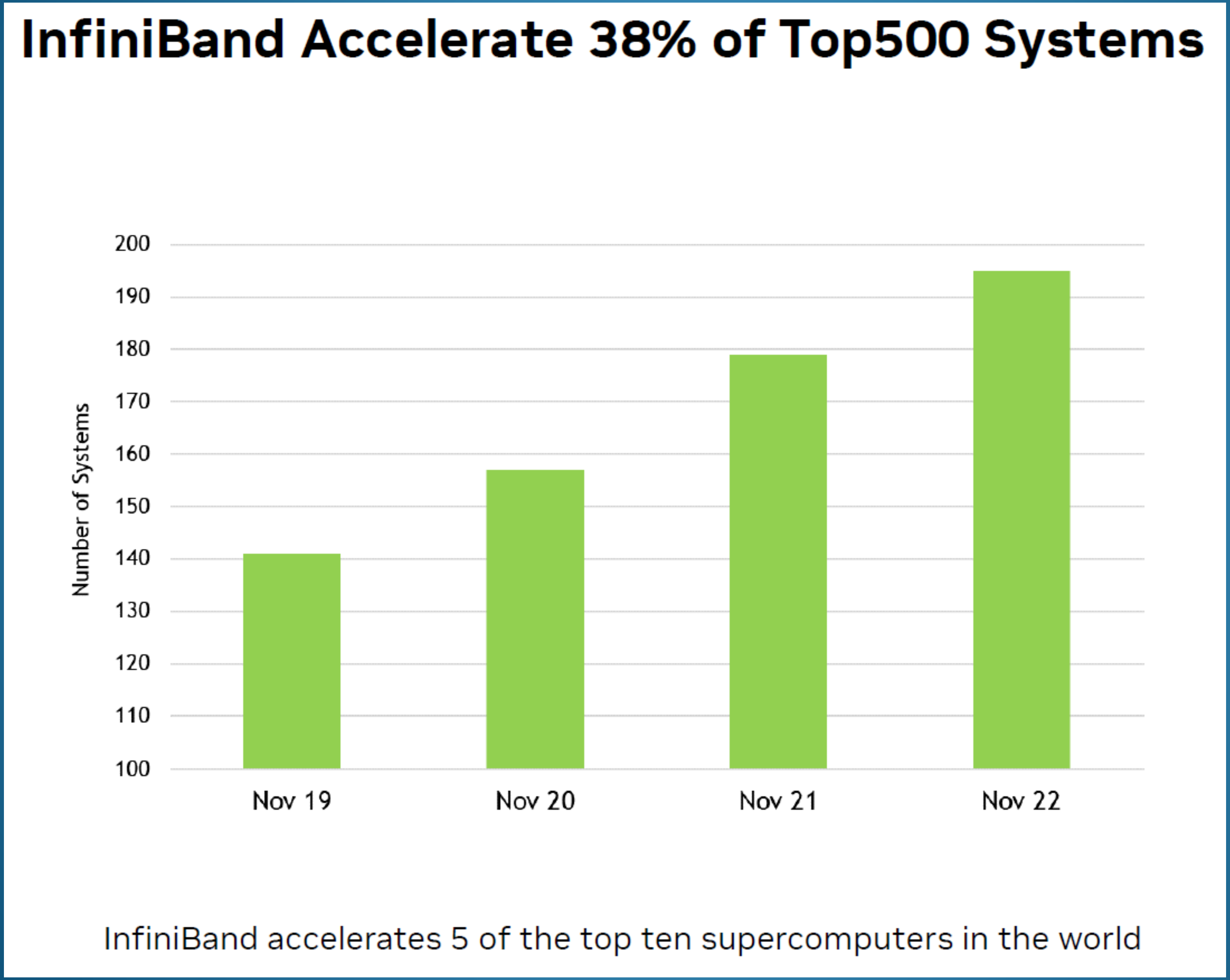
IB In the top500 (2008-09)



Growth rate from Nov 06 to Nov 07 (year)

- InfiniBand: +52%
- All Proprietary: -52%
- GigE: +26%

IB In the top500 (2023)



InfiniBand Accelerate 38% of Top500 Systems

InfiniBand accelerates 5 of the top ten supercomputers in the world

Resources

- Switch Vendors
  - http://www.infinibandta.org/kshowcase/view/catalogs_by_category?categories=2dce0fa800733b814dd6ba794b1afbffc77762fa
- Software Stacks
  - http://www.openib.org or
  - http://www.openfabrics.org
- Cable Vendors (slightly dated)
  - http://www.infinibandta.org/itinfo/IL/IL_Cable_2004-01.pdf
- Similar tutorials:
  - https://www.naddod.com/blog/top-10-advantages-of-infiniband

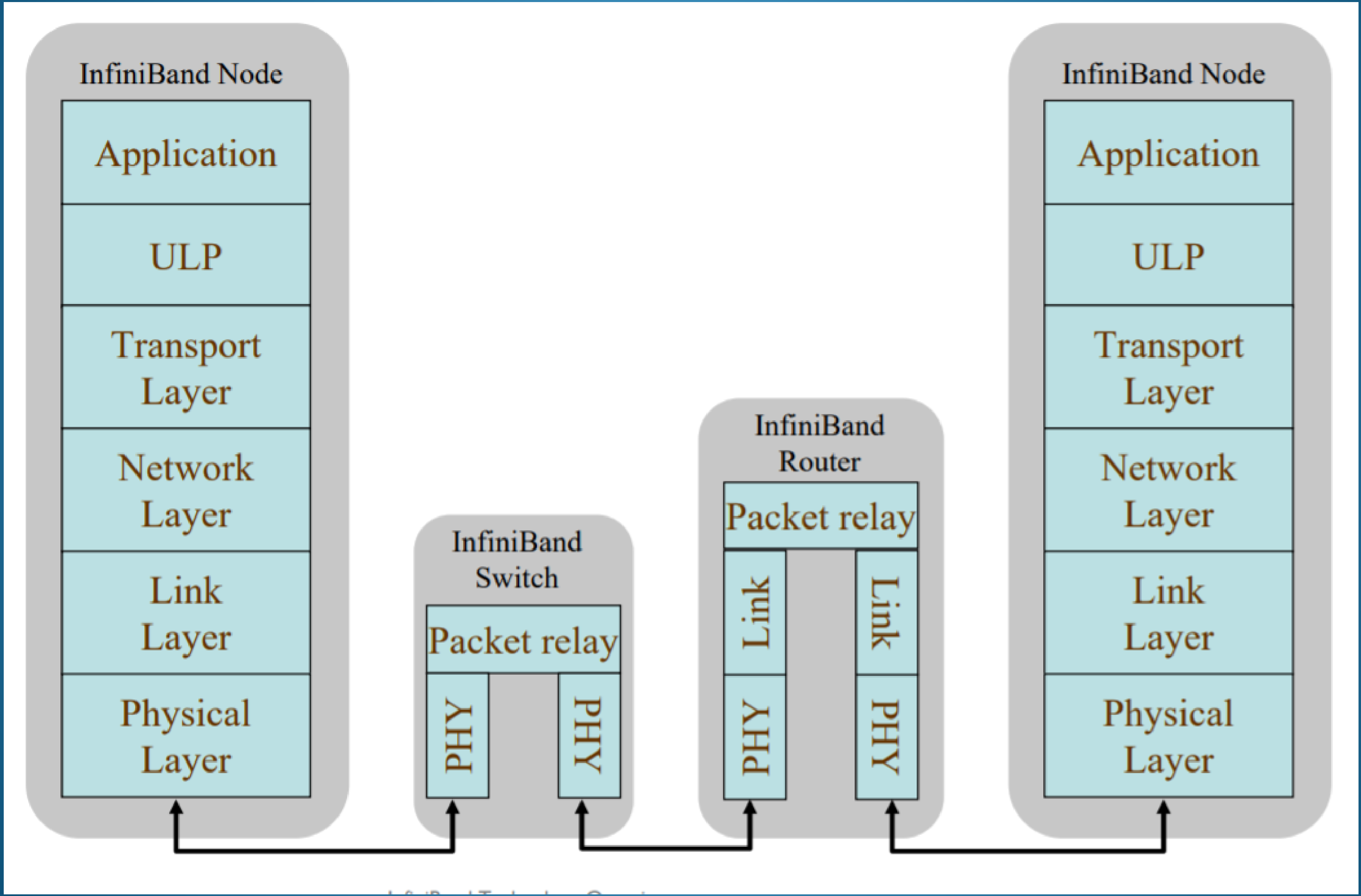
Revision History

Version 5

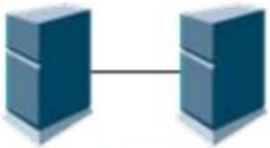
- Added 40 Gbit press release
- Added web links
- Added Slide Numbers
- Customized for FCA
- Add 2019 updates
- Add 2023 roadmap/updates

Boring Backup Slides

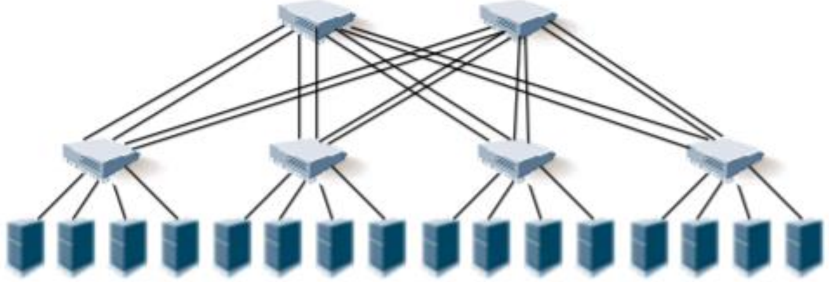
IB protocol


InfiniBand Node
Application
ULP
Transport Layer
Network Layer
Link Layer
Physical Layer
InfiniBand Switch
Packet relay
PHY
PHY
InfiniBand Router
Packet relay
Link
Link
PHY
PHY
InfiniBand Node
Application
ULP
Transport Layer
Network Layer
Link Layer
Physical Layer

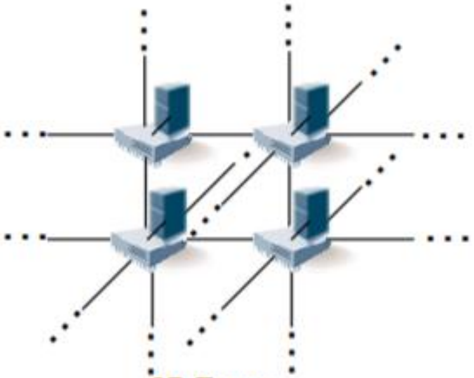
IB Topologies



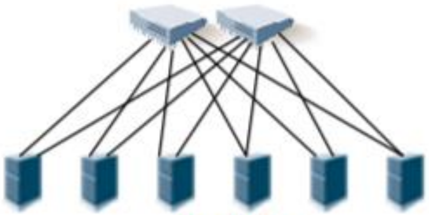
Back to Back



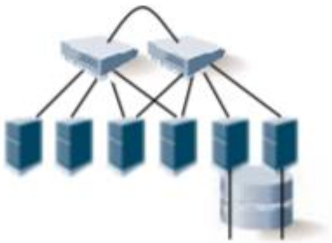
2 Level Fat Tree



3D Torus



Dual Star



Hybrid

- Example topologies commonly used
- Architecture does not limit topology
- Modular switches are based on fat tree architecture

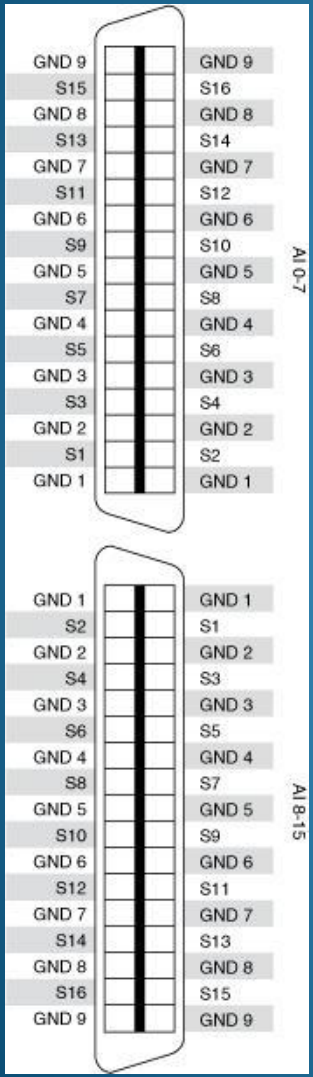
IB signals/pin

- InfiniBand cables use a connector based on the microGiGaCN series developed by Fujitsu. The connector provides excellent signal integrity with a shield plate between input signal pairs.
- Each signal pair is shielded internally, resulting in tight 10% impedance matching, and near-end crosstalk of less than 4% at 50psec rise time.
- Additionally, this design minimizes skew, crosstalk and EMI.
- IB cables are typically LVDS, consumes very little power (milliwatts/pair)

IB signals/pin

- What is LVDS?
- Low-voltage differential signaling, is an electrical signaling system that can run at very high speeds over cheap, twisted-pair copper cables.
- Introduced in 1994, and has since become very popular in computers, where it forms part of very high-speed networks and computer buses.
- Standards document
  - ANSI/TIA/EIA-644-A (published in 2001)
  - Also used in
    - HyperTransport, FireWire, Futurebus , Ultra-2 SCSI, Serial ATA, RapidIO, and SpaceWire, amongst many others……

IB signals/pin



Pin	Connector AI 0-7	Connector AI 8-15
S1	AI 7 +	AI 15 +
S2	AI 7 –	AI 15 –
S3	AI 6 +	AI 14 +
S4	AI 6 –	AI 14 –
S5	AI 5 +	AI 13 +
S6	AI 5 –	AI 13 –
S7	AI 4 +	AI 12 +
S8	AI 4 –	AI 12 –
S9	AI 3 +	AI 11 +
S10	AI 3 –	AI 11 –
S11	AI 2 +	AI 10 +
S12	AI 2 –	AI 10 –
S13	AI 1 +	AI 9 +
S14	AI 1 –	AI 9 –
S15	AI 0 +	AI 8 +
S16	AI 0 –	AI 8 –
GND 1–9	*Ground*	*Ground*
Shield	*Ground*	*Ground*

Source:
National Instruments

Drivers of Modern HPC Cluster Architectures

Multi-/Many-core Processors

High Performance Interconnects – InfiniBand (DPU), Slingshot
<1usec latency, 200-400Gbps Bandwidth>

Accelerators
high compute density, high performance/watt
>9.7 TFlop DP on a chip

SSD, NVMe-SSD, NVRAM

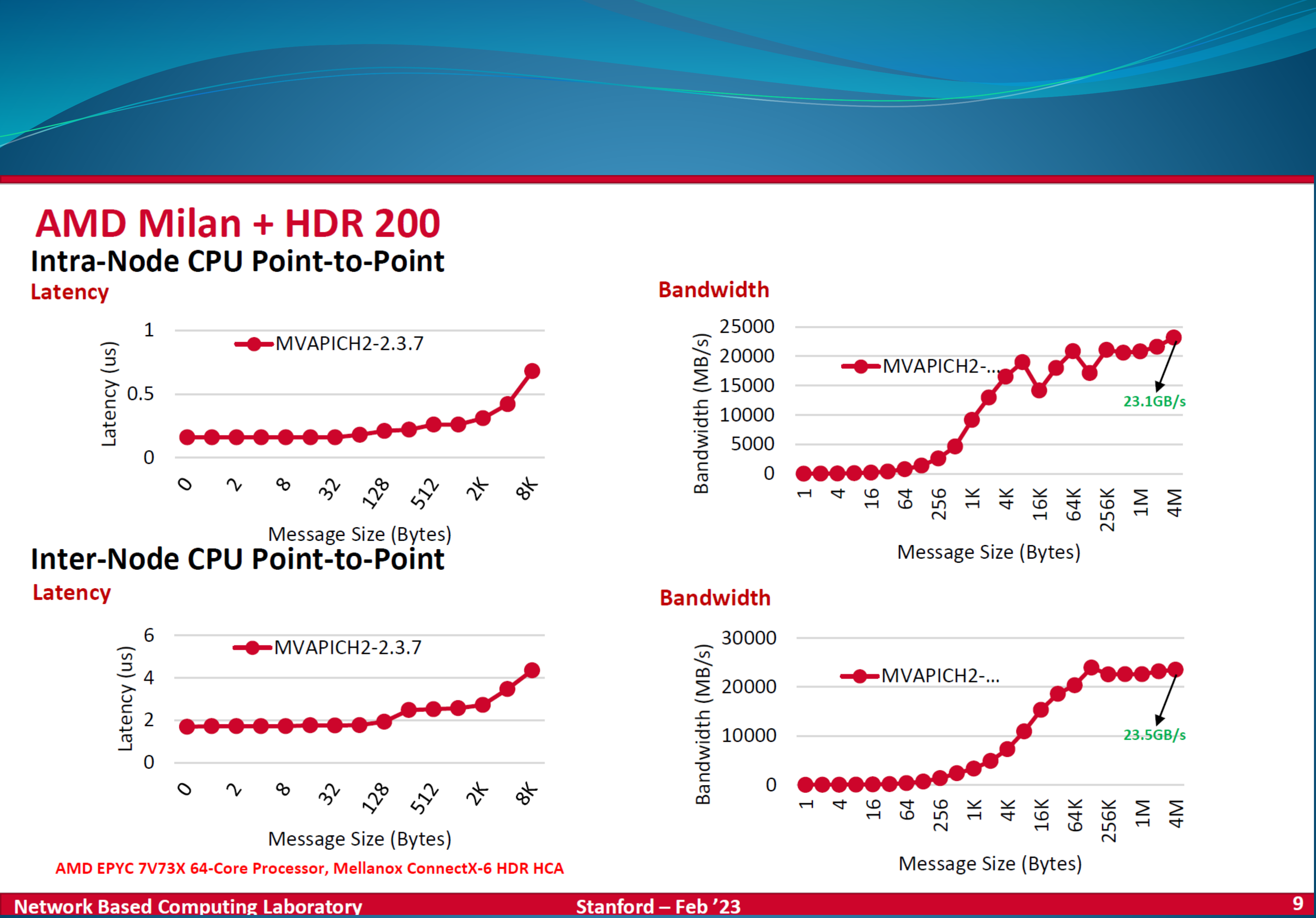
- Multi-core/many-core technologies
- Remote Direct Memory Access (RDMA)-enabled networking (InfiniBand, RoCE, Slingshot)
- Solid State Drives (SSDs), Non-Volatile Random-Access Memory (NVRAM), NVMe-SSD
- Accelerators (NVIDIA GPGPUs)
- Available on HPC Clouds, e.g., Amazon EC2, NSF Chameleon, Microsoft Azure, etc.

Frontier

Fugaku

Summit

Lumi



AMD Milan + HDR 200

Intra-Node CPU Point-to-Point

Latency

Bandwidth

Inter-Node CPU Point-to-Point

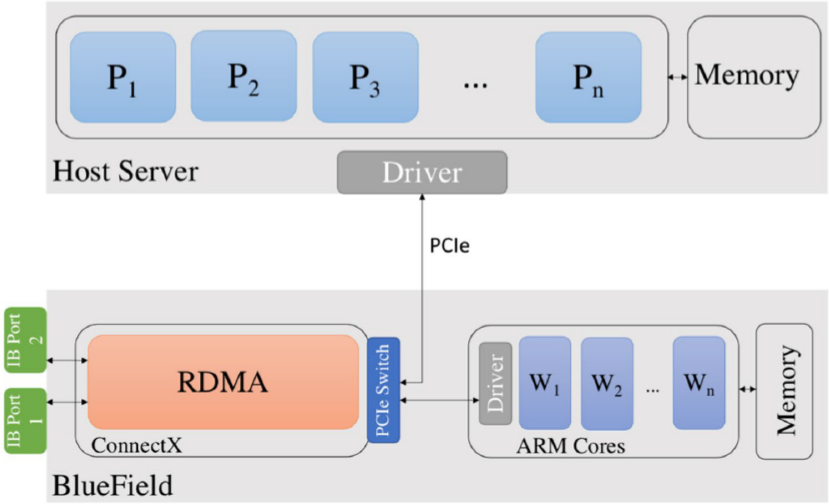
Latency

Bandwidth

AMD EPYC 7V73X 64-Core Processor, Mellanox ConnectX-6 HDR HCA

Accelerating Applications with BlueField-2 Datacenter Processing Unit (DPU)

- ConnectX-6 network adapter with 200Gbps InfiniBand
- System-on-chip containing eight 64-bit ARMv8 A72 cores with 2.75 GHz each
- 16 GB of memory for the ARM cores



P_1
P_2
P_3
...
P_n
Memory
Host Server
Driver
PCIe
IB Port 2
IB Port 1
RDMA
ConnectX
PCIe Switch
Driver
W_1
W_2
...
W_n
ARM Cores
Memory
BlueField