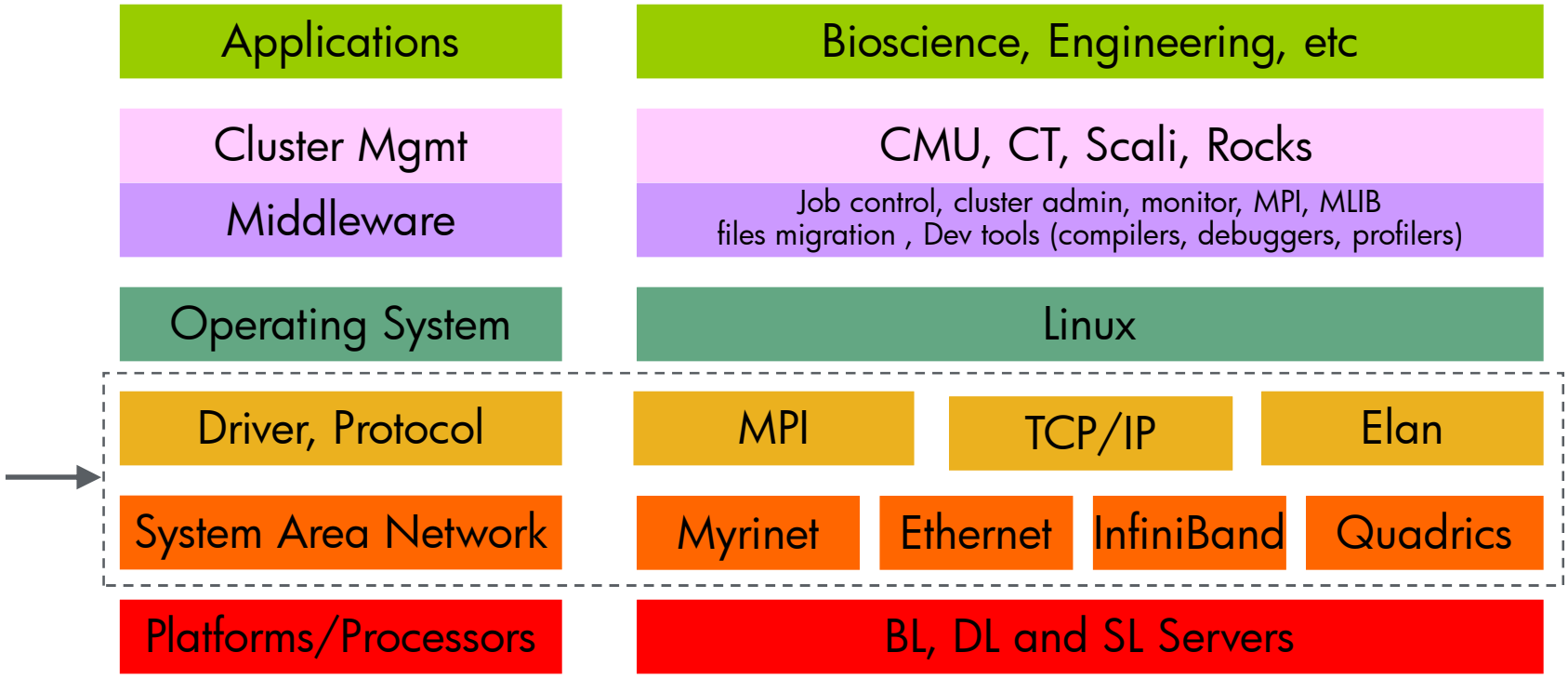
High Performance Computing clusters

Scale up,

Scale out,

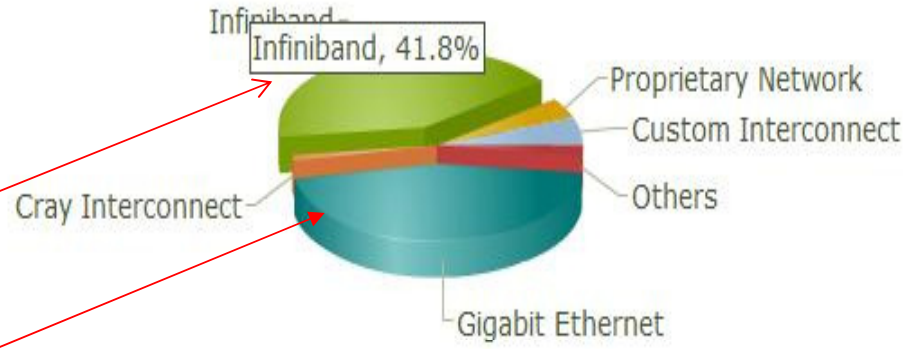
Scale simply!

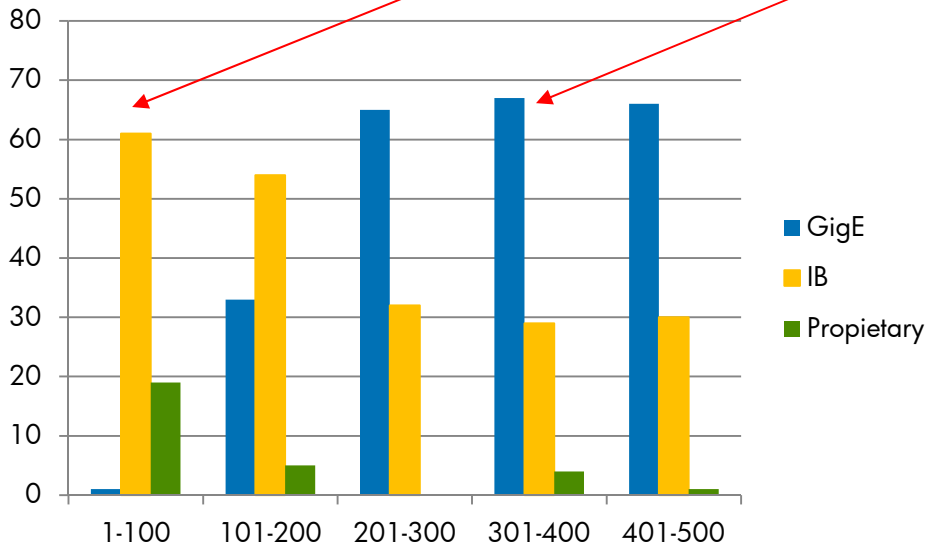
Linux HPC Clusters : Interconnects



Top500 list breakdown by interconnects

InfiniBand is dominant as a *high speed* interconnect: *low latency, high bandwidth, and scalability*

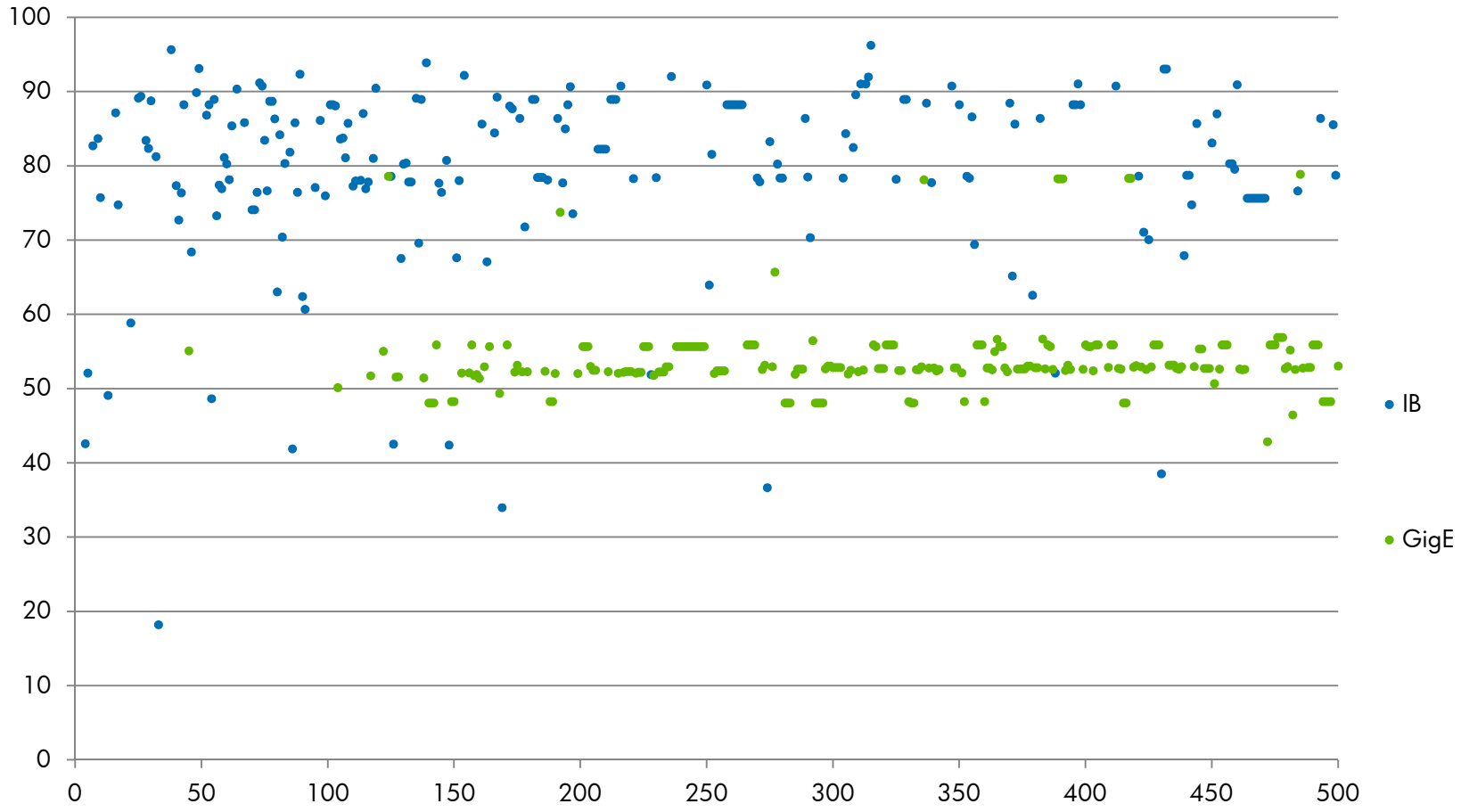




Ethernet is dominant as a *general purpose* interconnect: *connectivity, pervasive technology, and cost*

Source: TOP500 list, November 2011

InfiniBand increases system performance



Source: TOP500 list, November 2011

What is InfiniBand (IB)?

- Is a an industry standard, channel-based architecture that features high- speed, low latency interconnects for cluster computing infrastructure

www.infinibandta.org

IBA Players - 2011

Servers

Switches and Infrastructure

Storage

Embedded, Communications, Military, Industrial

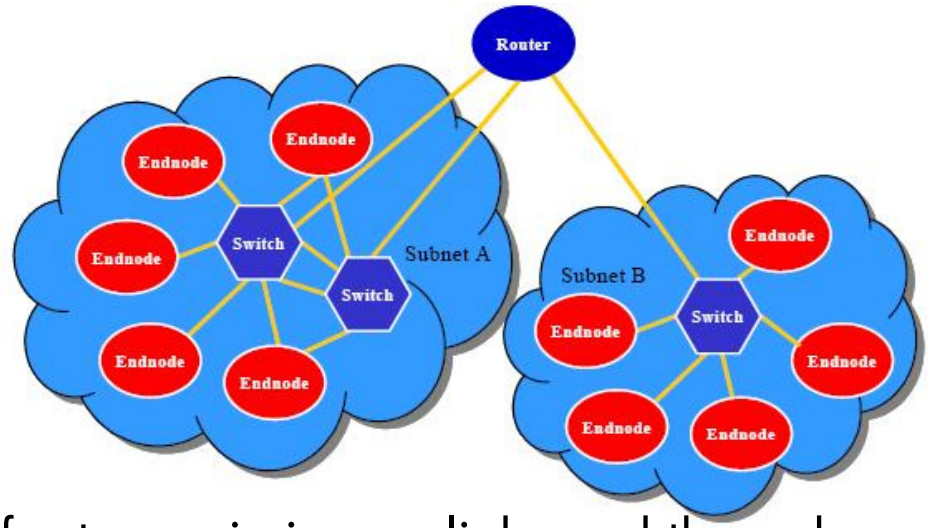
Components of an IB architecture

- Switches
- Routers
- Servers
- HCA – Host channel Adapters
- Subnet Managers

- + cables, connectors, QSFP

Switches and Routers

- IB switches route messages from their source to their destination based on routing tables.
- The exact format, content, and organization of these tables in the switch hardware is vendor-specific.



Messages are divided into packets for transmission on links and through switches.
The Maximum Transfer Unit: 256B, 1, 2, 4KB
Switch size, number of ports, links supported, multicast support are vendor-specific.

IB Routers → forward packets from one subnet to another without consuming or generating packets.

End Nodes

- Generally they are the servers that access the IB
- Theory: are the host systems (servers) or devices (network adapters, storage subsystems) that access the IB.
- Any two end-nodes can communicate to each other on the IB fabric using the IB specifications.

HP c7000 enclosure with blades

HP BL460c Server Blade

HP 4x DDR IB HCA mezzanine card

HP 4x DDR IB Interconnect Switch

Channel Adapters

Host Channel Adapter

- An HCA can be a card installed in an expansion slot or integrated onto the host's system board.
- An HCA can communicate directly with another HCA, with a target channel adapter, or with an InfiniBand switch.
- Has a collection of features that are defined to be available to host programs. (called verbs)

Target Channel Adapter

- A TCA is used to connect an external device (storage unit or I/O interface) to an InfiniBand infrastructure

Subnet Management

An InfiniBand subnet requires a Subnet Manager :

Role of the Subnet Manager :

- discovery of all the IB links
- Link management (activating, routing tables ..)
- Configuring the ports
- Monitoring and Reporting Performance
- …

No Subnet Manager => Infiniband Subnet is Down

Manager and Agents

- IBA Management is composed from managers and agents.
- Managers are active entities
- Agents are passive entities that respond to messages from managers.
- Every Subnet contains only one MASTER SUBNET MANAGER (on one node or switch)
- There is exactly one Master SM, and other SMs are Stand By SMs(or in Not Active State).

Where does the SM run ?

Subnet Management software can be located on :

- An Infiniband switch with an embedded Subnet Management Board (SMB)

  Such a switch is called 'Internally Managed Switch'

- A linux server running the Subnet Management software

  (available with OFED, the OpenFabrics Entreprise Distribution)

➔ Have at least two subnet managers for availability

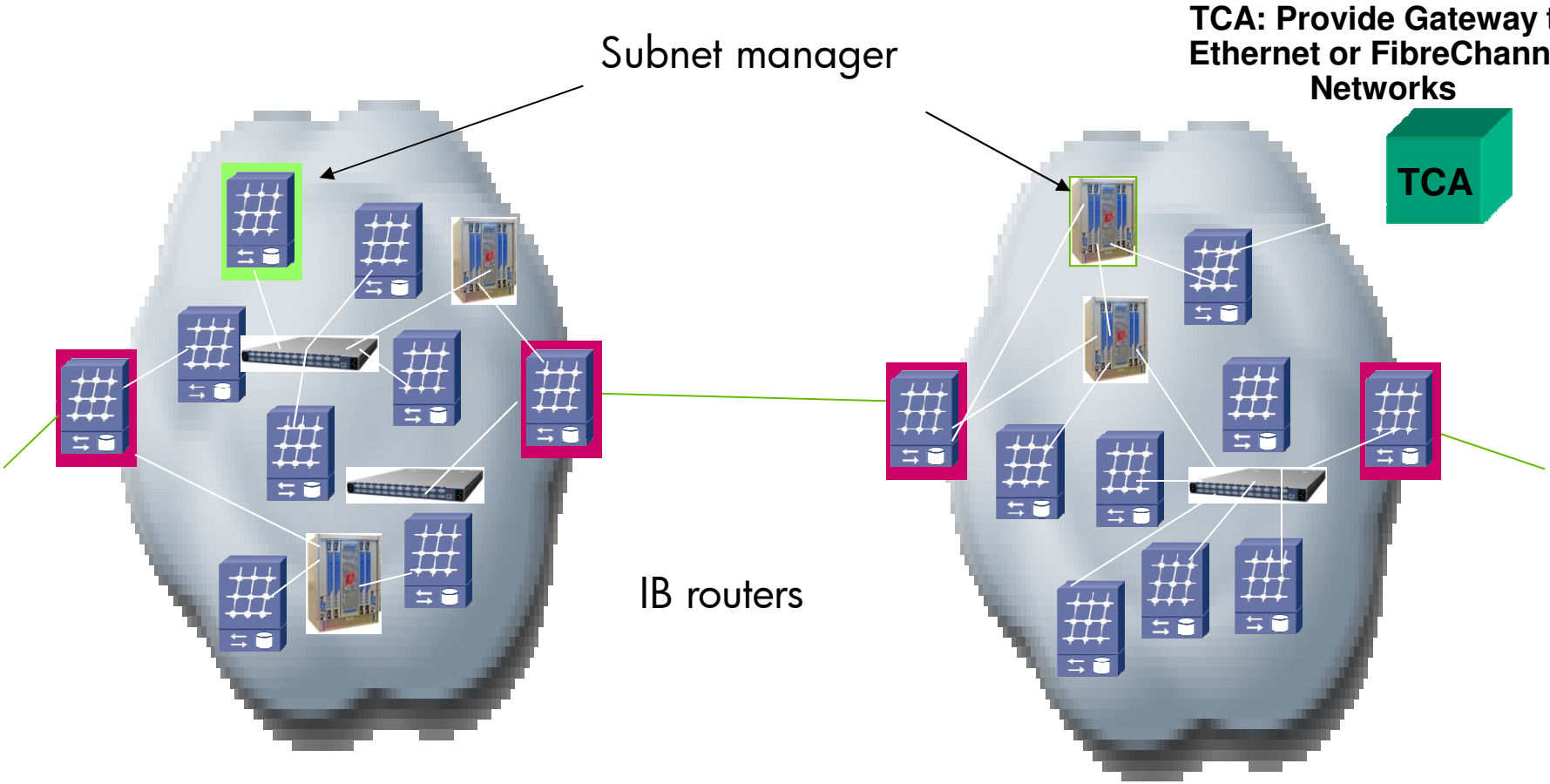
Switches with an embedded SMB are called 'Internally Managed Switches'

Switches without an embedded SMB are called 'Externally Managed Switches'

Master SM Functions

- It discovers and initializes the network
- Assigns local IDs (LIDs) to all elements
- Determines path MTUs
- Loads the switch routing tables that determines the paths from endnode to endnode
- Asks for information each Subnet Mangement Agent from each host.
- Scans regulary the subnet to detect additions (hot plugs) and deletions (hot unplug)

Infiniband Architecture



Each subnet must have a subnet manager

InfiniBand Addressing

Each component must be addressed, similar to Ethernet MAC addresses

GID at the Fabric level :

- Subnet Prefix — 64 bits
- GUID — 64 bits - Global Unique Id
- GID = Subnet Prefix + GUID

LID at the Subnet level :

- LID — 16 bits - Local ID

Signal Links 8/10 → 64/66

- Signal Rate is 2.5 Gbit/s or
- Infiniband supports Single, Double and Quad Data Rate

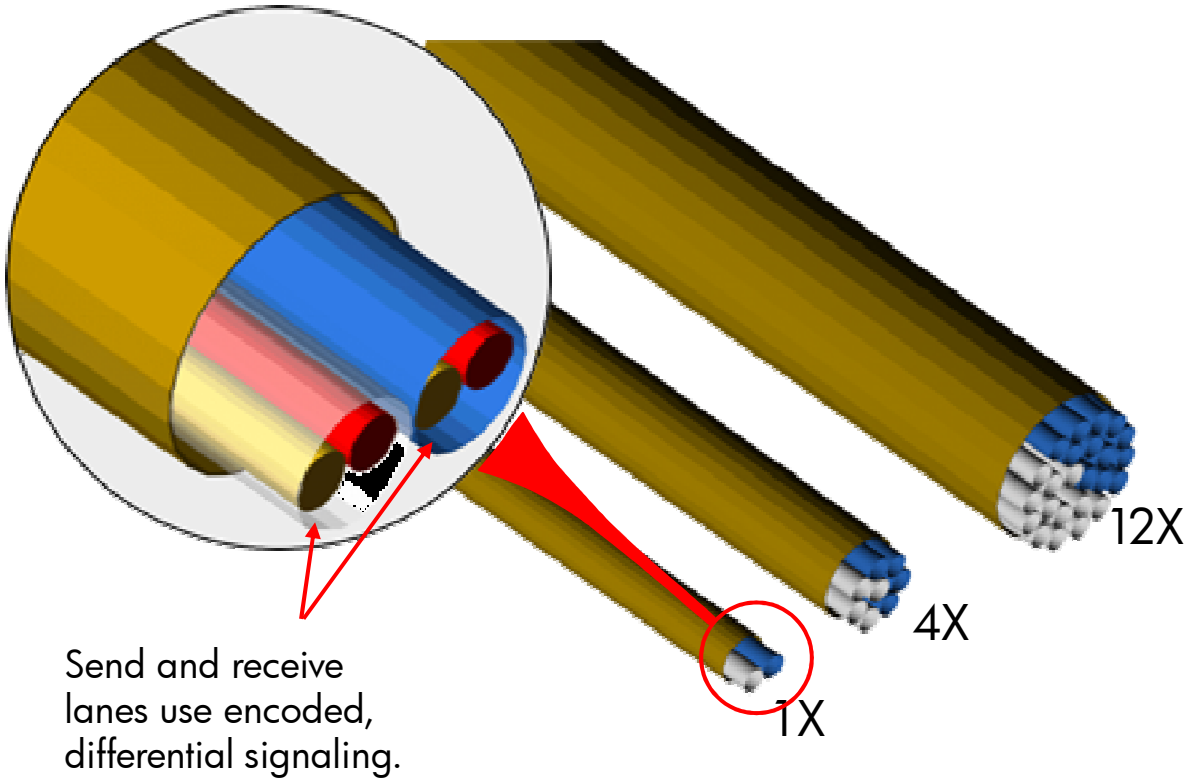
	Signal Rate	8B/10B Encoding*
Single Data Rate	2.5 Gbit/s	2 Gbit/s
Double Data Rate	5 Gbit/s	4 Gbit/s
Quad Data Rate	10 Gbit/s	8 Gbit/s

- Infiniband will support FDR(released this year) and EDR (2013)

	Signal Rate	64B/66B Encoding*
Fourteen Data Rate	14.0625 Gbit/s	14 Gbit/s
EnhancedData Rate	25.78125 Gbit/s	25 Gbit/s

Link Width

- Link width is equal with the number of channel pairs used by an IB link
- The link width can have 1, 4 or 12 Channels Pairs
- Each pair is defined by a send channel and a receive one



Link Speed

- Link Speed = Link Width * Signal Rate
- The following bandwidth can be reached

	Single (SDR)	Double (DDR)	Quad (QDR)	Fourteen (FDR)	Enhanced (EDR)
1X	2 Gbit/s	4 Gbit/s	8 Gbit/s	14 Gbit/s	25 Gbit/s
4X	8 Gbit/s	16 Gbit/s	32 Gbit/s	56 Gbit/s	100 Gbit/s
12X	24 Gbit/s	48 Gbit/s	96 Gbit/s	168 Gbit/s	300 Gbit/s

- The CURRENT OLD Technology us 4X QDR
- The CURRENT NEW Technology is 4xFDR.

QUESTION: WHY DID WE STOP AT 4X QDR.
Why not 12 X QDR?

ANSWER : THE PCI ☺

PCI	Raw bit Rate	Link Bw	BW/lane/ way	x8 (IB boards)	x16 (GPU boards)
PCIe 1.x	2.5GT/s	2Gb/s	~250MB/s	2GB/s	4GB/s
PCIe 2.x	5.0GT/s	4Gb/s	~500MB/s	4GB/s	8GB/s
PCIe 3.0	8.0 GT/s	8 Gb/s	~1GB/s	8GB/s	16GB/s

IB	Single (SDR)	Double (DDR)	Quad (QDR)	Fourteen (FDR)	Enhanced (EDR)
1X	2 Gbit/s	4 Gbit/s	8 Gbit/s	14 Gbit/s	25 Gbit/s
4X	8 Gbit/s	16 Gbit/s	32 Gbit/s	56 Gbit/s	100 Gbit/s
12X	24 Gbit/s	48 Gbit/s	96 Gbit/s	168 Gbit/s	300 Gbit/s

Point-to-point QDR bandwidths

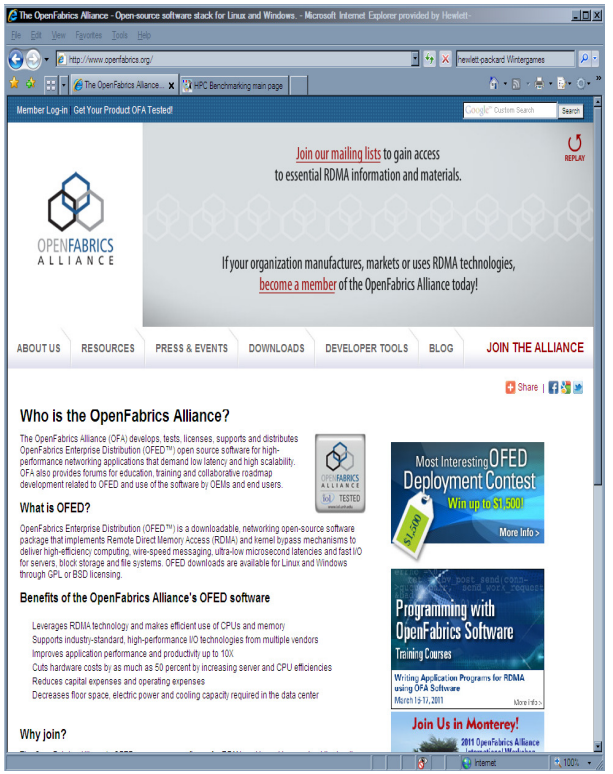
		Mellanox HCA	Mellanox SW	Voltaire SW	QLogic HCA	QLogic SW
		HCA	SW	SW	HCA	SW
Mellanox	HCA	3.2GB/s	3.2GB/s	3.2GB/s	?	3.2GB/s
Mellanox	SW	3.2GB/s	4.0GB/s	4.0GB/s	X	X
Voltaire	SW	3.2GB/s	4.0GB/s	4.0GB/s	3.2GB/s	1.0-4.0GB/s
QLogic	HCA	?	X	3.2GB/s	3.2GB/s	3.2GB/s
QLogic	SW	3.2GB/s	X	1.0-4.0GB/s	3.2GB/s	4.0GB/s

FDR – Roadmap & Ecosystem

- Signal rate enables preservation of most of the ecosystem & very fast time to market
- Preserved from QDR
  - Chassis & Midplane
  - Cables (max distance might be slightly reduced)
  - Physical connectors (QSFP+)
- New
  - HCA & Servers (PCI Gen 3)
  - Line & Fabric cards
- Availability → Q3 2011

Hardware is nothing w/o software

- OpenFabrics Alliance: http://www.infinibandta.org/



- Naboo planet found here, among many other Promoters:

About OpenFabric.org, OFED

- **OpenFabrics.org** is a 'not for profit 'alliance committed to develop Open Source transport independent stacks using RDMA technology.
- Target market : HPC and storage.
- Focus on Infiniband and iWARP, Linux and Windows

- OFED (OpenFabrics Enterprise Distribution) is the name of the software packages, released by the OpenFabric Alliance
- OFED is the strategic direction for Infiniband
  1. Adopted by all Infiniband hardware vendors
  2. Included in Linux kernel

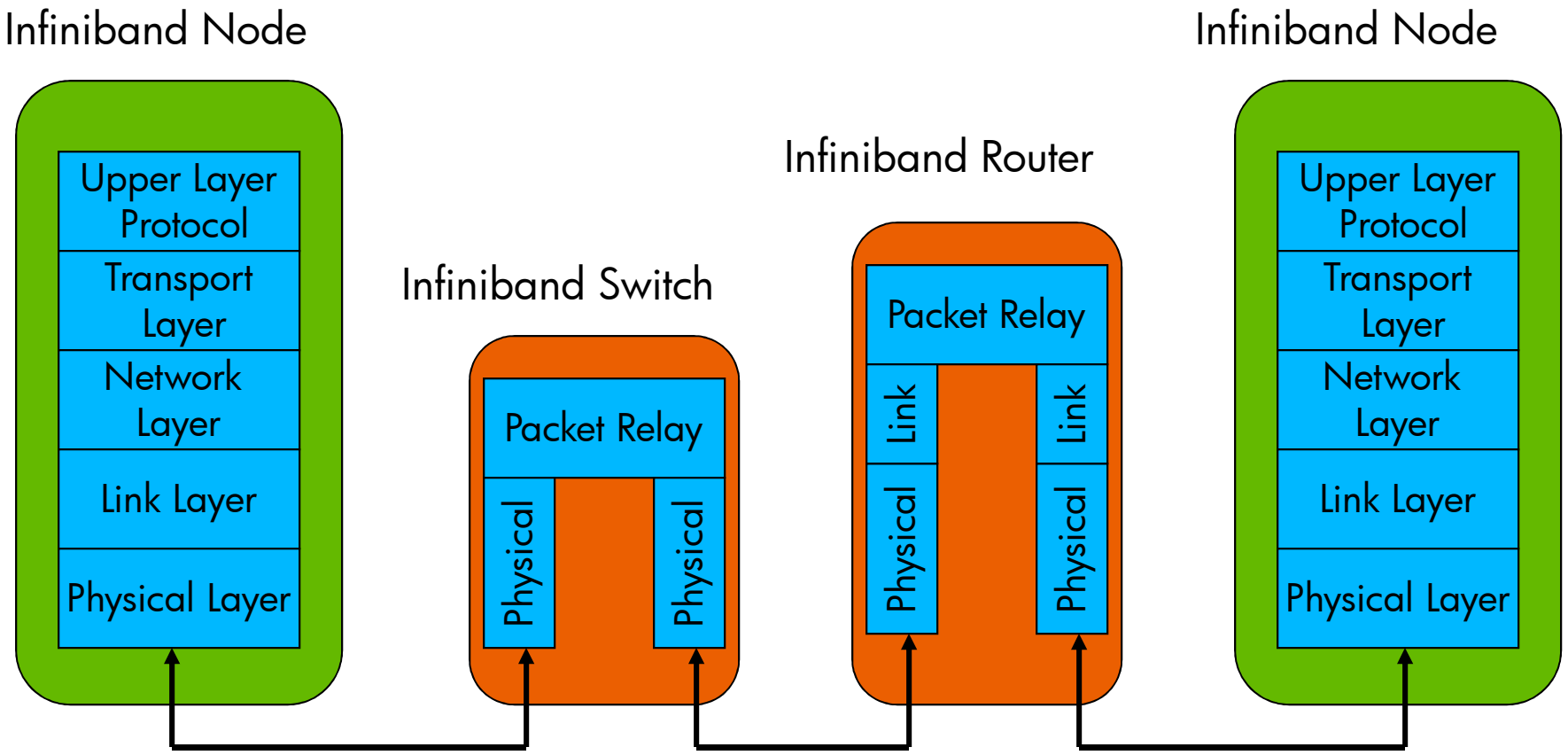
IB Software

- InfiniBand, like Ethernet, uses a multi-layer processing stack to transfer data between nodes.
- IBA provides OS-bypass features
  - communication processing duties
  - RDMA operations as core capabilities
  - offers greater adaptability through a variety of services and protocols.
- Drivers and HCA stacks are available for Linux, Microsoft Windows, HP-UX, Solaris

IB Layers

Application Layer
Transport Layer
Internet Layer
Link Layer

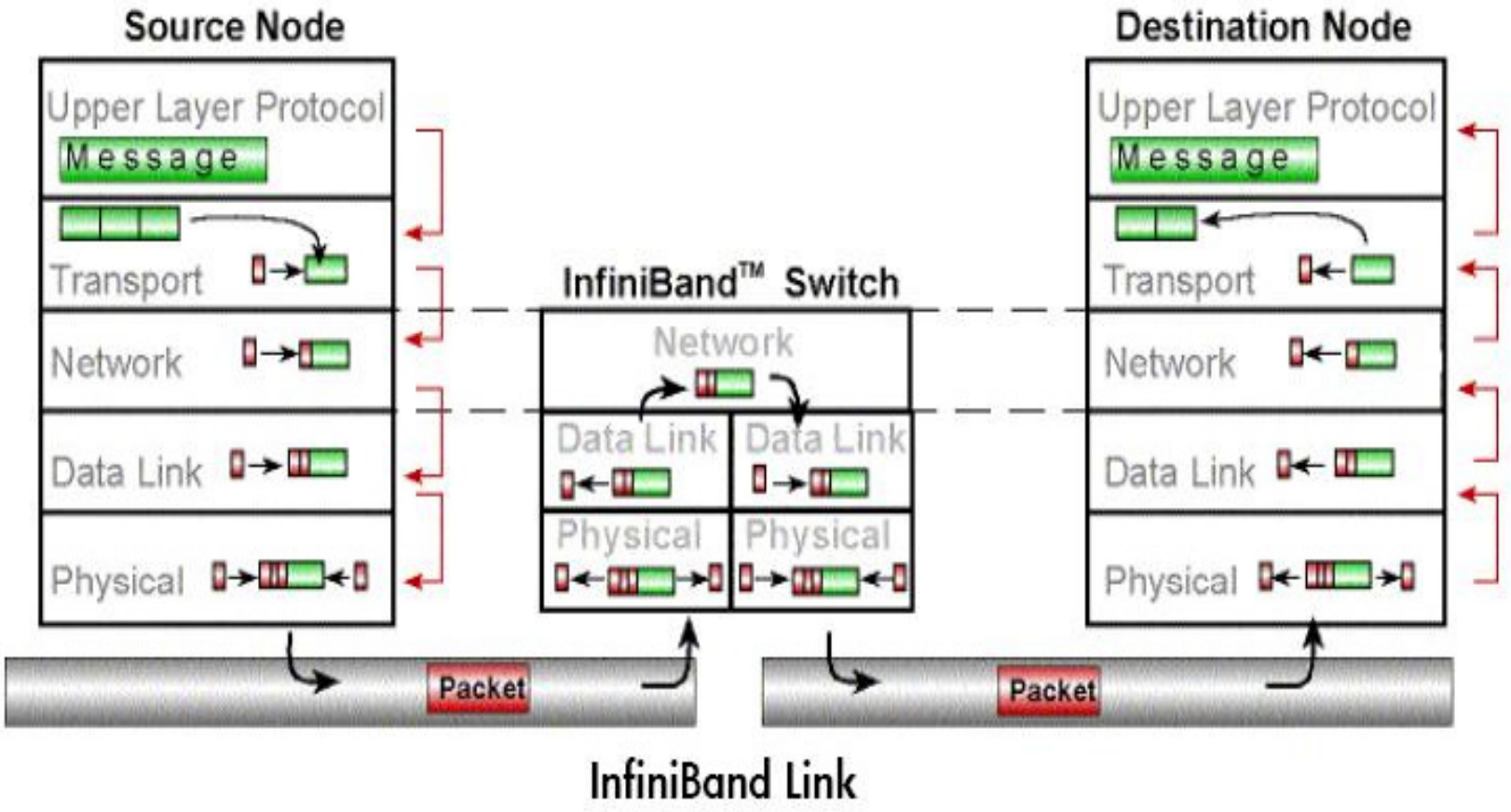
Application Layer
Presentation Layer
Session Layer
Transport Layer
Network Layer
Data Link Layer
Physical Layer



IB Layers

- ULP (Upper Layer Protocol): works close to the OS and application. It defines how much software overhead will be required by data transfer
- Transport Layer: is responsible for the communication between the applications. It splits the messages into data payloads and encapsulates each data payload and an identifier of the destination node into one or more packets.
- Network Layer: selects a route to the destination node and attaches the route information to the packets.
- Data Link Layer: attaches a local identifier (LID) to the packet for communication at the subnet level.
-
- Physical Layer: transforms the packet into an electromagnetic signal based on the type of network media—copper or fibre.

Message Passing over IB Layers

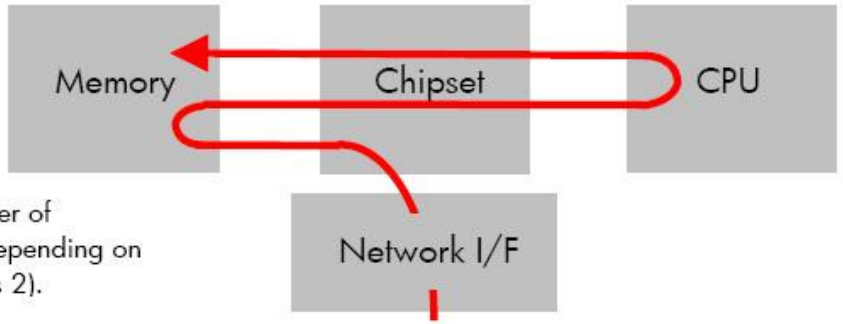


The real FORCE : Remode Direct Memory Access

- RDMA is a data exchange technology that improves network performance by streamlining data processing operations.
- RDMA provides a faster path for applications to transmit messages between network devices
- Can be applied to both Ethernet, TCP and IB supporting SDP, iSER, NFS, SRPand MPI.

RDMA

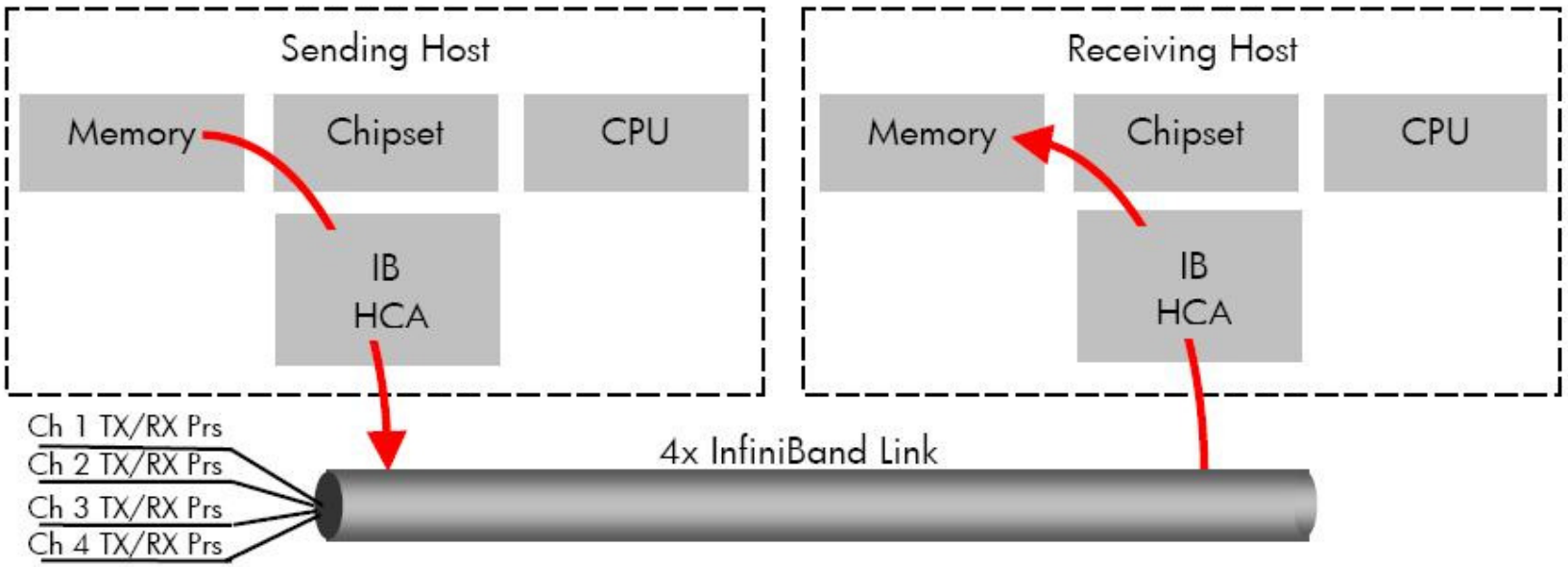
- RDMA was basically developed to move data from memory from one computer into the memory of another with minimum involvement from their processors
- The RDMA protocol allows a system to place transferred data directly into its final memory destination without additional or interim data copies into system buffers and minimal work by the OS kernel
- This “zero copy” or “direct data placement” (DDP) capability provides the most efficient network communication possible between systems.



NOTE: The actual number of memory copies varies depending on OS (Example: Linux uses 2).

RDMA - IB

- The Principle is the same as in Ethernet.
- The difference is that the HCA requires prior loading of both software drivers (vendor specific) and the communication stack that is OS specific



Message Passing Interface

- **MPI** protocol is a library of calls used by applications in a parallel computing environment to communicate between nodes.
- Code is executed across multiple nodes simultaneously
- MPI facilitates the communication and synchronization among these jobs across the entire cluster.

There are several implementations of MPI on the market:

- HP-MPI
- Intel MPI
- Publicly available versions such as MVAPICH2 and Open MPI

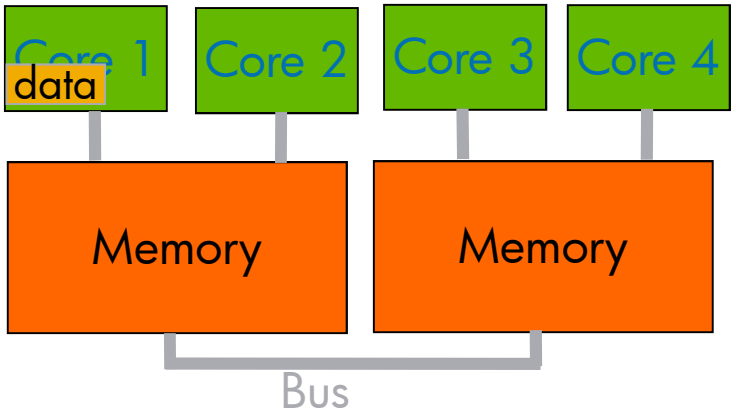
MPI has become the de-facto IB ULP standard

MPI Communication

Movement of data depends on relative location of destination and interconnect. Paths are:

- Communication within a Node (shared memory)
- Communication from Node to Node over TCP/IP
- Communication from Node to Node over high speed interconnects InfiniBand, Quadrics, Myrinet

MPI Communication within a Node



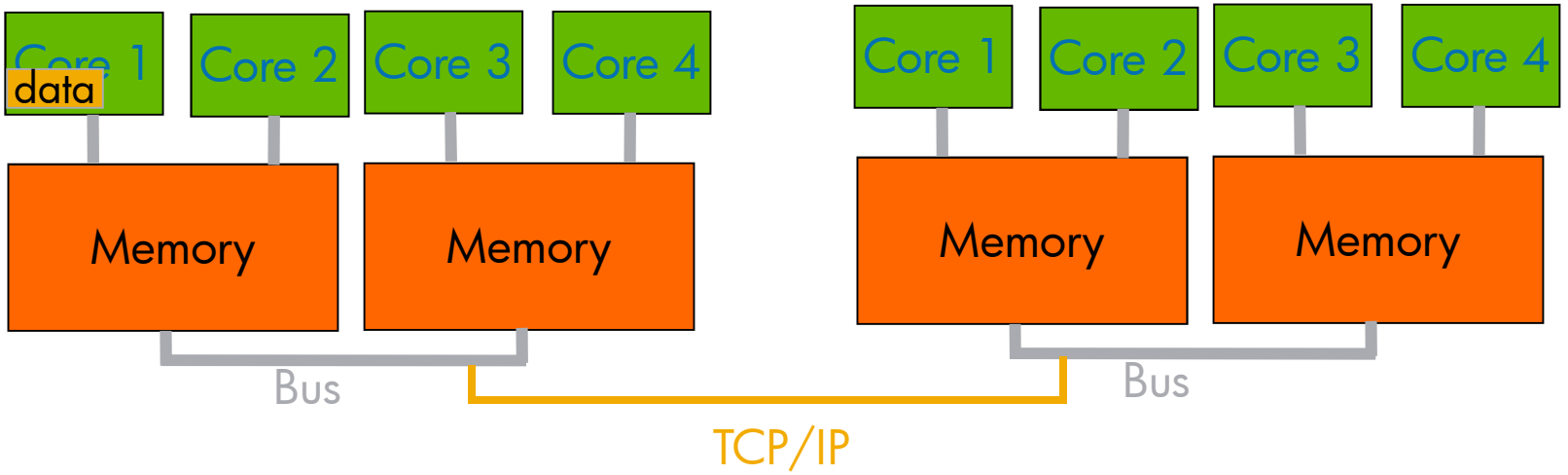
To Send data from Core 1 to Core 4:

- Core 1 -> Core 1 Local Memory
- Core 1 Local Memory* -> System Shared Memory**
- System Shared Memory -> Core 4 Local Memory
- Core 4 Local Memory -> Core 4

*The operating system makes Local Memory available to a single process

**The operating system makes Shared Memory available to multiple processes

MPI Communication to another Node via TCP/IP

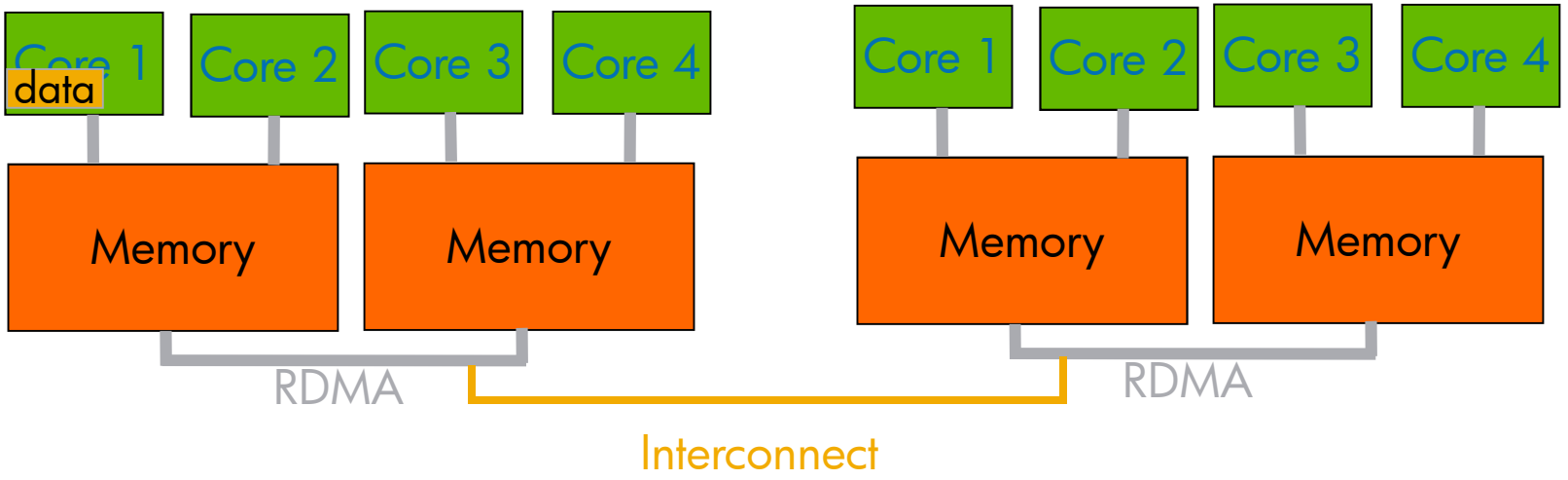


To Send data from Core 1, Node 1 to Core 1, Node 2:

- Core 1, Node 1 -> Core 1, Node 1 Local Memory
- Core 1, Node 1 Local Memory -> Node 1 Shared Memory
- Node 1 Shared Memory -> Interconnect
- Interconnect -> Node 2 Shared Memory
- Node 2 Shared Memory -> Core 1, Node 2 Local Memory
- Core 1, Node 2 Local Memory -> Core 1, Node 2

The core is used to send data to the TCP/IP Interconnect

MPI Communication to another Node via INFINIBAND



To Send data from Core 1, Node 1 to Core 1, Node 2:

- Core 1, Node 1 -> Core 1, Node 1 Local Memory
- Core 1, Node 1 Local Memory -> Node 1 Shared Memory
- Node 1 Shared Memory -> Interconnect
- Interconnect -> Node 2 Shared Memory
- Node 2 Shared Memory -> Core 1, Node 2 Local Memory
- Core 1, Node 2 Local Memory -> Core 1, Node 2

10G vs InfiniBand – when to use what?

- InfiniBand is best for:
  - Customers looking for lowest latency end-to-end
    - MPI-based HPC applications, or other applications that are/can be implemented on Verbs API
  - Customers need more than 10Gbps on the fabric
    - 4X QDR IB provides 40Gbps (32Gbps data) bandwidth
- 10GE Ethernet is best for:
  - Customers deploying scale-out computing for enterprise and virtualization applications
  - Need more performance than 1G Ethernet, but do not want to add InfiniBand into their environment

THANK YOU

Non blocking fabrics

Fully non blocking fabrics

Fully nonblocking fabrics

- Fabric can be designed to be fully nonblocking

⇒ **Full bandwidth** from every server to every other server anywhere in the fabric

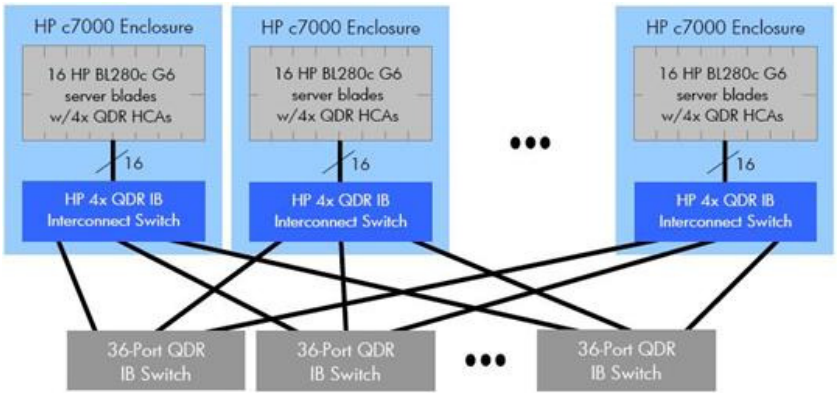
⇒ By design, all single switches are fully nonblocking

or you can design the Fabric with a block factor.

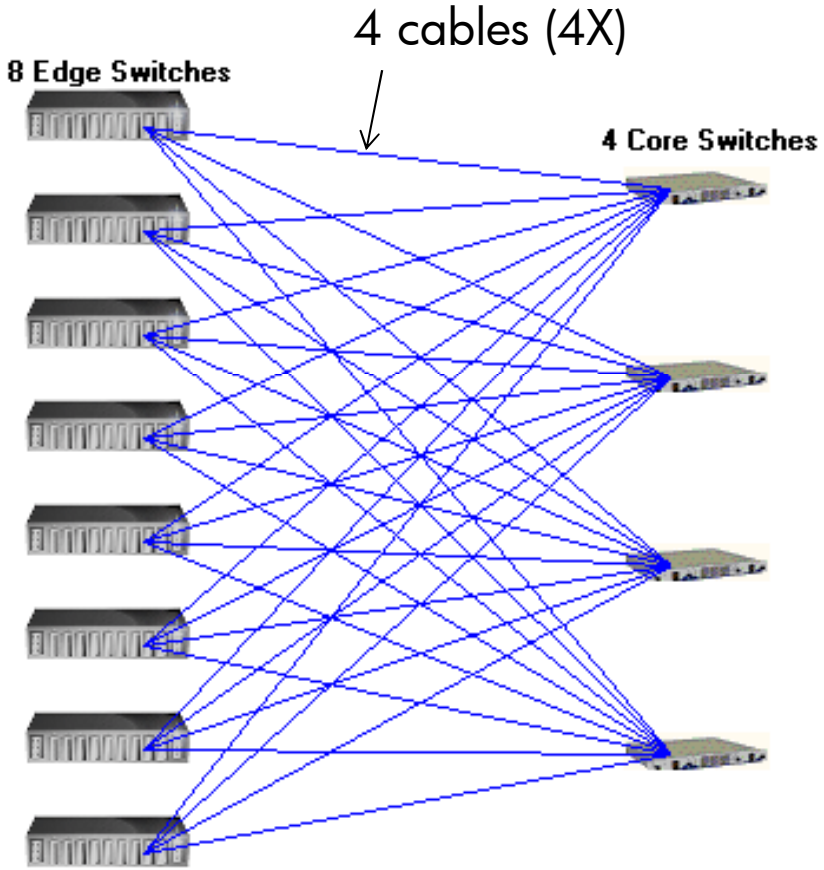
8 Blade enclosures (C7000)
a 128-node Blade cluster using 36-port switches
full Non-blocking bandwidth layout

16 blades BL460c per C7000, 1 HBA per Bl460c
1 blade infiniband switch 36 ports per C7000
4 core switches

16 uplinks from each blade IB switch
each blue line is 4 4X links



total bandwidth : 2560 Gb/S



C7000

Non blocking fabrics

Fully non blocking fabrics

Blocking Factor

- Fabric can be designed to be partially blocking. There are less switches and it introduces less bandwidth

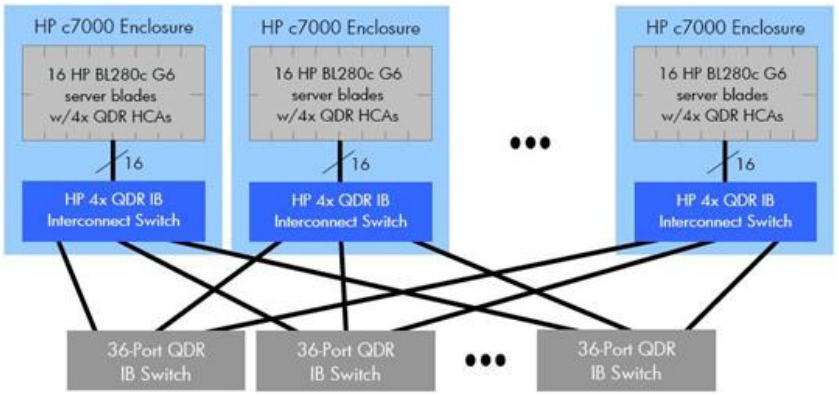
Typically reduced bandwidth comes from the Leaf switch by using more ports to connect servers than to connect to the Root switch

- ➔ No impact on latency
- ➔ Blocking will be visible when there is contention on the links between the Leaf switches and Root switches
- ➔ Servers may have full bandwidth if there is no contention

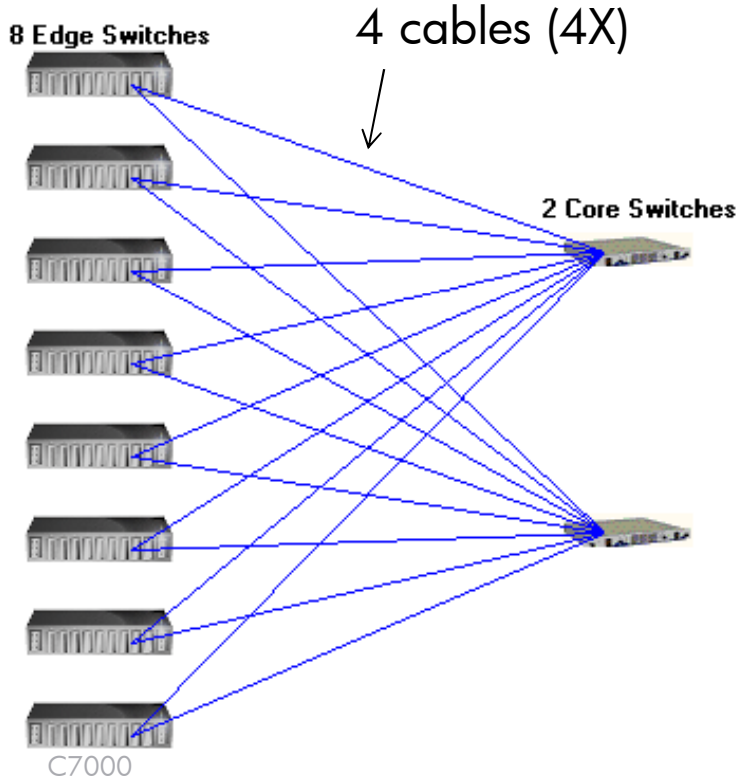
8 Blade enclosures (C7000)
a 128-node Blade cluster using 36-port switches with half bandwidth (2:1 oversubscription)

16 blades BL460c per C7000, 1 HBA per Bl460c
1 blade infiniband switch 36 ports per C7000
2 core switches

8 uplinks from each blade IB switch
each blue line is 4 4X links

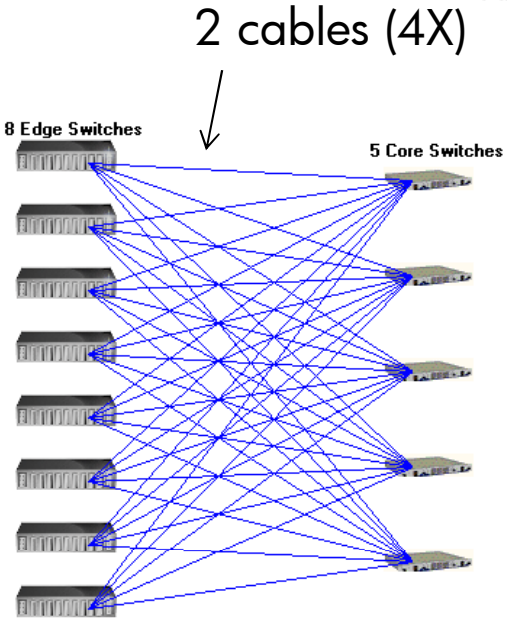


total bandwidth : 1280 Gb/S

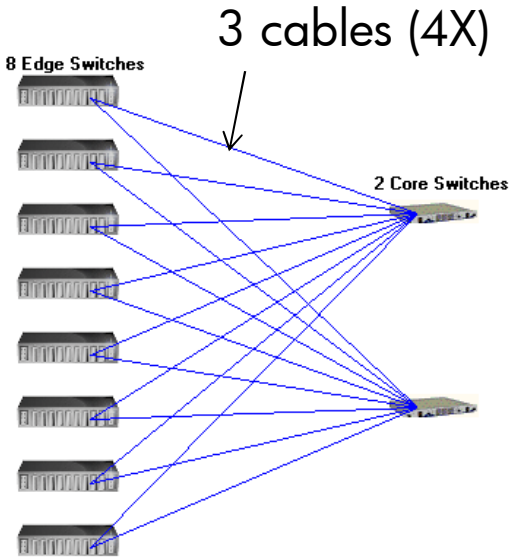


Lot of variants are possible

~60% blocking factor
- 16 blades BL460c per C7000
- 1 blade infiniband switch 36 ports per C7000
- 5 core switches
- each blue line is 2 4X links
- total bandwidth : 1600 Gb/S



~33% blocking factor
- 16 blades BL460c per C7000
- 1 blade infiniband switch 36 ports per C7000
- 2 core switches
- each blue line is 3 4X links
- total bandwidth : 960 Gb/S



which one ?

At HP, we support 2:1 or 1:1 bandwidth configurations in blades or rack-mount server clusters, other on request.

For many applications, 2:1 oversubscription is the cost effective hardware configuration

- 1:1 is preferred for HPC datacenters who run large number of various HPC applications
- 2:1 is preferred for HPC datacenters who focus on a few workloads where latency is a primary factor for performance

Other factors affect the performance

- the application itself
- MPI implementation
- IB Routing algorithm,
- Batch scheduler
- User experience, etc, …