University of BRISTOL

Lecture 8: InfiniBand Architecture for HPC

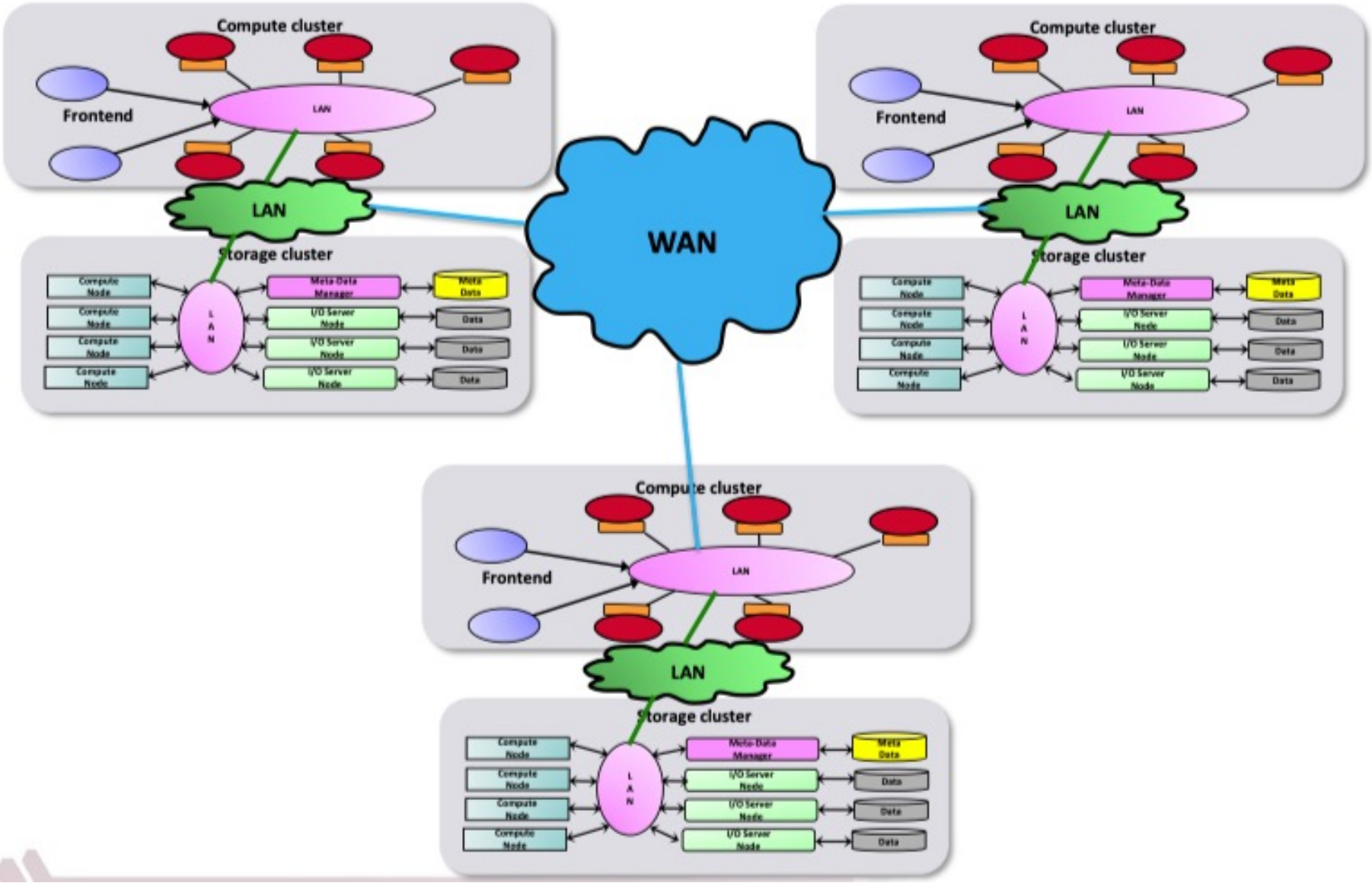
Part 1

Application and computing systems

University of BRISTOL

- Diverse range of applications
  - Processing and dataset characteristics vary
- Growth of high-performance computing
  - Growth in processor performance
    - Multi-core CPUs, chip density doubles every 18 months
  - Growth in commodity networking:
    - increase in speed/features + reducing cost
- Different kinds of systems
  - Clusters, grid, cloud, data centers

Grid Computing Environment



Networking and I/O Requirements

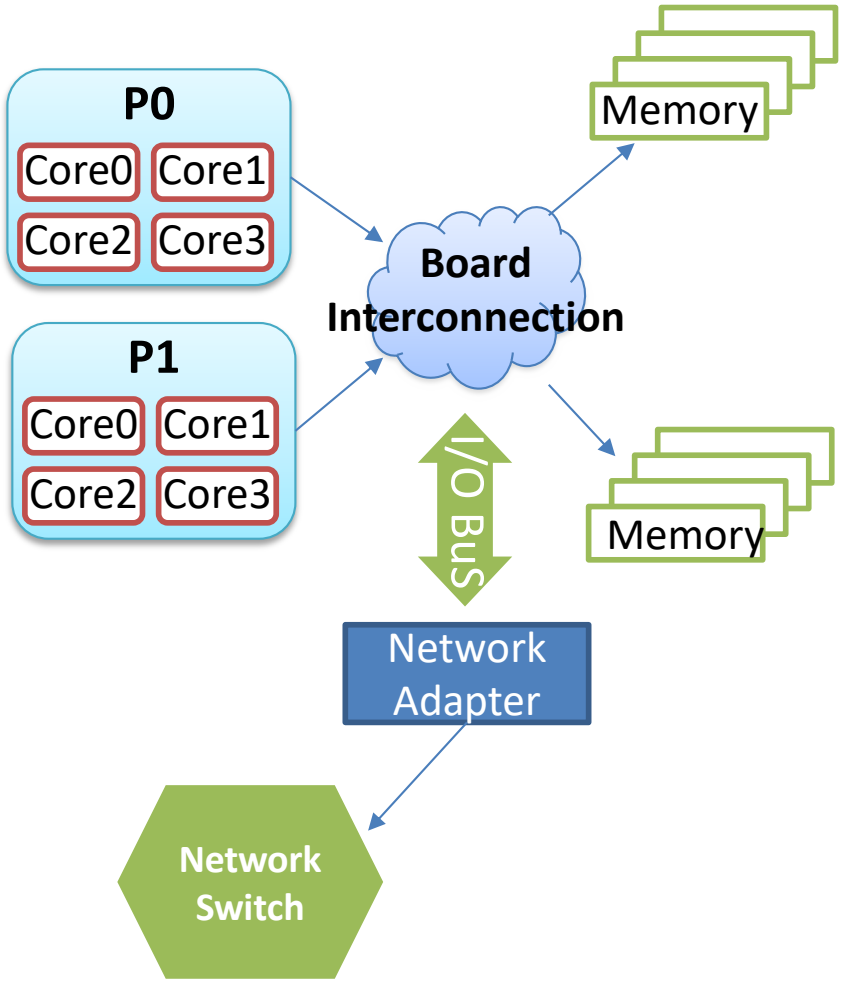
University of BRISTOL

- System Area Networks
  - Excellent performance (low latency, high bandwidth and low CPU utilization) for inter-processor communication (IPC)
  - High performance I/O
- WAN connectivity in addition to intra-cluster SAN/LAN connectivity
- Varies Quality of Service (QoS) for interactive applications
- RAS (Reliability, Availability, and Serviceability)

Key elements of the computing architecture

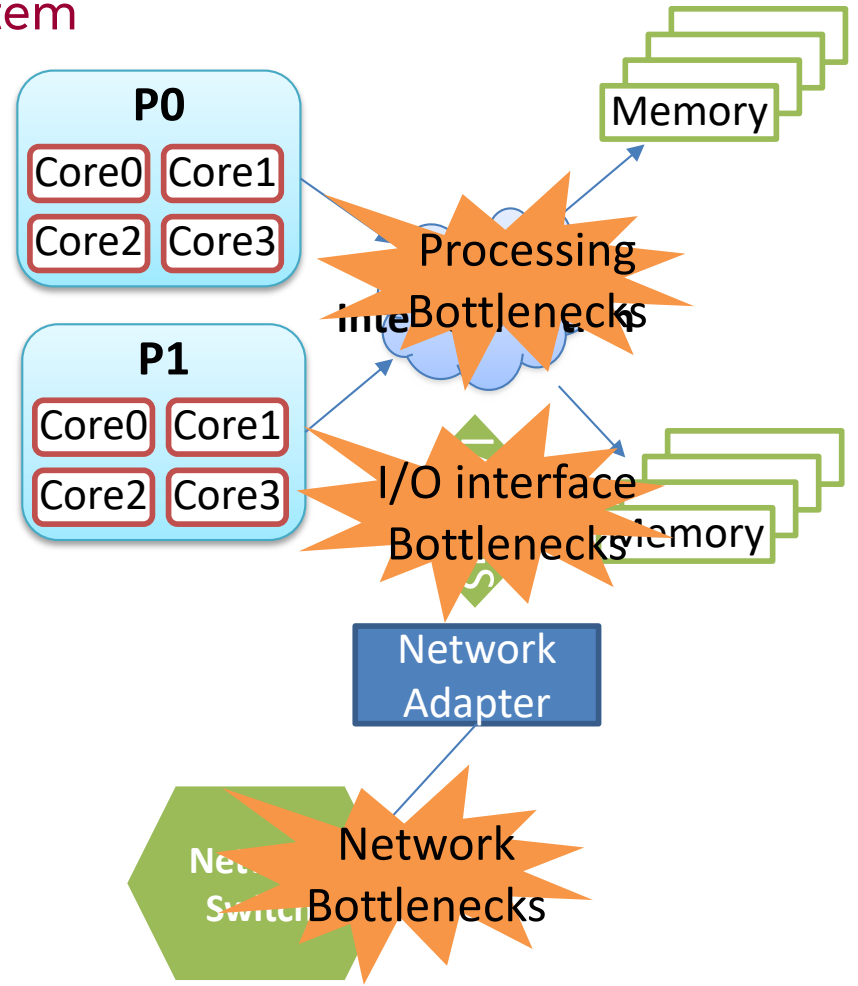
University of BRISTOL

- System Balance in Computing Platforms!
  - Central Processing Unit (CPU) throughput
  - Memory sub-system performance
  - Input/Output (I/O) sub-system performance
- Operations contribute to the overall compute time
  - Get the data from the other node (I/O)
  - Fetch the data from memory
  - Process the data and generate a result



Major components in computing systems

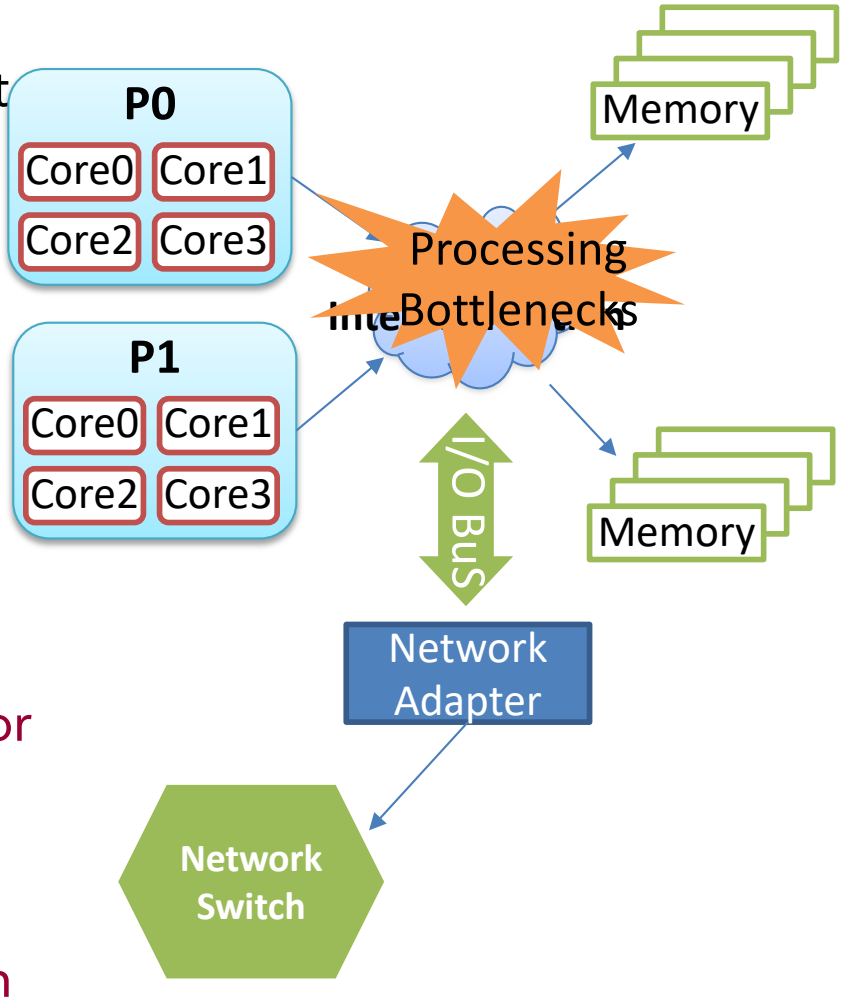
- Hardware components
  - Processing cores and memory subsystem
  - I/O bus or links
  - Network adapters/switches
- Software components
  - Communication stack



Processing bottlenecks in Traditional Protocols (e.g., TCP/IP, UDP/IP)

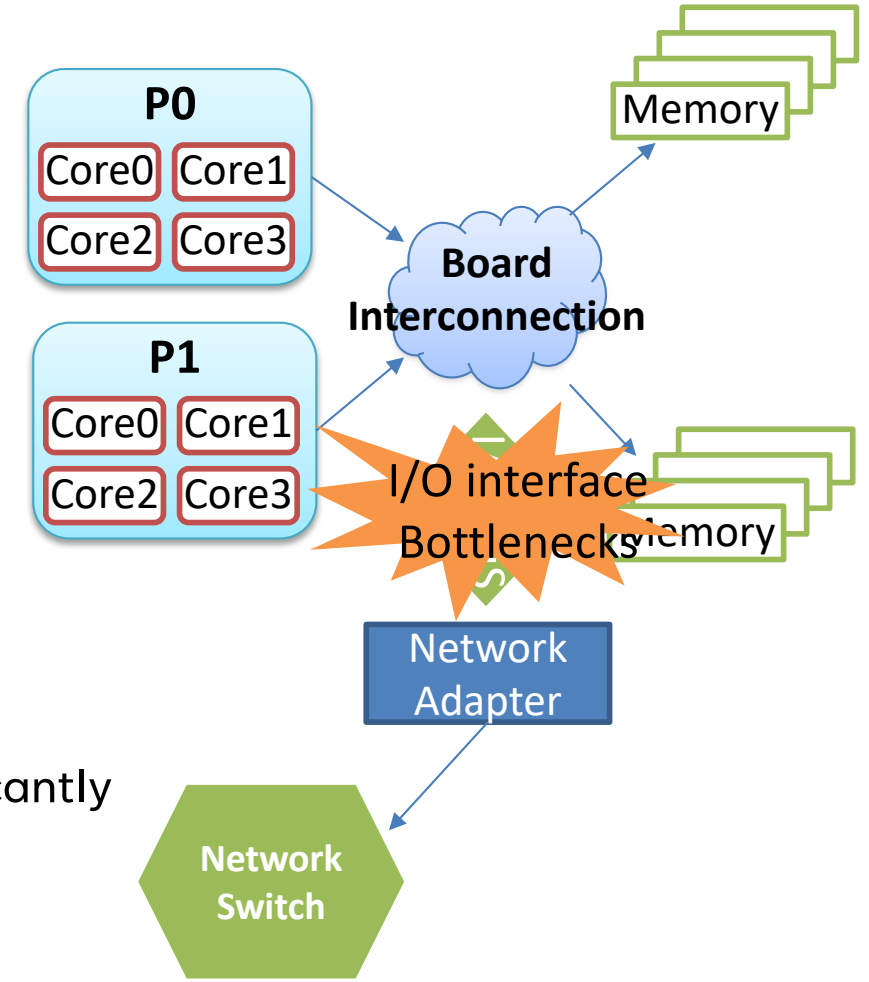
University of BRISTOL

- Generic architecture for all networks
- Host processor handles almost all aspect of communications
  - Data buffering (copies on sender and receiver)
  - Data integrity (checksum)
  - Routing aspects (IP routing)
- Signalling between different layers
  - Hardware interrupt on packet arrival or transmission
  - Software signals between different layers to handle protocol processing in different priority levels



Bottlenecks in Traditional I/O interfaces and networks

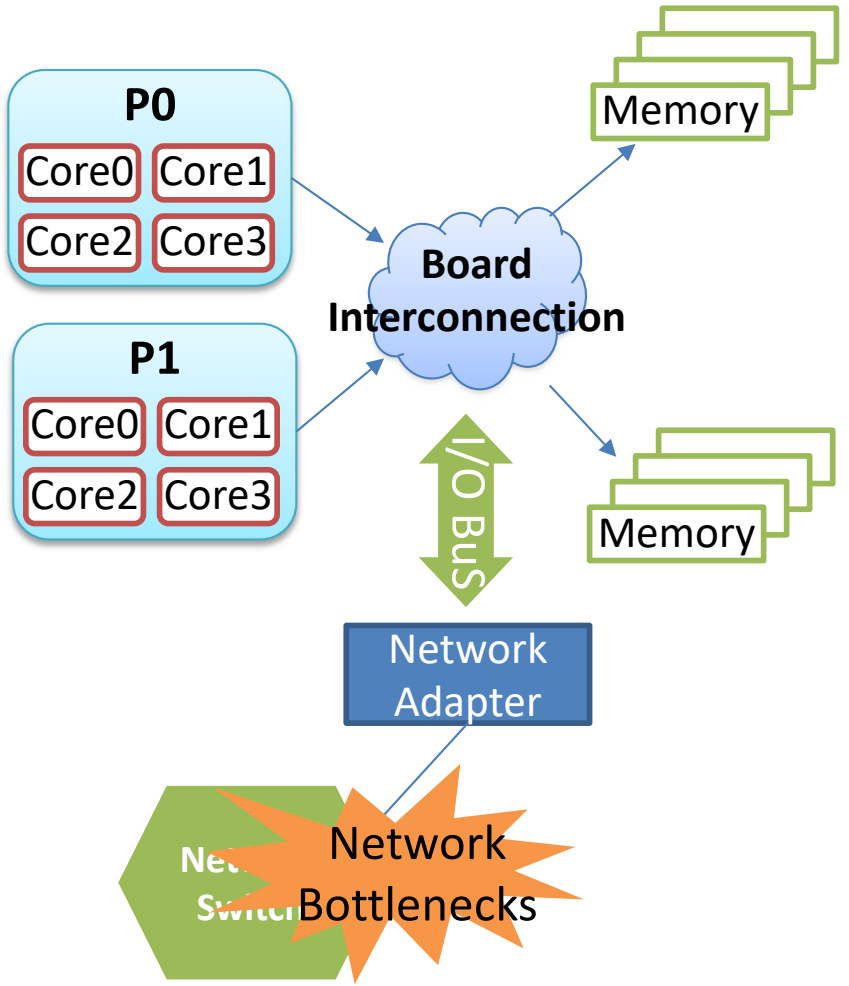
- Traditionally relied on bus-based technologies (e.g., PCI, PCI-X)
  - One bit per wire
  - Performance increase through
    - Clock speed
    - Bus width
  - Not scalable
    - Crosstalk
    - Skew between wires
    - Signal integrity limit bus width significantly



E.g., Ports using a bus speed doubled to 66 MHz and a bus width doubled to 64 bits, the pin count increased to 184 from 124)

Bottlenecks on traditional networks

- Network speeds saturated at around 1Gbps
  - Features provided were limited
  - Commodity networks were not considered scalable enough for very large-scale systems



Ethernet (1979 -)	10 Mbit/sec
Fast Ethernet (1993 -)	100 Mbit/sec
Gigabit Ethernet (1995 -)	1000 Mbit /sec
ATM (1995 -)	155/622/1024 Mbit/sec
Myrinet (1993 -)	1 Gbit/sec
Fibre Channel (1994 -)	1 Gbit/sec

InfiniBand (IB) and High-Speed Ethernet (HSE)

University of BRISTOL

- Industry Network Standards
  - InfiniBand and High-speed Ethernet were introduced into the market to address these bottlenecks
- InfiniBand aimed at all three bottlenecks
  - Protocol processing
  - I/O bus
  - Network speed
- Ethernet aimed at directly handling the network speed bottleneck and relying on complementary technologies to alleviate the protocol processing and I/O bus bottlenecks

How to tackle network speed bottlenecks with IB and high-speed Ethernet

University of BRISTOL

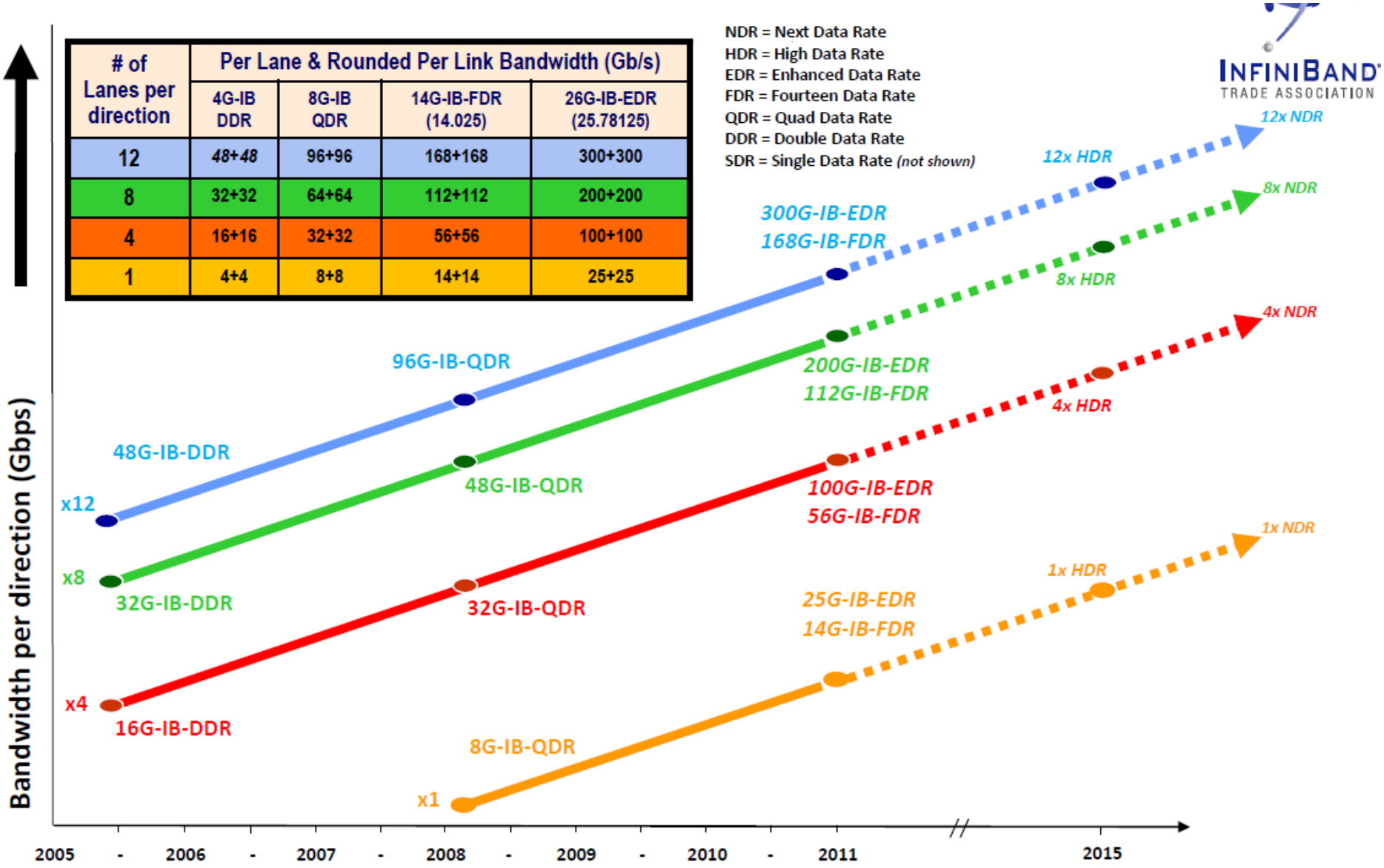
- Bit serial differential signalling
  - Independent pairs of wires to transmit independent data (lane)
  - Scalable to any number of lanes
  - Easy to increase clock speed

Ethernet (1979 -)	10 Mbit/sec
Fast Ethernet (1993 -)	100 Mbit/sec
Gigabit Ethernet (1995 -)	1000 Mbit /sec
ATM (1995 -)	155/622/1024 Mbit/sec
Myrinet (1993 -)	1 Gbit/sec
Fibre Channel (1994 -)	1 Gbit/sec
InfiniBand (2001 -)	2 Gbit/sec (1X SDR)
10-Gigabit Ethernet (2001 -)	10 Gbit/sec
InfiniBand (2003 -)	8 Gbit/sec (4X SDR)
InfiniBand (2005 -)	16 Gbit/sec (4X DDR)
	24 Gbit/sec (12X SDR)
InfiniBand (2007 -)	32 Gbit/sec (4X QDR)
40-Gigabit Ethernet (2010 -)	40 Gbit/sec
InfiniBand (2011 -)	56 Gbit/sec (4X FDR)
InfiniBand (2012 -)	100 Gbit/sec (4X EDR)

InfiniBand Link Speed Standardization Roadmap

University of BRISTOL

# of Lanes per direction	Per Lane & Rounded Per Link Bandwidth (Gb/s)			
	4G-IB DDR	8G-IB QDR	14G-IB-FDR (14.025)	26G-IB-EDR (25.78125)
12	*48+48*	96+96	168+168	300+300
8	32+32	64+64	112+112	200+200
4	16+16	32+32	56+56	100+100
1	4+4	8+8	14+14	25+25



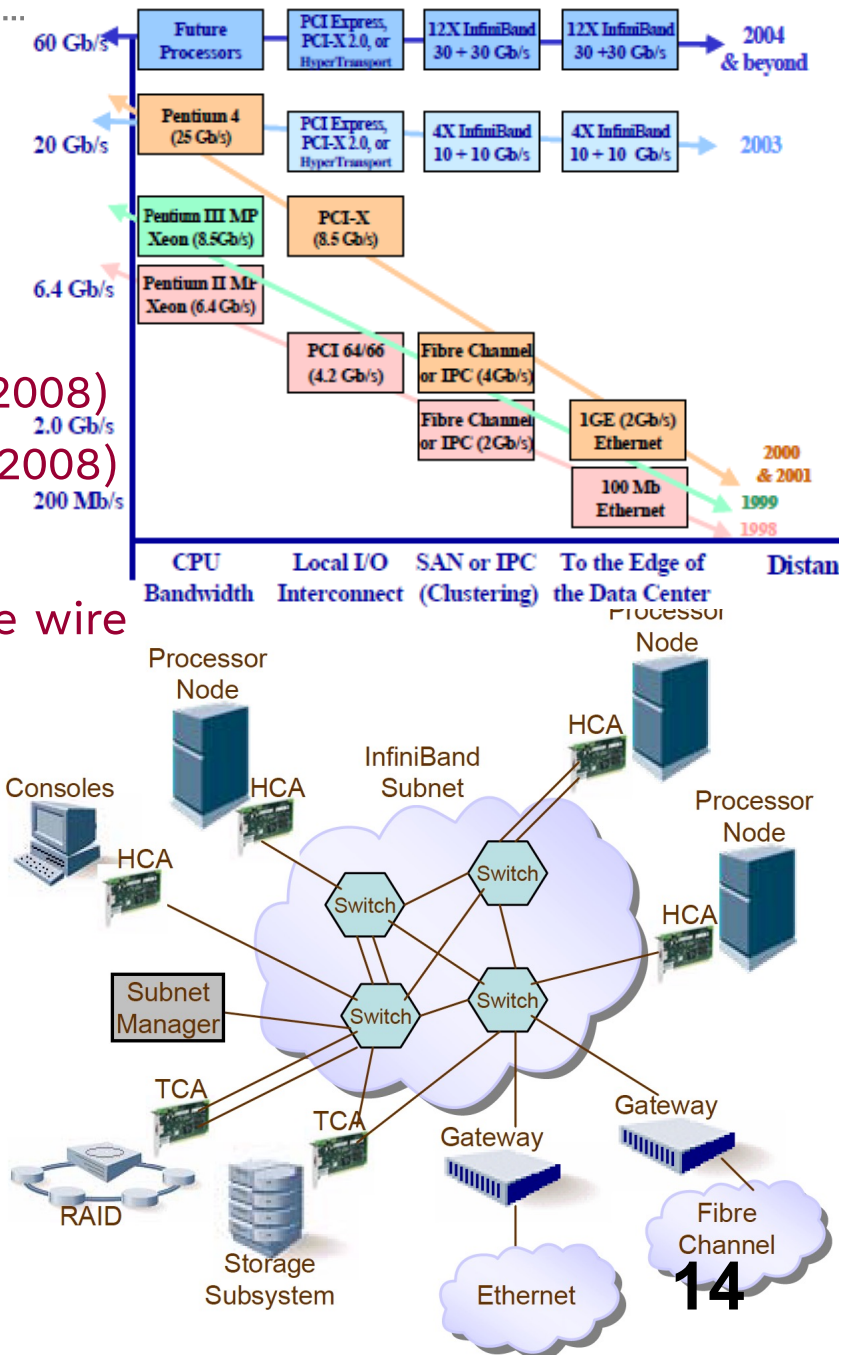
IB hardware acceleration

University of BRISTOL

- Support entire protocol processing completely in hardware (hardware protocol offload engines).
  - Some IB models have multiple hardware accelerators. E.g., Mellanox IB adapters
- Protocol Offload Engines
  - Completely implement ISO/OSI 2-4 layers in hardware (link layer, network layer and transport layer)
- No software signalling between communication layers
  - All layers are implemented on a dedicated hardware unit, and not on a shared host CPU
- Additional hardware supported features also present
  - RDMA, Multicast, QoS, Fault tolerance, and many more
- Converged (Enhanced) Ethernet (CEE or CE)
  - Combine a number of optical Ethernet standards into one umbrella

InfiniBand Storage Solutions

- InfiniBand standard – strong alternative to
  - Fiber Channel (SAN)
  - Ethernet (NAS)
- Superior performance
  - 20Gb/s host/target ports (moving to 40Gb/s 2008)
  - 60Gb/s switch to switch (moving to 120Gb/s 2008)
- Unified fabric for datacenter
  - Storage, networking and clustering in a single wire
- Cost effective
  - Compelling price/performance advantage
- Less power consumption
  - Less than 5w per 20Gb/s port
- Other
  - High reliable fabric
  - Multi-pathing
  - Automatic failover



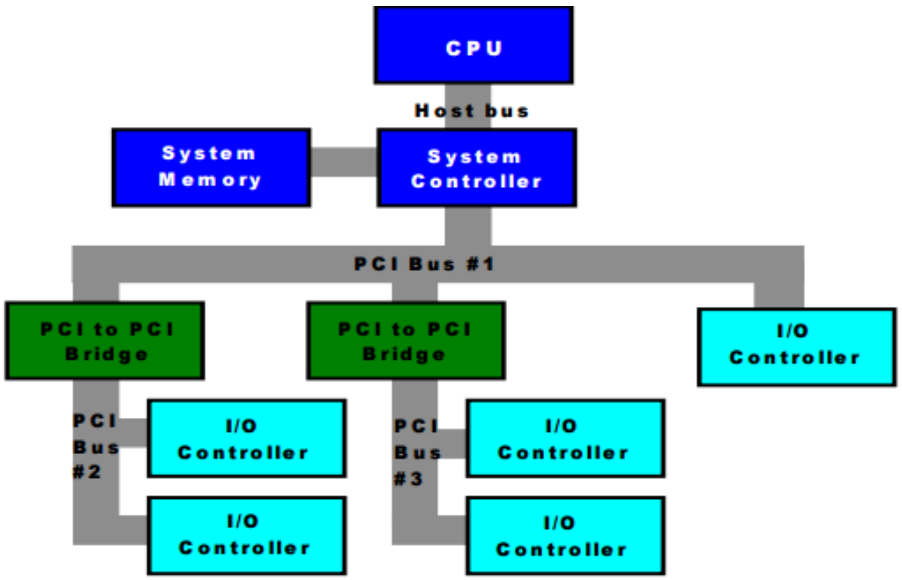
Interplay with I/O technologies

University of BRISTOL

- InfiniBand initially intended to replace I/O bus technologies with networking-like technologies
- Both IB and HSE today come as network adapters that plug into existing technologies
- Recent trend in I/O interfaces show I/O matching head-to-head with network speeds

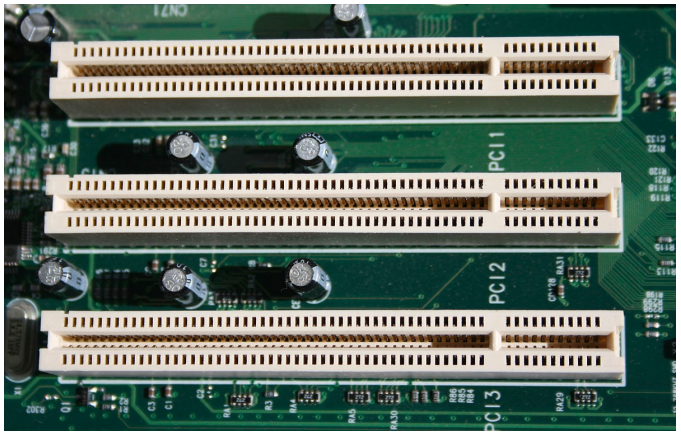
PCI	1990	33MHz/32bit: 1.05Gbps (shared bidirectional)
PCI-X	1998 (v1.0) 2003 (v2.0)	133MHz/64bit: 8.5Gbps (shared bidirectional) 266-533MHz/64bit: 17Gbps (shared bidirectional)
AMD HyperTransport (HT)	2001 (v1.0), 2004 (v2.0) 2006 (v3.0), 2008 (v3.1)	102.4Gbps (v1.0), 179.2Gbps (v2.0) 332.8Gbps (v3.0), 409.6Gbps (v3.1) (32 lanes)
PCI-Express (PCIe) by Intel	2003 (Gen1), 2007 (Gen2) 2009 (Gen3 standard)	Gen1: 4X (8Gbps), 8X (16Gbps), 16X (32Gbps) Gen2: 4X (16Gbps), 8X (32Gbps), 16X (64Gbps) Gen3: 4X (~32Gbps), 8X (~64Gbps), 16X (~128Gbps)
Intel QuickPath Interconnect (QPI)	2009	153.6-204.8Gbps (20 lanes)

Typical PCI interface



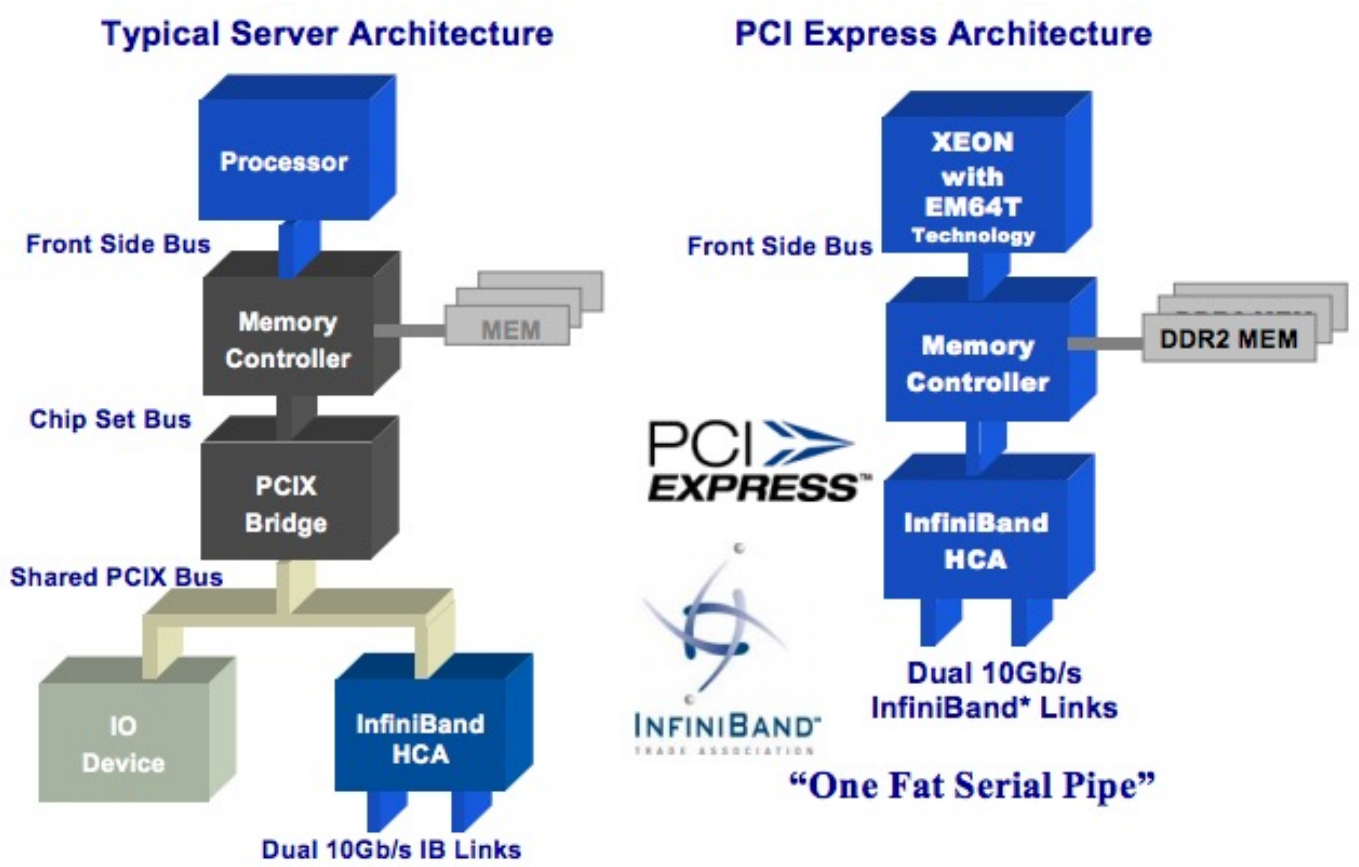
Example: Transfer 64K Bytes of Data
Processor: Intel 2.8GHz Xeon™ CPUs
Chipset/Memory: DDR 300MHz, 128 bit wide
I/O: PCI-X and Gigabit Ethernet

Sub-System	Execution Time Contribution
I/O	1092.3us
Memory	27.3us
Processing	70.2us
Total Execution Time	1189.8us



Sub-System	Execution Time Contribution
I/O (InfiniBand & PCI Express)	72.8us
Memory (300MHz/128bits)	27.3us
Processing (2.8GHz CPU)	70.2us
Total Execution Time	170.3us

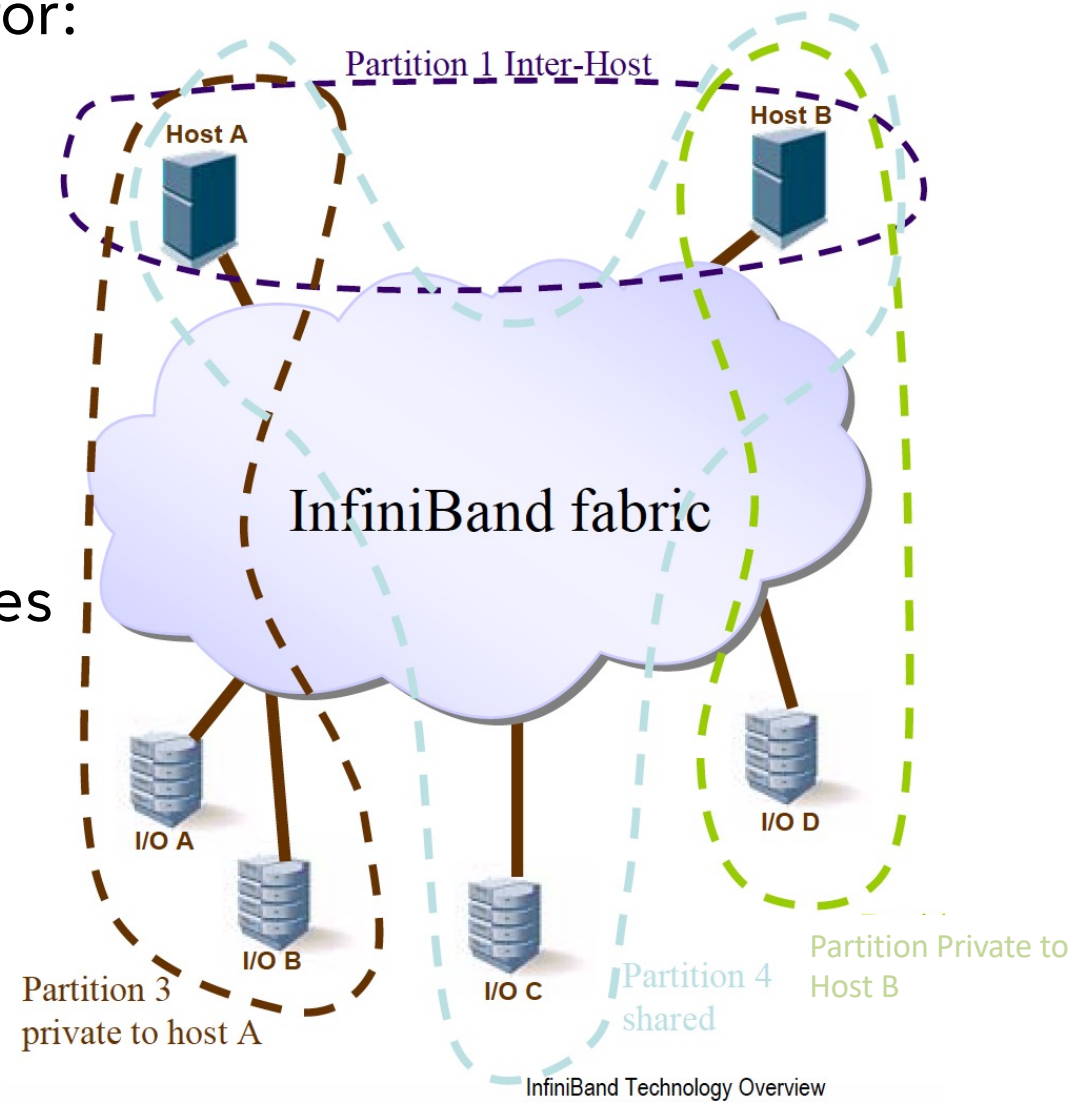
PCI Express Architecture vs Typical Server Architecture



With InfiniBand and PCI Express there is one fat serial pipe directly between servers to the CPU and memory sub-systems. This results in ***reduced chip count and complexity*** and ***improves both bandwidth and latency***, and as will be shown ***overall system level balance and performance.***

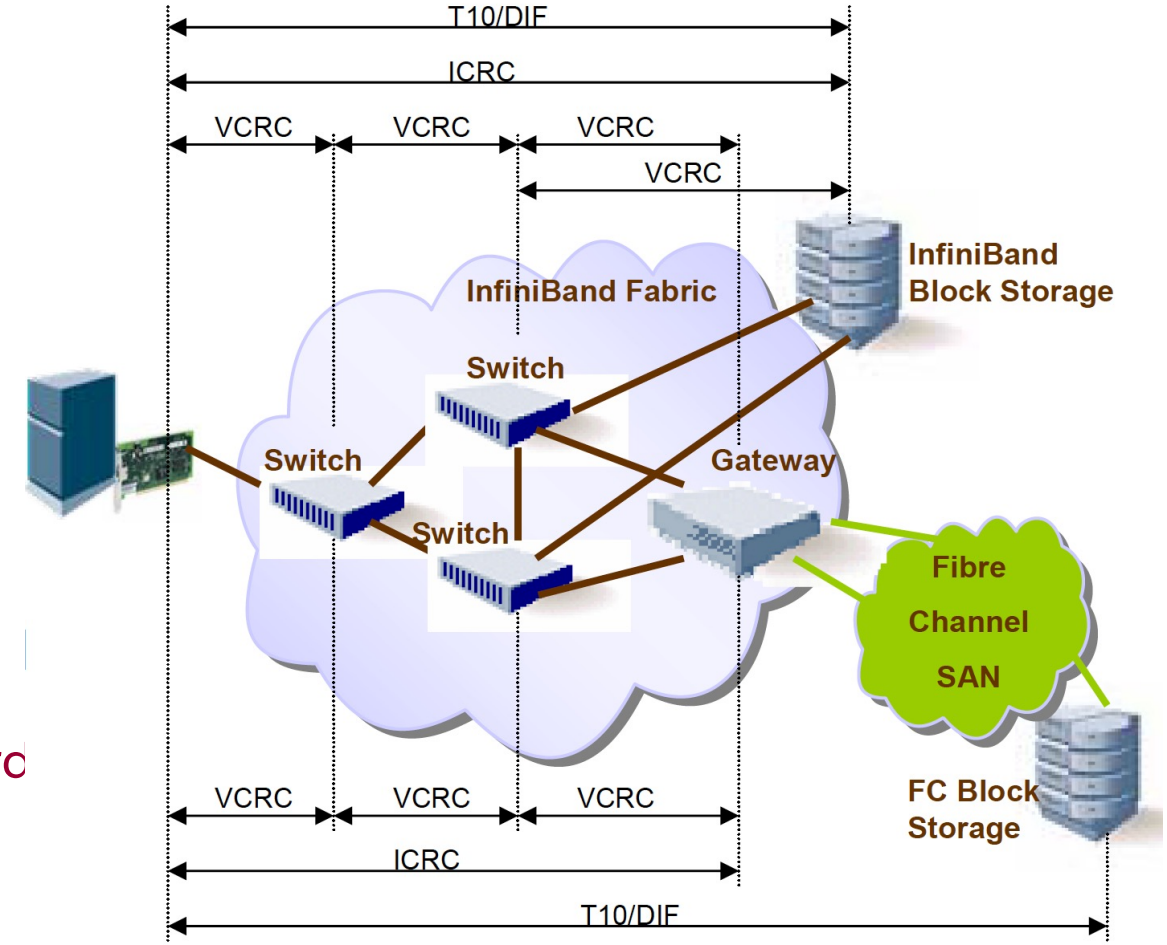
InfiniBand Network Partition

- Define different partitions for:
  - Different customers
  - Different application
  - Security purposes
  - Quality of Service
- Partition filtering at switches
- Logically divide the fabric into isolated domains
- Similar to
  - FiberChannel Zoning
  - 802.1Q VLANs



IB Data Integrity

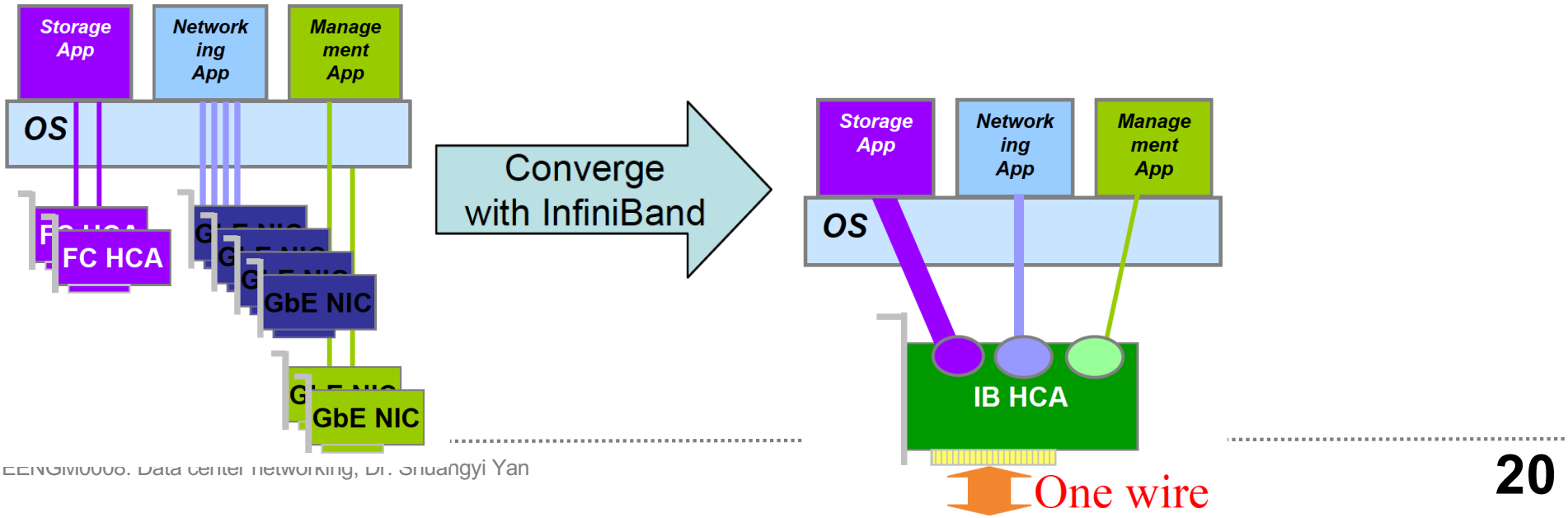
- Hop by hop
  - VCRC – 16 bit CRC
  - CRC16 0x100B
- End to end
  - ICRC – 32 bit CRC
  - CRC32 0x04C11DB7
  - Same CRC as Ethernet
- Application level
  - T10/DIF Logical Block Guard
    - Per block CRC
  - 16 bit CRC 0x8BB7



I/O Consolidation

- Slower I/O
- Different service needs – different fabrics
- No flexibility
- More ports to manage
- More power
- More space
- Higher TCO

- High bandwidth pipe for capacity provisioning
- Dedicated I/O channels enable convergence
  - For Networking, Storage, Management
  - Application compatibility
  - QoS – differentiates different traffic types
  - Partitions – logical fabrics, isolation
- Gateways – Share remote Fiber Channel and ETH ports
  - Design based on average load across multiple servers
  - Scale incrementally – add Ethernet/FC/Server blades
  - Scale independently

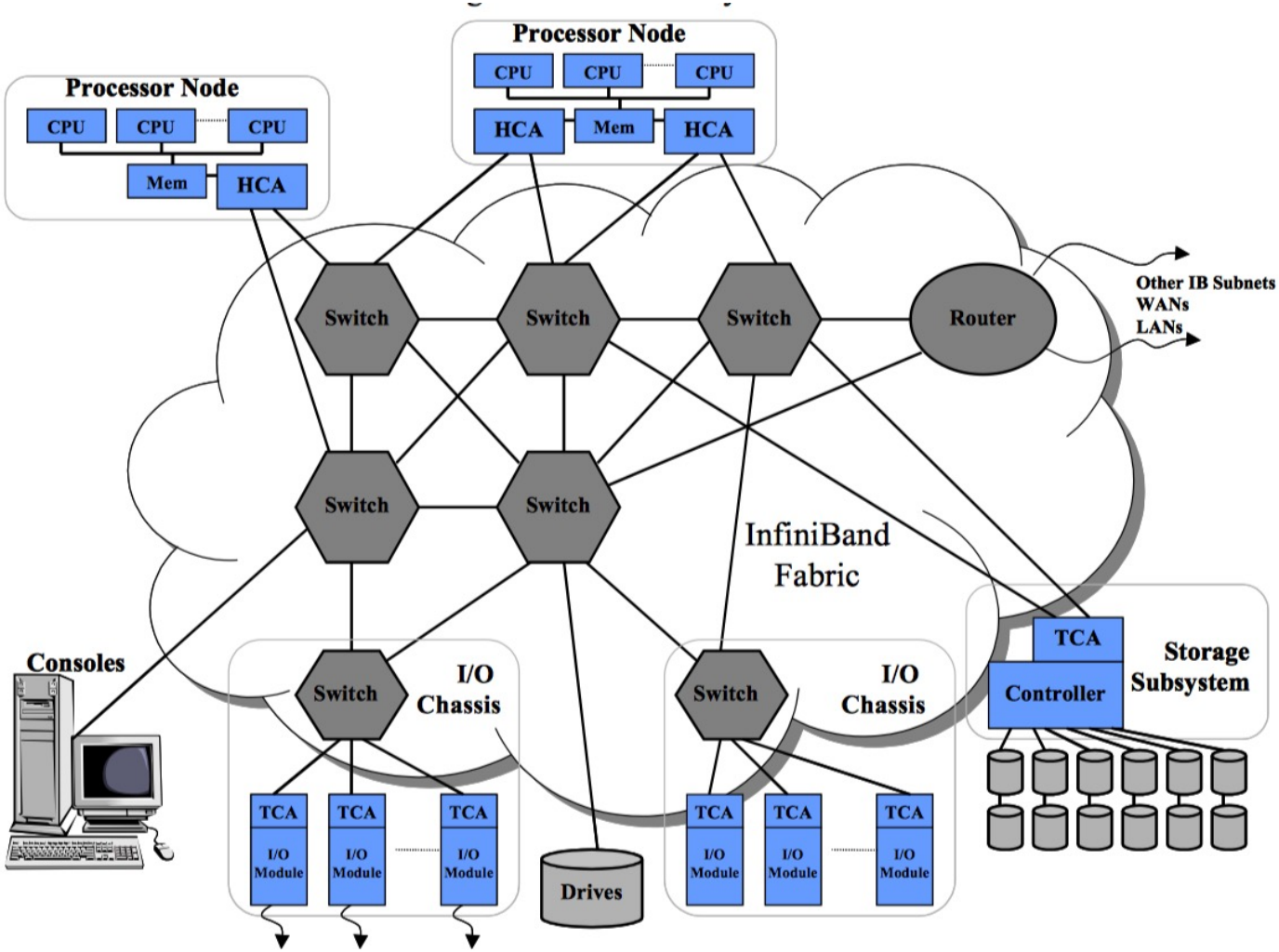


High Availability and Redundancy

University of BRISTOL

- Multi-port HCAs
  - Covers link failure
- Redundant fabric topologies
  - Covers link failure
- Link layer multi-pathing (LMC)
- Automatic Path Migration (APM)
- ULP High Availability
  - Application level multi-pathing (SRP/iSER)
  - Teaming/Bonding (IPoIB)
  - Covers HCA failure and link failure

InfiniBand System Fabric



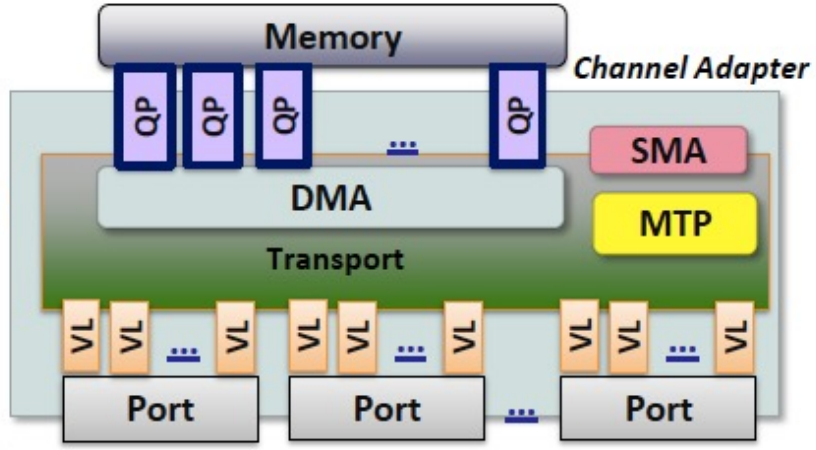
Ref: https://www.mellanox.com/pdf/whitepapers/IB_Intro_WP_190.pdf

University of BRISTOL

Introduction of InfiniBand
Part 2

IB Components

- Channel Adapters
  - Host Channel Adapter (HCA)
    - Device that terminates an IB link and executes transport-level functions and support the verbs interface
    - Present in servers.



  - Target Channel Adapter (TCA)
    - Present on I/O devices, i.e. RAID (Redundant Array of Inexpensive/Independent Disks), JBOD (just a bunch of disks) subsystem.

Host Channel Adapter (HCA)

University of BRISTOL

- Equivalent to a NIC (Ethernet)
  - GUID (Global Unique ID)
- Converts PCI to InfiniBand
- CPU offload of transport operations
- End-to-end QoS and congestion control
- HCA bandwidth options:

• SDR - Single Data Rate	2.5 Gb/s * 4	= 10	
• DDR - Double Data Rate	5 Gb/s * 4	= 20	
• QDR - Quadruple Data Rate	10 Gb/s * 4	= 40	
• FDR - Fourteen Data Rate	14 Gb/s * 4	= 56	
• EDR - Enhanced Data Rate	25 Gb/s * 4	= 100	

IB Components

- Switch
  - A device that moves packets from one link to another of the same IB Subnet

- Router
  - A device that transports packets between different IBA subnets

- Bridge/Gateway
  - InfiniBand to Ethernet

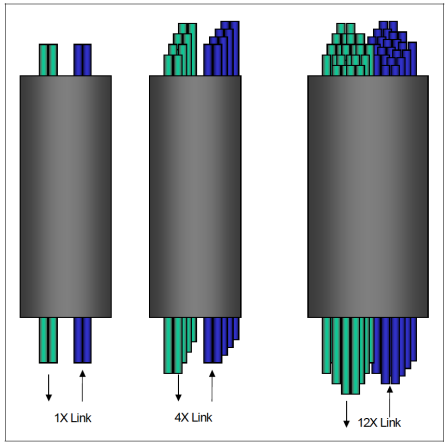
- Link & Repeaters

(Courtesy Intel)

IB Physical Layer

University of BRISTOL

- IB defines 3 link speeds at physical layer:
  - 1X, 4X, 8X 12X including auto-negotiation
  - Speed (SDR/DDR/QDR) including auto-negotiation
  - Each link is a four-wire serial differential connection
    - Two wires in each direction (2 complementary signals to eliminate electromagnetic noise) providing full duplex connection
  - 8/10 encoding (actual raw data bandwidth drops to 2.0 Gbps from a 2.5 Gbps theoretical) to maintain DC balance and offer limited run length of 0's and 1's



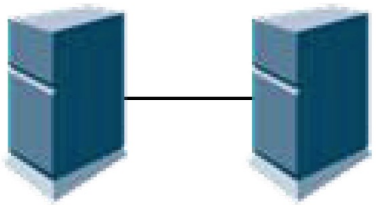
Link Speed (10^9 bit/sec)

Lane Speed → / Link Width ↓	SDR (2.5GHz)	DDR (5GHz)	QDR (10GHz)
1X	2.5	5	10
4X	10	20	40
8X	20	40	80
12X	30	60	120

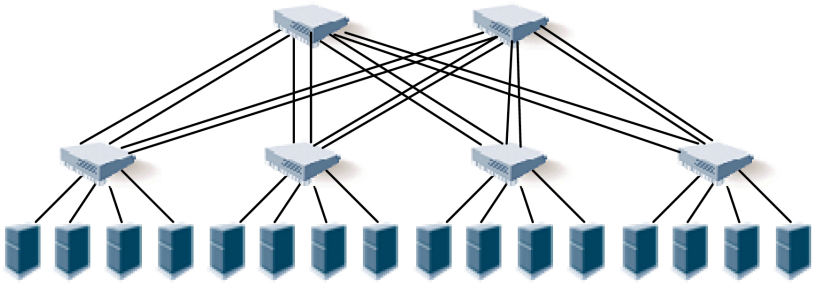
- Control symbols used for Lane de-skew, auto-negotiation, training, clock tolerance, framing

InfiniBand (IB) Topologies

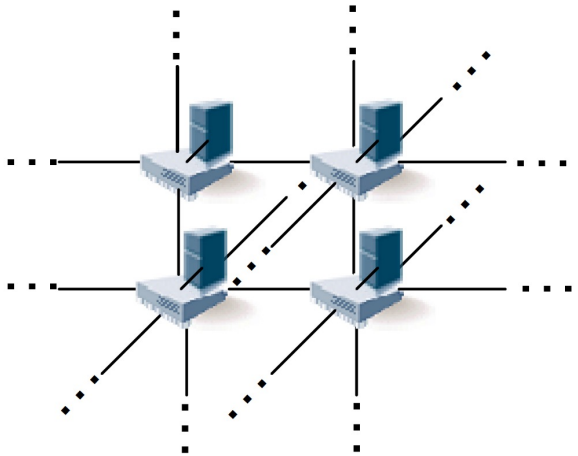
- Example topologies commonly used
- Modular switches are based on fat tree architecture



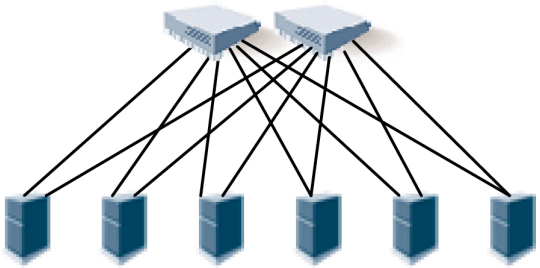
Back to Back



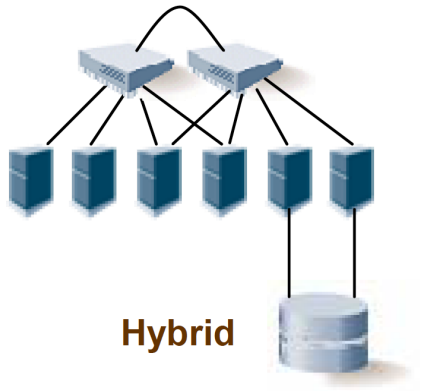
2 Level Fat Tree



3D Torus



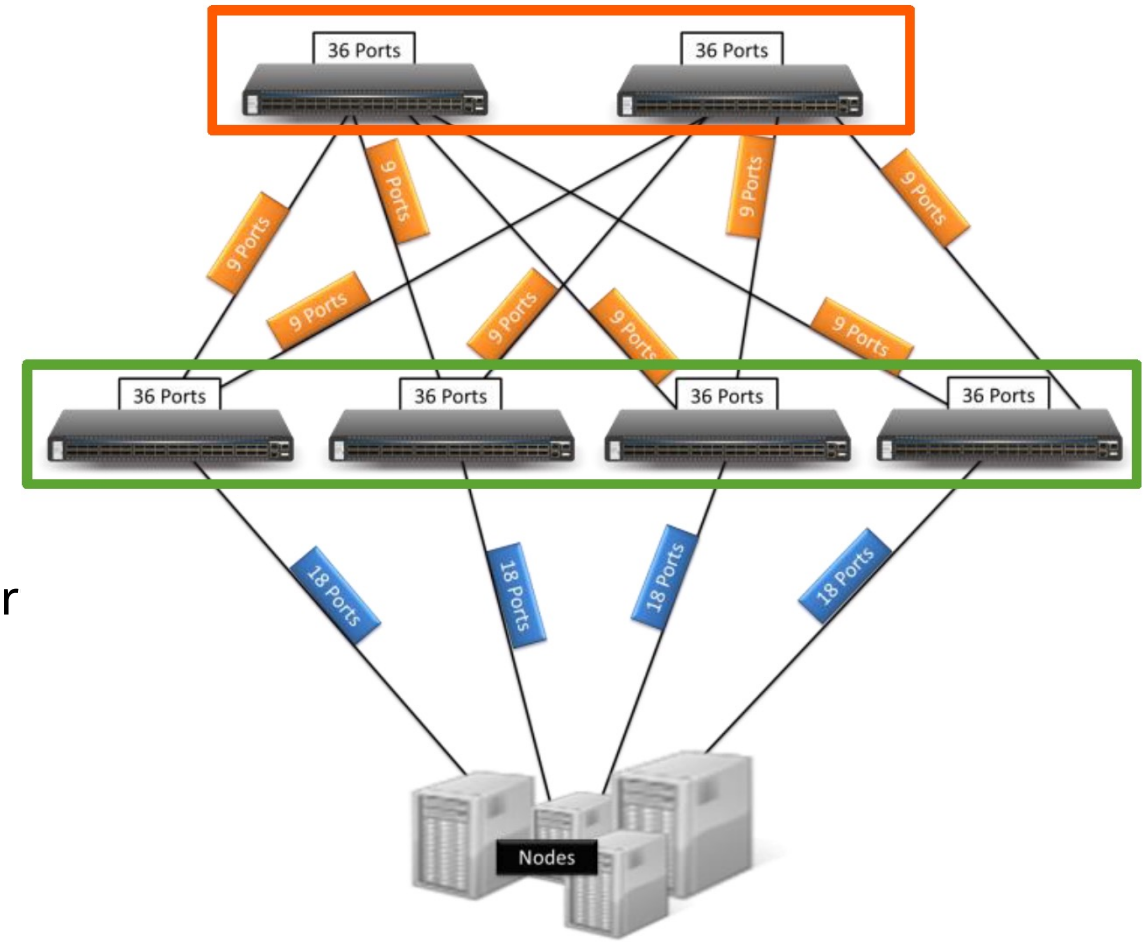
Dual Star



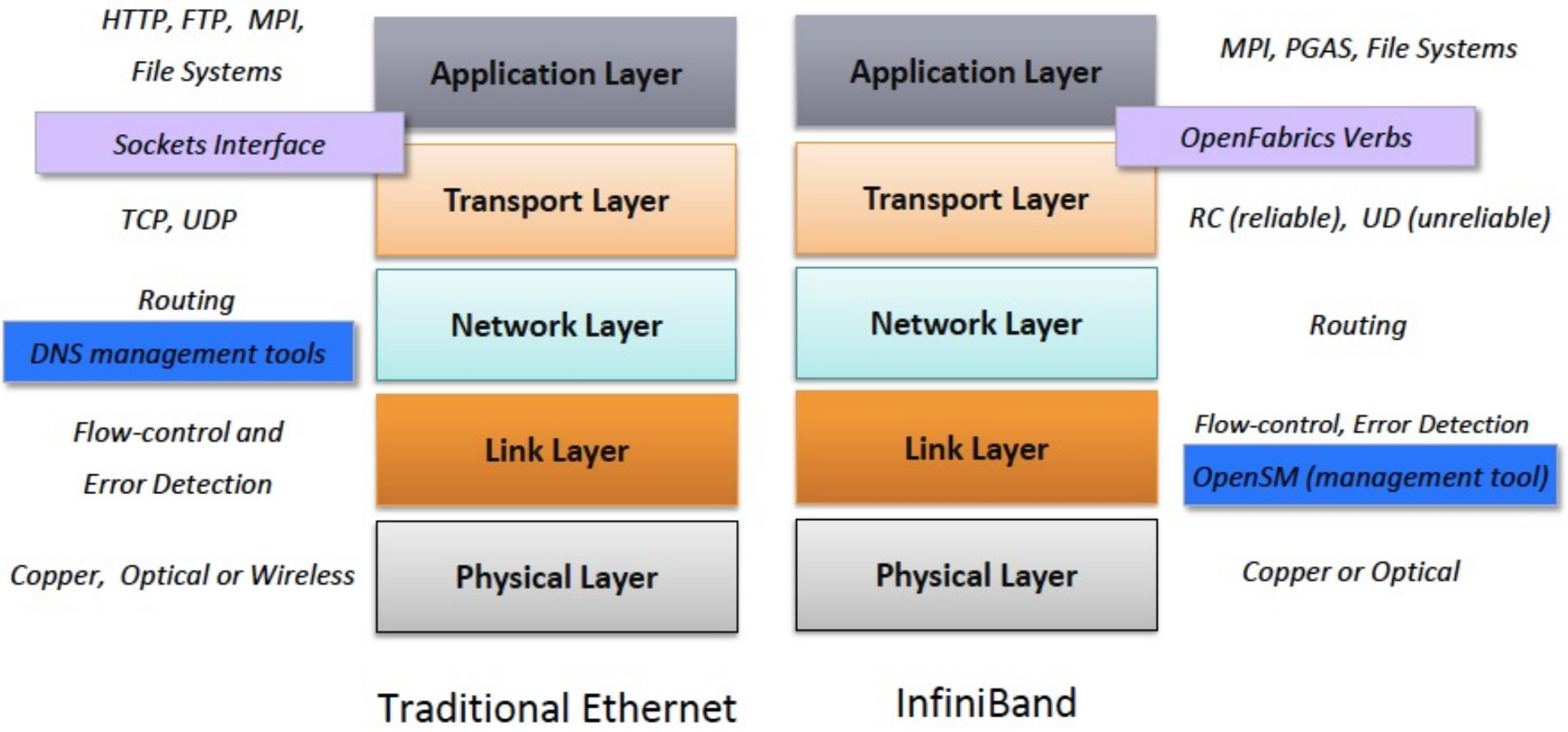
Hybrid

IB Fabric Building Block

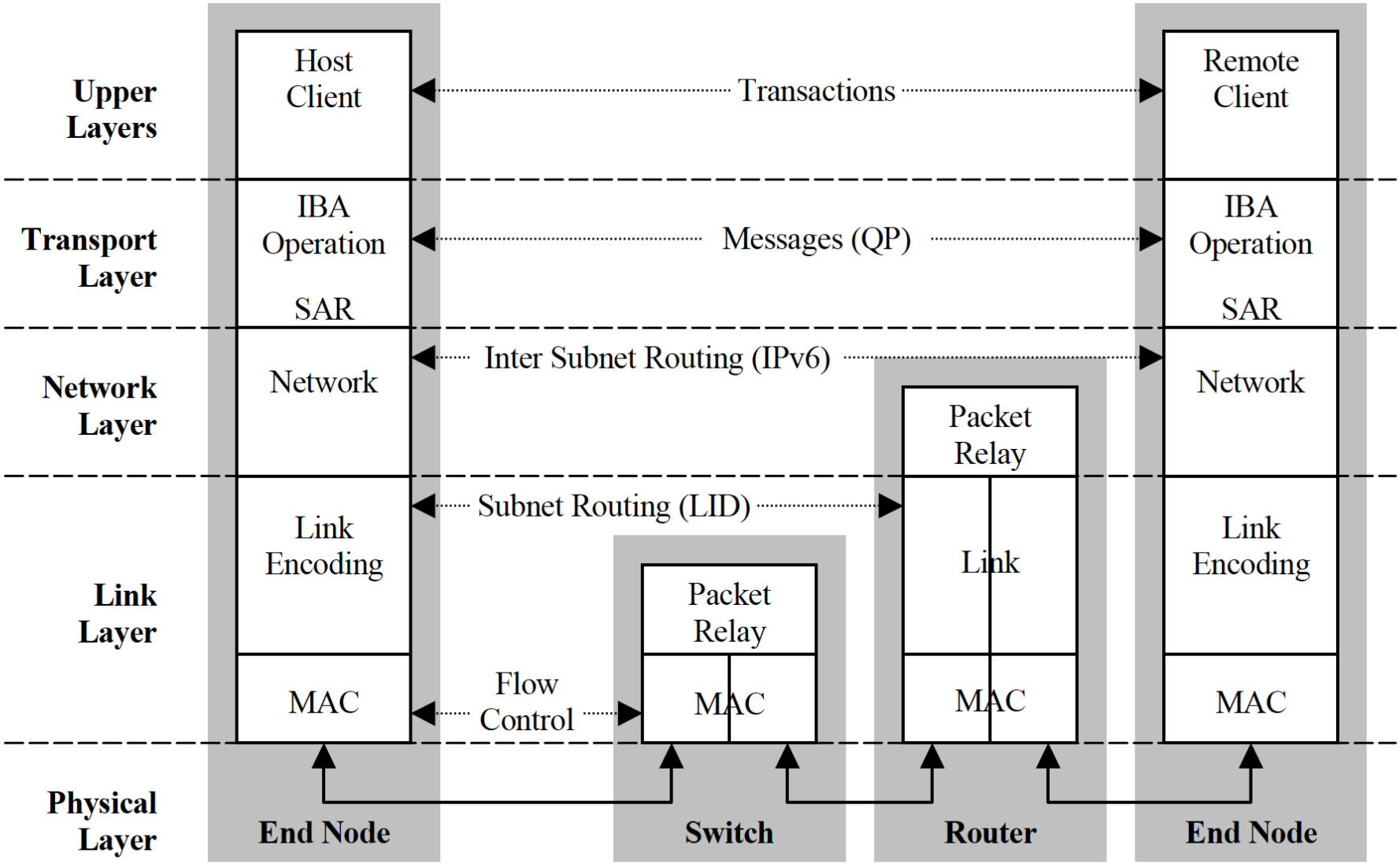
- Basic Block fro every IB switch module is a single 36-port IB switch chip
- Higher port-count switching module can be created using multiple chips
- Examples(on the right): 72 por switch is created using 6 identical chips:
  - 4 chips will function as lines
  - 2 chips will function as core



Comparing InfiniBand with Traditional Networking Stack

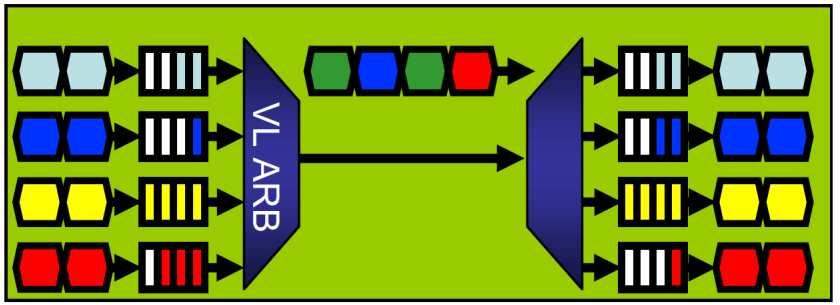


InfiniBand Protocol Layers

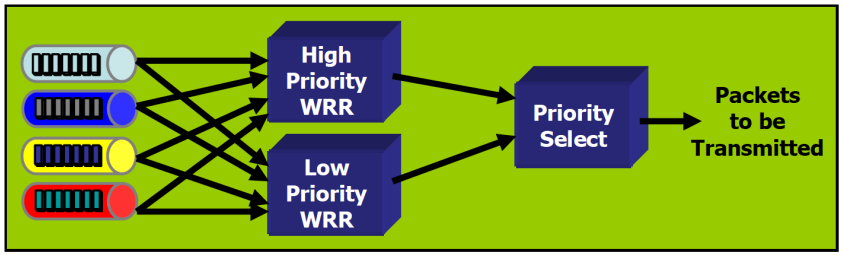


IB Link Layer

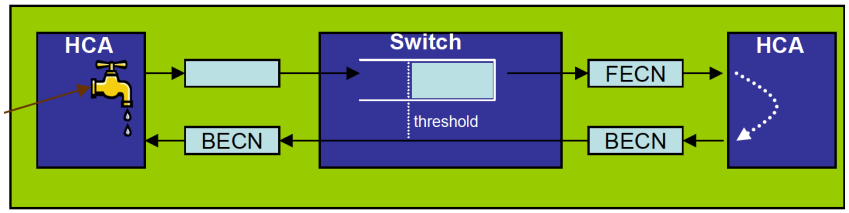
- The link layer (along with the transport layer) is the heart of the InfiniBand Architecture.
- The link layer encompasses packet layout, point-to-point link operations, and switching within a local subnet.
- Addressing and Switching
  - Local Identifier (LID) addressing
  - Unicast LID – 48K addresses
  - Multicast LID – up to 16K addresses
  - Efficient linear lookup
  - Multi-path support through LMC
- Independent Virtual Lanes
  - Flow control (lossless fabric)
  - Server Level
  - VL arbitration for QoS
- Congestion Control
  - Forward/Backward Explicit Congestion Notification (FECN/BECN)
- Data Integrity
  - Invariant CRC
  - Variant CRC



Independent Virtual Lanes (VLs)



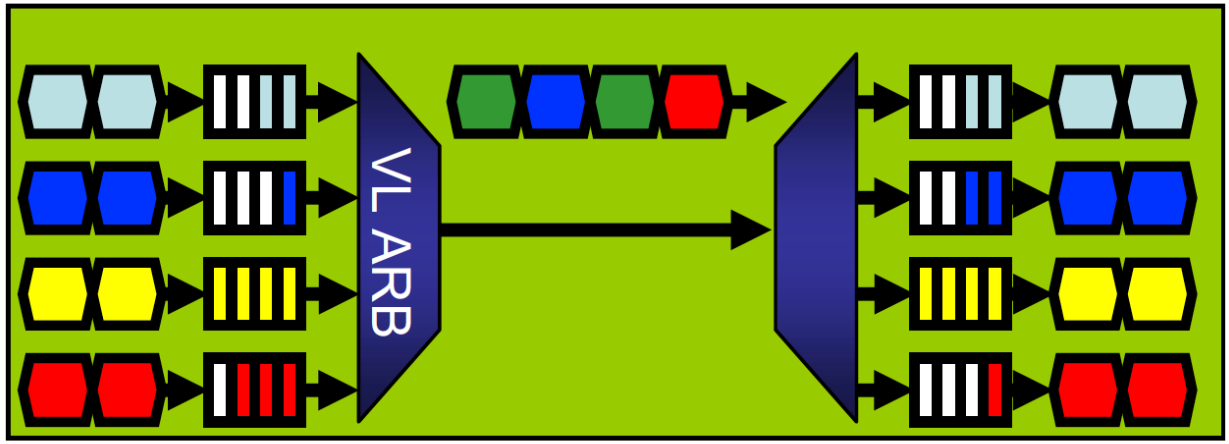
H/L Weighted Round Robin (WRR) VL Arbitration



Efficient FECN/BECN Based Congestion Control

IB Link Layer – Cont'd

- QoS is supported by InfiniBand through virtual lanes (VL).
  - These VLs are separate logical communication links that share a single physical link.
  - Each link can support up to 15 standard VLs and one management lane (VL 15). VL15 is the highest priority and VL0 is the lowest.
  - Management packets use VL15 exclusively.



Independent Virtual Lanes (VLs)

IB Link Layer – Cont'd

University of BRISTOL

- The link layer encompasses packet layout, point-to-point link operations, and switching within a local subnet.
- Packets
  - There are two types of packets in the link layer: management and data packets.
    - Management packets are used for link configuration and maintenance. Device information, such as virtual lane support, is determined with management packets.
    - Data packets carry up to 4 KB of a transaction payload.
- Switching
  - Within a subnet, packet forwarding and switching is handled at the link layer.
    - All devices within a subnet have a 16-bit local identifier (LID) assigned by the subnet manager.
    - All packets sent within a subnet use the LID for addressing.
    - Link Level switching forwards packets to the device specified by a destination LID within a local route header (LRH) in the packet. The LRH is present in all packets.

IB Link Layer – Cont'd

University of BRISTOL

- Credit-based flow control
  - Link-level flow control is used to manage data flow between two point-to-point links.
    - Flow control is handled on a per-VL basis, allowing separate virtual fabrics to maintain communication utilizing the same physical media.
    - Each receiving end of a link supplies credits to the sending device on the link to specify the amount of data that can be received without loss of data.
    - Credit passing between each device is managed by a dedicated link packet to update the number of data packets the receiver can accept. Data is not transmitted unless the receiver advertises credits indicating that receive buffer space is available.

IB Link Layer – Cont'd

University of BRISTOL

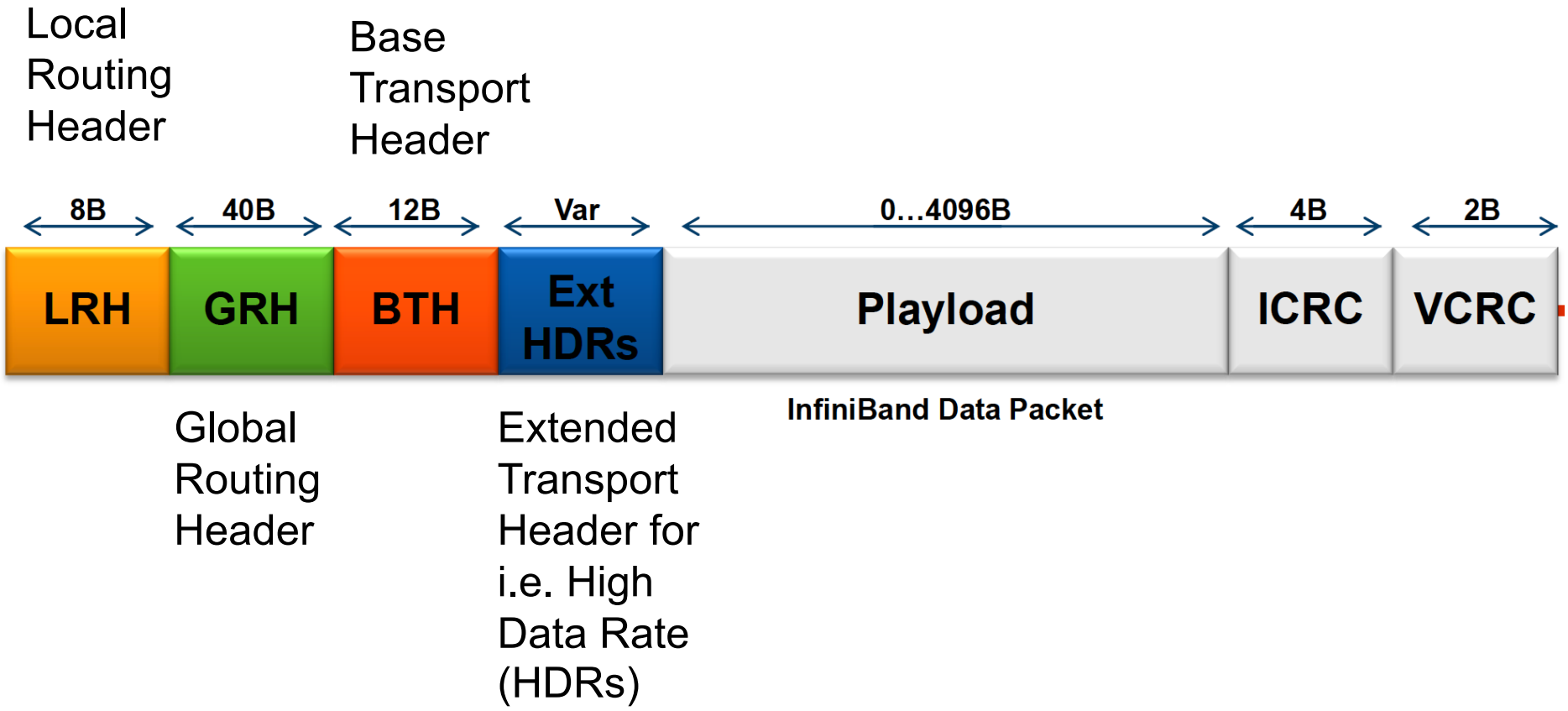
- Data integrity
  - At the link level, there are two CRCs per packet, variant CRC (VCRC) and invariant CRC (ICRC), that ensure data integrity.
    - The 16-bit VCRC includes all fields in the packet and is recalculated at each hop.
    - The 32-bit ICRC covers only the fields that do not change from hop to hop. T
    - he VCRC provides link-level data integrity between two hops and the ICRC provides end-to-end data integrity. In a protocol such as Ethernet, which defines only a single CRC, an error can be introduced within a device that then recalculates the CRC.
    - The check at the next hop would reveal a valid CRC, even though the data has been corrupted. InfiniBand includes the ICRC so that when a bit error is introduced, the error will always be detected.

IB Layer 2 Switch addressing: Local Identifier (LID)

University of BRISTOL

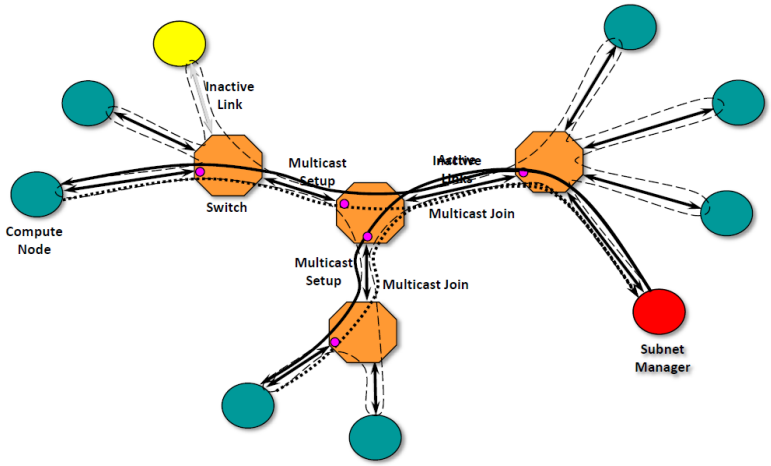
- Local Identifier: 16-bit Layer 2 Address
  - Assigned by the Subnet Manager when port becomes active
  - Not persistent through reboots
- Global Identifier (GID) addressing = {64 bit GID prefix, 64 bit GUID}
  - GUID = Global Unique Identifier – 64 bit
  - GUID 0 – assigned by the manufacturer
  - GUID 1 … (N-1) – assigned by the Subnet Manager

InfiniBand Data Packet



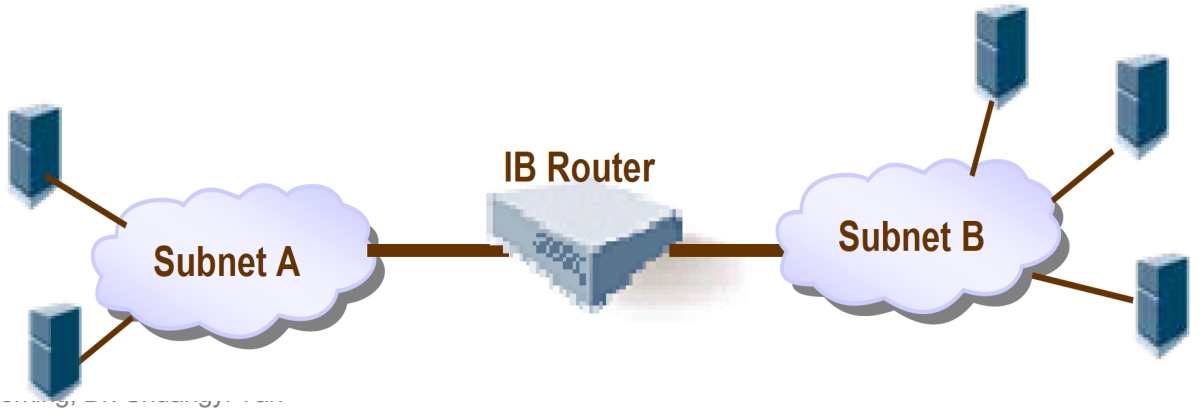
Subnet Manager

- Agents
  - Processes or hardware units running on each adapter, switch, router
  - Provide capability to query and set parameters
- Managers
  - Make high-level decisions and implement it on the network fabric using the agents
- Messaging schemes
  - Used for interactions between the manager and agents
- Messages



IB Network Layer

- The network layer handles routing of packets from one subnet to another.
  - Packets that are sent between subnets contain a global route header (GRH).
  - The GRH contains the 128-bit IPv6 address for the source and destination of the packet.
  - The packets are forwarded between subnets through a router, based on each device's 64-bit Globally Unique Identifier (GUID).
  - The router modifies the LRH with the proper local address within each subnet.
  - Therefore, the last router in the path replaces the LID in the LRH with the LID of the destination port.



Global Unique Identifier (GUID) – Physical Address

University of BRISTOL

- Any InfiniBand node requires GUID & LID (Local Identifier) addresses

- GUID (Global Unique Identifier)- 64 bits address, “Like a Ethernet MAC address” ?
  - Assigned by IB vendor
  - Persistent through reboots

- IB Switch “Multiple” Address GUIDS
  - Node = Is meant to identify the HCA as a entity
  - Port = Identifies the port as a port
  - System = Allows to combine multiple GUIDS creating one entity

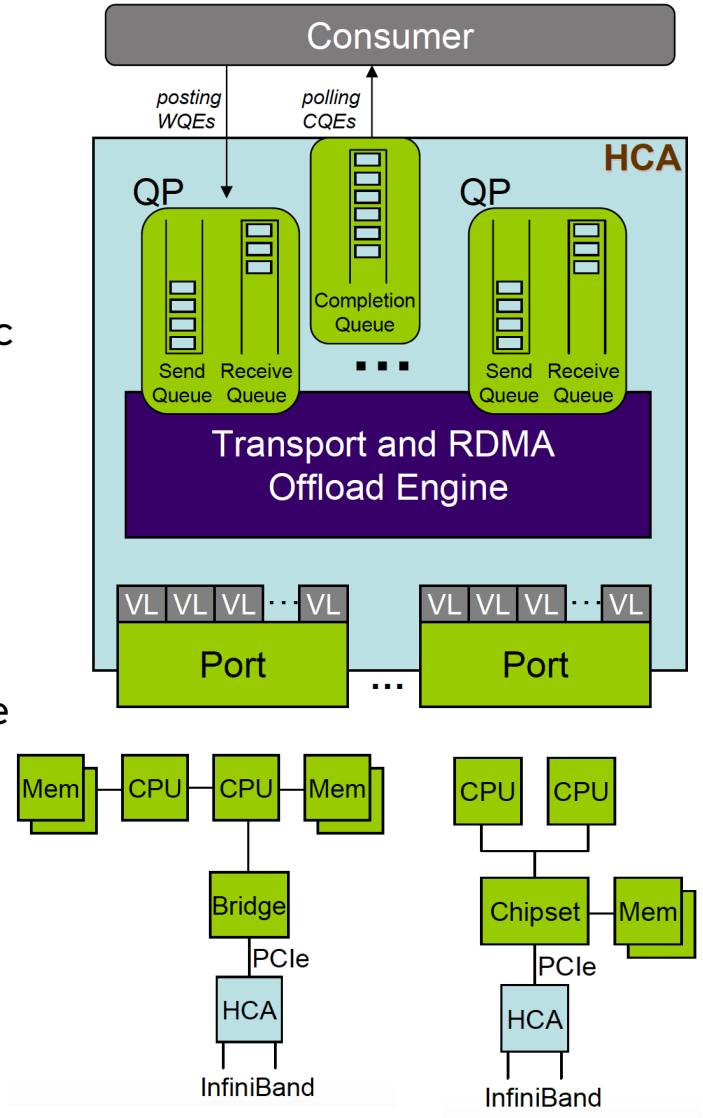
IB Transport Layer

University of BRISTOL

- The transport layer is responsible for:
  - in-order packet delivery, partitioning, channel multiplexing, and transport services
    - reliable connection, unreliable connection
    - reliable datagram, unreliable datagram, raw datagram
- InfiniBand architecture offers a significant improvement for the transport layer: all functions are implemented in the hardware.

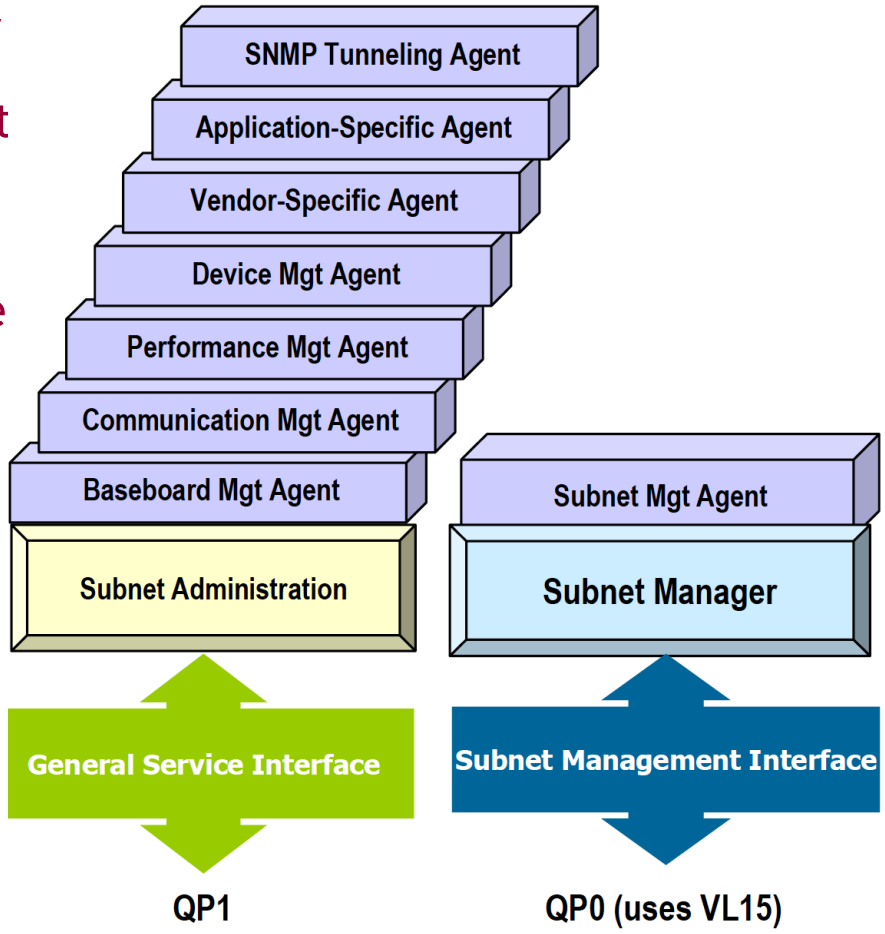
IB Transport Layer – Cont'd

- Queue Pair (QP) – transport endpoint
  - Asynchronous interface
    - Send Queue, Receive Queue, Completion Queue
  - Full transport offload
    - Segmentation, reassembly, timers, retransmission, etc
  - Operations supported
    - Send/Receive – messaging semantics
    - RDMA Read/Write – enable zero copy operations
    - Atomics – remote Compare & Swap, Fetch and Add
    - Memory management – Bind/Fast Register/Invalidate
- Kernel bypass
  - Enables low latency and CPU offload
  - Enabled through QPs, Completion Queues (CQs),
  - Protection domains (PD), Memory Regions (MRs)



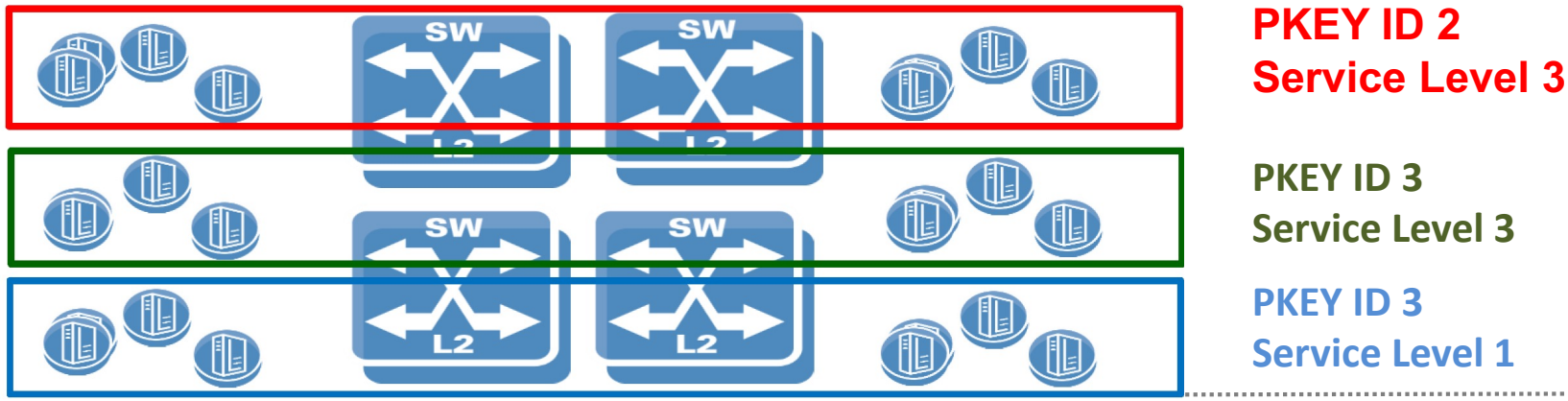
Management Model

- Subnet Manager (SM)
  - Configures/Administers fabric topology
  - Implemented at an end-node or a swit
  - Active/Passive model when more than
  - Talks with SM Agents in nodes/switche
- Subnet administration
  - Provides path records
  - QoS management
- Communication Management
  - Connection establishment processing



InfiniBand Network Partition

- Define different partitions for:
  - Different customers
  - Different application
  - Security purposes
  - Quality of Service
- A partition is defined as a collection of Channel Adapters (i.e. Host Channel Adapter (HCA)), router and switch ports that are permitted to communicate with one another. A port may be a member of multiple partitions simultaneously.
- Each partition has an Partition Key (PKEY) Identifier. A partition is represented by a 15-bit partition ID theoretically permitting up to 32K partitions. One or more PKEYs are assigned to a port by the subnet's Partition Manager (PM).



Main applications for InfiniBand

University of BRISTOL

- Application Clustering
  - A cluster is simply a group of servers connected by load balancing switches working in parallel to serve a particular application.
- Storage Area Networks
  - Simplified between storage and server.
    - Removal of the Fibre Channel network allows servers to directly connect to a storage area network without a costly HBA.
  - Remote DMA (RDMA) support simulta-neous peer to peer communication and end to end flow control.
    - InfiniBand overcomes the defi-ciencies of Fibre Channel without the need of an expensive, complex HBA.
- Inter-Tier Communication
- Inter-Processor Communication
  - The higher bandwidth connections (4X, 12X) defined by InfiniBand provide backbone capabilities for IPC clusters without the need of a secondary I/O interconnect.

Summary

University of BRISTOL

- Introduce InfiniBand
  - Understand the bottleneck for high performance computing
  - How the design in InfiniBand handle all bottlenecks