InfiniBand and 10-Gigabit Ethernet for Dummies

A Tutorial at SC '08

by

Dhabaleswar K. (DK) Panda
The Ohio State University
E-mail: panda@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~panda

Pavan Balaji
Argonne National Laboratory
E-mail: balaji@mcs.anl.gov
http://www.mcs.anl.gov/~balaji

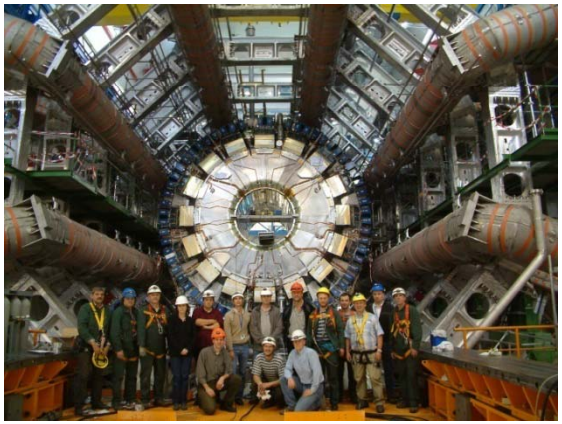
Matthew Koop
The Ohio State University
E-mail: koop@cse.ohio-state.edu
http://www.cse.ohio-state.edu/~koop

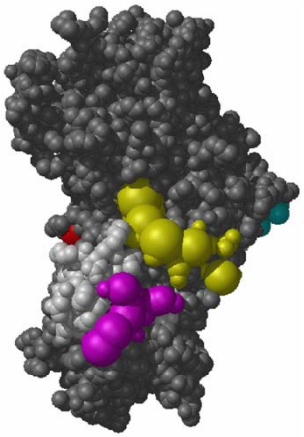
Presentation Overview

- Introduction
- Why InfiniBand and 10-Gigabit Ethernet?
- Overview of IB and 10GE and their Novel Features
- IB and 10GE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
  - MPI, SDP, File Systems, Data Center and Virtualization
- Conclusions and Final Q&A

Current and Next Generation Applications and Computing Systems

- Big demand for
  - High Performance Computing (HPC)
  - File-systems, multimedia, database, visualization
  - Internet data-centers
- Processor performance continues to grow
  - Chip density doubling every 18 months
  - Multi-core chips are emerging
- Commodity networking also continues to grow
  - Increase in speed and features
  - Affordable pricing
- Clusters are increasingly becoming popular to design next generation computing systems
  - Scalability, Modularity and Upgradeability with compute and network technologies



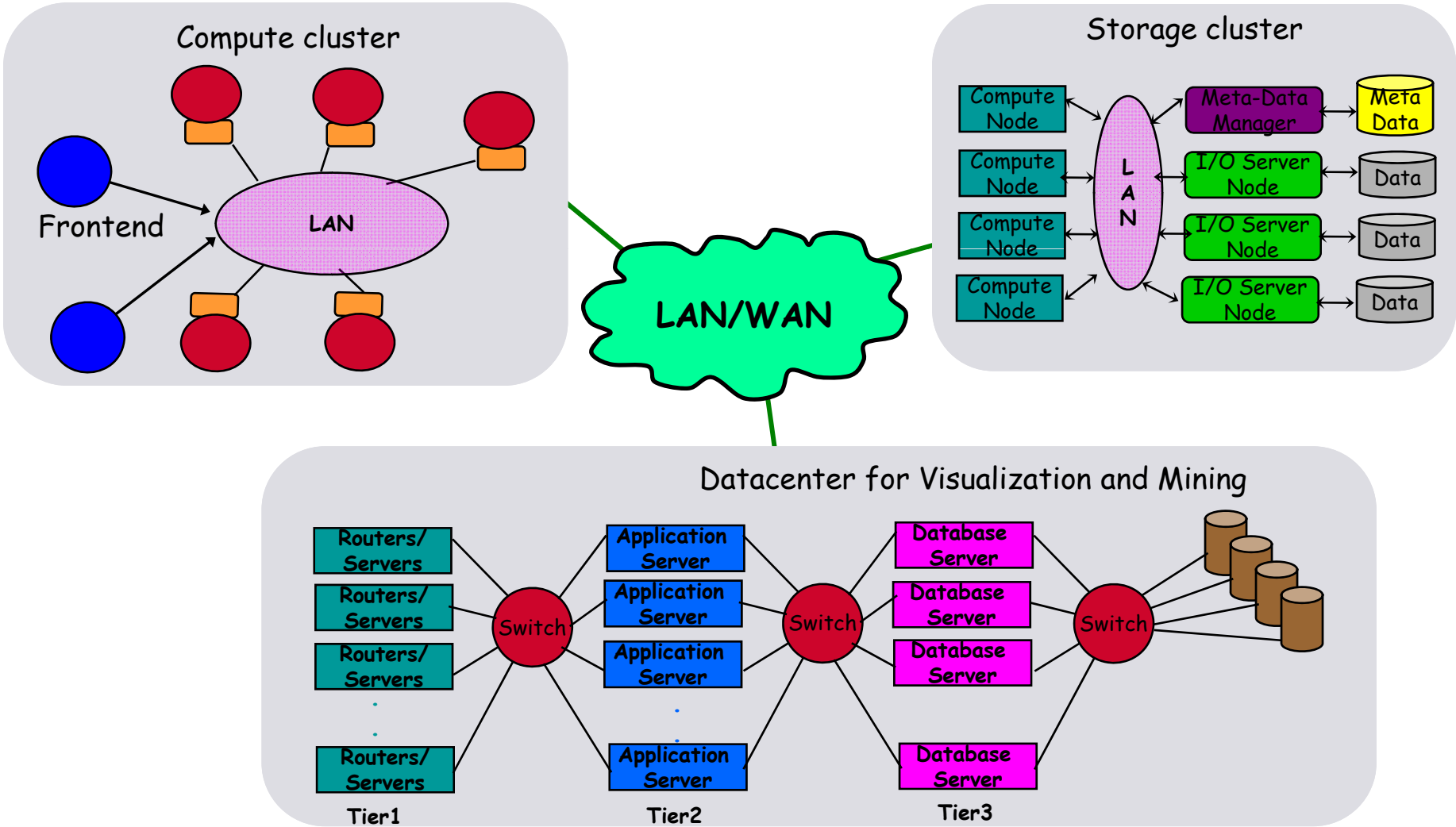


Trends for Computing Clusters in the Top 500 List

- Top 500 list of Supercomputers (www.top500.org)

June 2001: 33/500 (6.6%)	June 2005: 304/500 (60.8%)
Nov 2001: 43/500 (8.6%)	Nov 2005: 360/500 (72.0%)
June 2002: 80/500 (16%)	June 2006: 364/500 (72.8%)
Nov 2002: 93/500 (18.6%)	Nov 2006: 361/500 (72.2%)
June 2003: 149/500 (29.8%)	June 2007: 373/500 (74.6%)
Nov 2003: 208/500 (41.6%)	Nov 2007: 406/500 (81.2%)
June 2004: 291/500 (58.2%)	June 2008: 400/500 (80.0%)
Nov 2004: 294/500 (58.8%)	Nov 2008: To be Announced

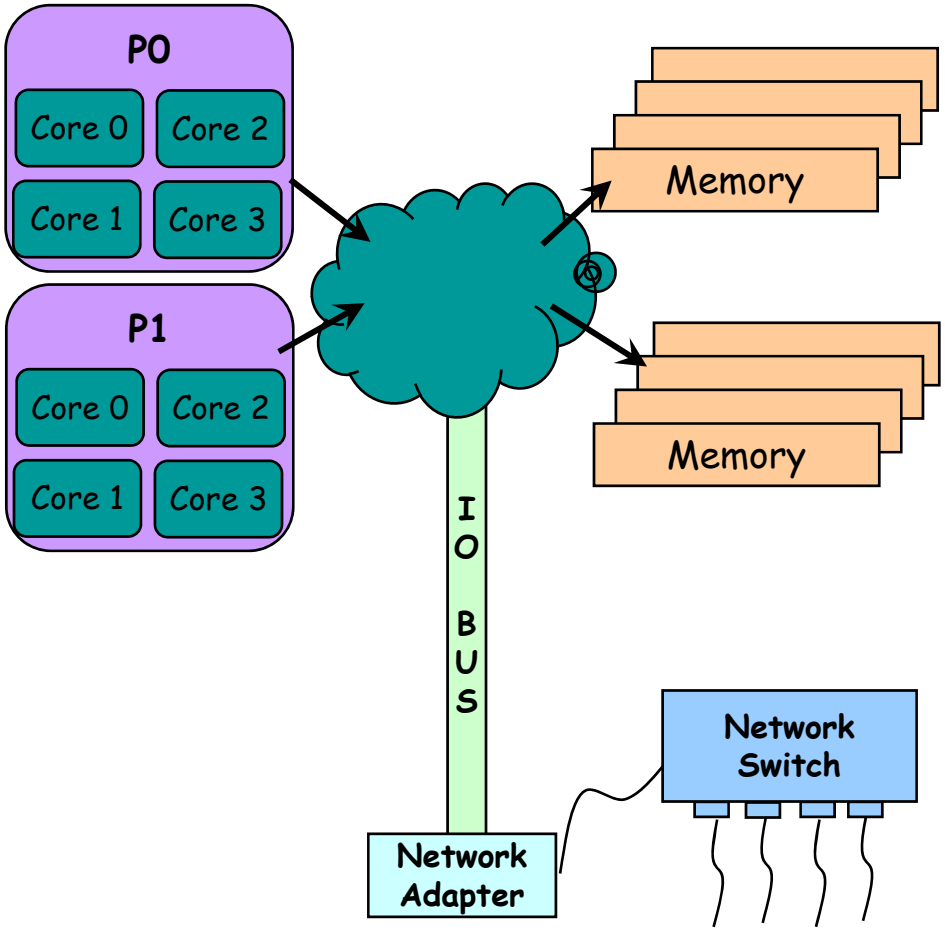
Integrated Environment with Multiple Clusters



Networking and I/O Requirements

- Good Systems Area Network with excellent performance (low latency and high bandwidth) for inter-processor communication (IPC) and I/O
- Good Storage Area Networks high performance I/O
- Good WAN connectivity in addition to intra-cluster SAN/LAN connectivity
- Quality of Service (QoS) for interactive applications
- RAS (Reliability, Availability, and Serviceability)
- With low cost

Major Components in Computing Systems



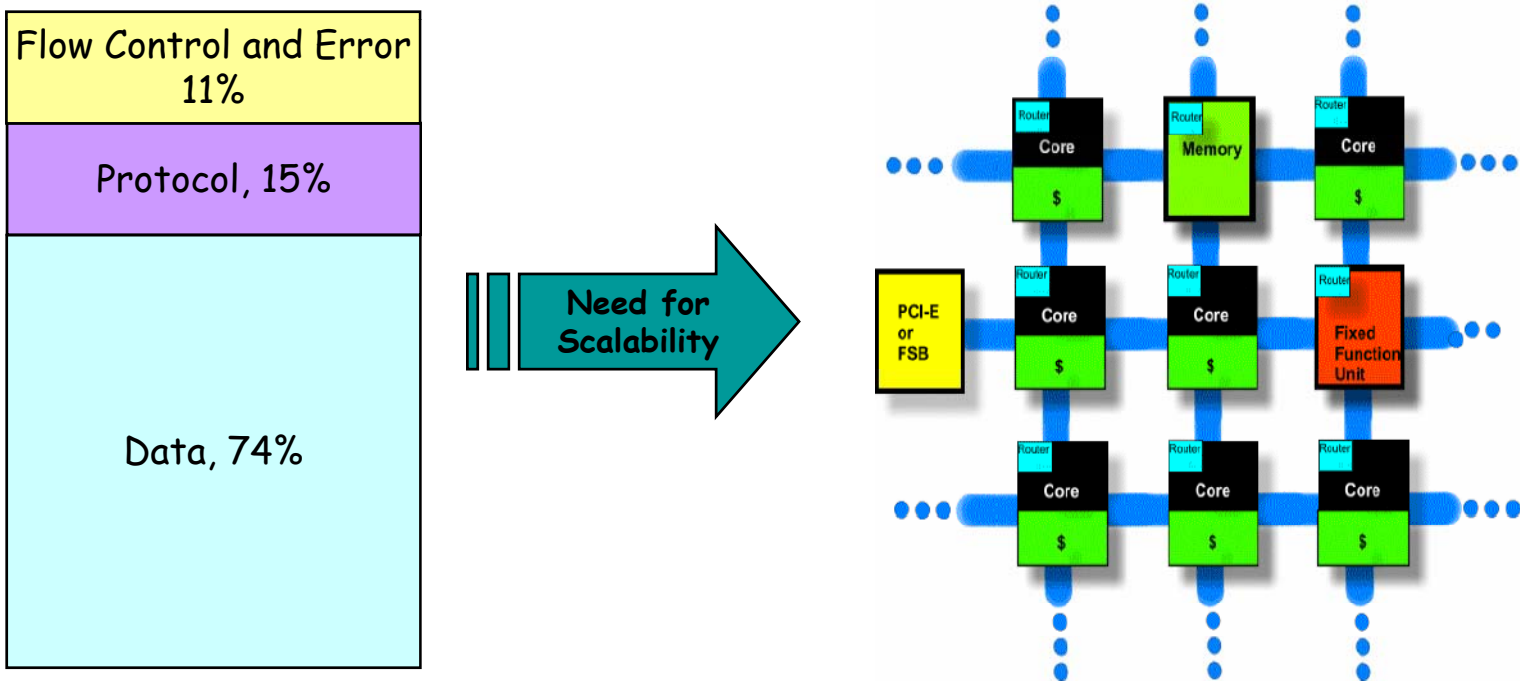
- Hardware Components
  - Processing Core and Memory sub-system
  - I/O Bus (PCI-X, PCIe, HT)
  - Network Adapter (InfiniBand, 10GE)
  - Network Switch (InfiniBand, 10GE)
- Software Components
  - Communication software

Processing Units

- Multi-processor systems have existed for many years
- Multi-core processors have also started coming into the market
- Quad-core processors are considered "commodity"
- Many-core processors upcoming
  - Intel planning to release an 80-core processor by 2011

Network-on-Chip Architectures

- Massive scale multi-cores rely on on-chip networks
  - 64-byte MTU; wormhole routing; end-to-end error control
  - Message passing on the chip



Courtesy Partha Kundu (Intel Corporation)

Trends in I/O Interfaces with Servers

- Network performance depends on
  - Networking technology (adapter + switch)
  - Network interface (last mile bottleneck)

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)
HyperTransport (HT) by AMD	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)
PCI-Express (PCIe) by Intel	2003 (Gen1) 2007 (Gen2)	Gen1: 4X (8Gbps), 8X (16Gbps), 16X (32Gbps) Gen2: 4X (16Gbps), 8X (32Gbps), 16X (64Gbps)
PCIe Gen3	Upcoming (due in 2009)	4X (32Gbps), 8X (64Gbps), 16X (128Gbps)
Intel QuickPath	Upcoming	192-256Gbps per link

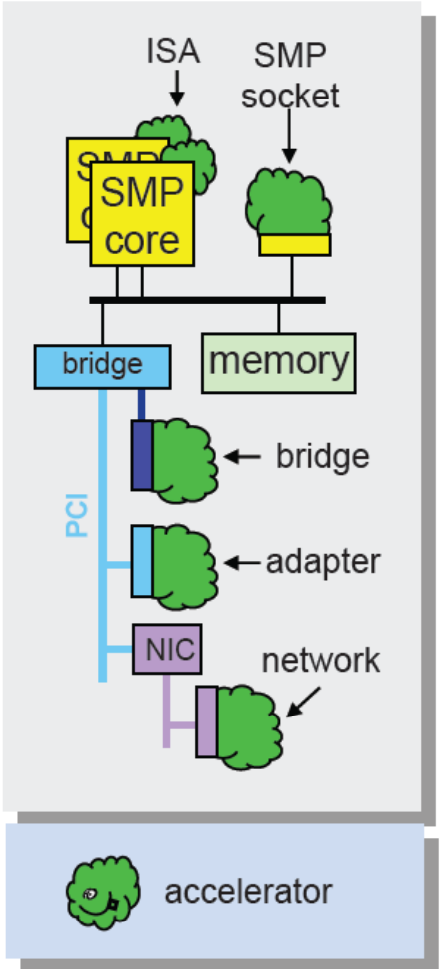
Upcoming I/O Interface Technologies

- PCI-Express Gen2 and Gen3
  - Founded by the PCI-SIG (with over 900 member organizations)
  - Gen2: 5Gbps signaling rate per lane → 4Gbps data rate
  - Gen3: 10Gbps signaling rate per lane → 8Gbps data rate
  - Standards available since Jan and Aug 2007
- HyperTransport 3.1
  - Founded by AMD, Apple, Cisco, and others
  - Increases clock speed (3.2GHz) and bandwidth (409.6Gbps per link)
  - Hot-plugging capability; Power Management Enhancements
- Intel QuickPath Interconnect (expected in late 2008)
  - Dedicated connectivity to CPUs and I/O devices like HT
  - Supporting 192 to 256Gbps bandwidth per link

Accelerators in High-end Computing

Tighter coupling
Lower overhead
Faster synchro.
Integrated SW
Broader use

Narrower use
Slower synchro.
Higher overhead
Loose coupling



High dev expense
Arch. dependent
Longer lead times
High SW expense
3rd party⇒lower RAS
Few entrants
(AMD: 4; Intel: 2)

Many entrants
Lower SW exp
Fast time-to-mkt
Arch. independent
Low dev expense

(memory attach, too – proposals)

(Courtesy Gregory Pfister, IBM)

Accelerator Classes

- Accelerators in every possible location on the system
  - On-par with the CPU (plugged in to the CPU socket)
    - DRC FPGA on the Hyper-transport for Cray XT5 family
    - STI Cell (PPE and SPEs connected with a dedicated bus)
  - Connected to the I/O bus
    - Network Traffic Accelerators (IB, 10GE family)
    - Computational libraries (ClearSpeed accelerator)
    - Graphics Accelerators (NVIDIA TESLA, AMD Firestream)
  - On memory devices
    - Research directions going on

Multi-core Processors vs. Accelerators

- Multiple cores on each processor
  - General purpose CPUs are becoming cheaper
- Will eventually outperform an accelerator
  - Becoming more difficult as CPU capability is increasing in parallelism, not speed
  - Accelerators themselves are becoming parallel too
  - By the time the CPU outperforms an accelerator, even more powerful accelerators will be available
- Accelerators and CPUs can be complementary!
  - Systems software need to be optimized to take advantage of both simultaneously

Presentation Overview

- Introduction
- Why InfiniBand and 10-Gigabit Ethernet?
- Overview of IB and 10GE and their Novel Features
- IB and 10GE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
  - MPI, SDP, File Systems, Data Center and Virtualization
- Conclusions and Final Q&A

Growth in Commodity Network Technology

Representative commodity networks; their entries into the market

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)

16 times in the last 7 years

Limitations of Traditional Host-based Protocols

- Ex: TCP/IP, UDP/IP
- Generic architecture for all network interfaces
- Host-handles almost all aspects of communication
  - Data buffering (copies on sender and receiver)
  - Data integrity (checksum)
  - Routing aspects (IP routing)
- Signaling between different layers
  - Hardware interrupt whenever a packet arrives or is sent
  - Software signals between different layers to handle protocol processing in different priority levels

Capabilities of Current High-Performance Networks

- Intelligent Network Interface Cards
- Support entire protocol processing completely in hardware (hardware protocol offload engines)
- Provide a rich communication interface to applications
  - ***User-level communication capability***
  - Gets rid of intermediate data buffering requirements
- No software signaling between communication layers
  - All layers are implemented on a ***dedicated*** hardware unit, and not on a ***shared*** host CPU

Previous High Performance Network Stacks

- Virtual Interface Architecture
  - Standardized by Intel, Compaq, Microsoft
- Fast Messages (FM)
  - Developed by UIUC
- Myricom GM
  - Proprietary protocol stack from Myricom
- These network stacks set the trend for high-performance communication requirements
  - Hardware offloaded protocol stack
  - Support for fast and secure user-level access to the protocol stack

IB Trade Association

- IB Trade Association was formed with seven industry leaders (Compaq, Dell, HP, IBM, Intel, Microsoft, and Sun)
- Goal: To design a scalable and high performance communication and I/O architecture by taking an integrated view of computing, networking, and storage technologies
- Many other industry participated in the effort to define the IB architecture specification
- IB Architecture (Volume 1, Version 1.0) was released to public on Oct 24, 2000
  - Latest version 1.2.1 released January 2008
- http://www.infinibandta.org

IB Hardware Acceleration

- Some IB models have multiple hardware acceleration units
  - E.g., Mellanox IB adapters
- Protocol Offload Engines
  - Completely implement layers 2-4 in hardware
- Additional hardware supported features also present
  - RDMA, Multicast, QoS, Network Fault Tolerance, and many more

10-Gigabit Ethernet Consortium

- 10GE Alliance formed by several industry leaders to take the Ethernet family to the next speed step
- Goal: To achieve a scalable and high performance communication architecture while maintaining backward compatibility with Ethernet
- http://www.ethernetalliance.org
- Upcoming 40-Gbps (Servers) and 100-Gbps Ethernet (Backbones, Switches, Routers): IEEE 802.3 WG
- Energy-efficient and power-conscious protocols
  - On-the-fly link speed reduction for under-utilized links

Ethernet Hardware Acceleration

- Interrupt Coalescing
  - Improves throughput, but degrades latency
- Jumbo Frames
  - No latency impact; Incompatible with existing switches
- Hardware Checksum Engines
  - Checksum performed in hardware → significantly faster
  - Shown to have minimal benefit independently
- Segmentation Offload Engines
  - Supported by most 10GE products because of its backward compatibility → considered "regular" Ethernet
  - Heavily used in the "server-on-steroids" model

TOE and iWARP Accelerators

- TCP Offload Engines (TOE)
  - Hardware Acceleration for the entire TCP/IP stack
  - Initially patented by Tehuti Networks
  - Actually refers to the IC on the network adapter that implements TCP/IP
  - In practice, it is usually referred to as the entire network adapter
- Internet Wide-Area RDMA Protocol (iWARP)
  - Standardized by IETF and the RDMA Consortium
  - Support acceleration features (like IB) for Ethernet
- http://www.ietf.org & http://www.rdmaconsortium.org

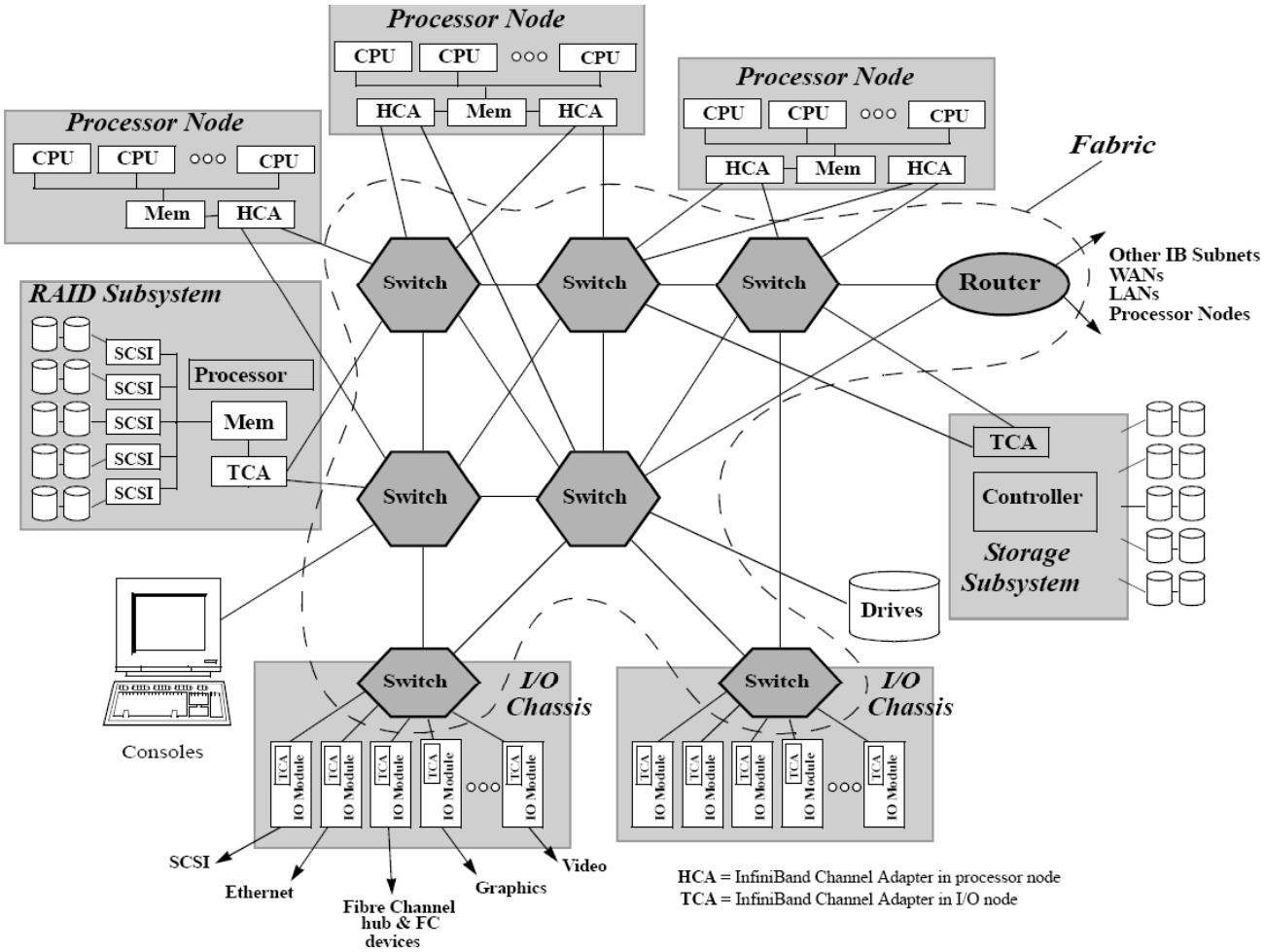
Presentation Overview

- Introduction
- Why InfiniBand and 10-Gigabit Ethernet?
- Overview of IB and 10GE and their Novel Features
- IB and 10GE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
  - MPI, SDP, File Systems, Data Center and Virtualization
- Conclusions and Final Q&A

IB and 10GE Overview

- InfiniBand
  - Architecture and Basic Hardware Components
  - Novel Features
    - Hardware Protocol Offload
      - Link, network and transport layer features
    - Communication Semantics
      - Memory registration and protection
      - Channel and memory semantics
  - IB Verbs Interface
  - Management and Services
    - Subnet Management
    - Hardware support for scalable network management

A Typical IB Network



Three primary components

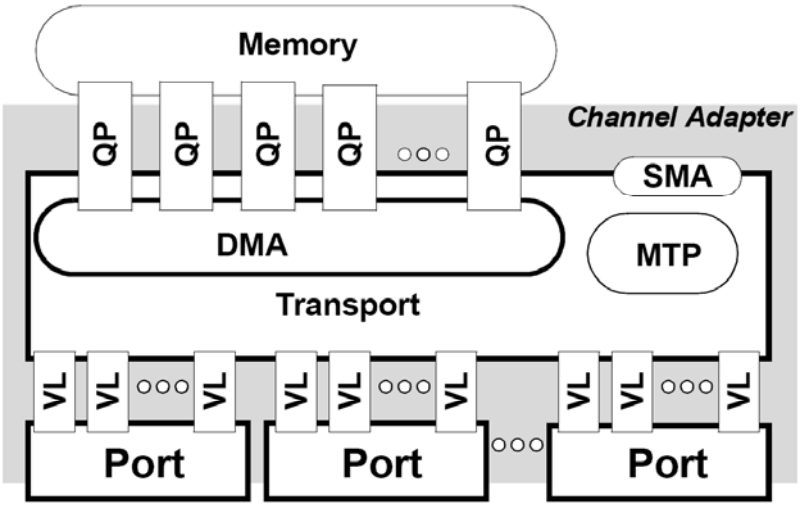
Channel Adapters

Switches/Routers

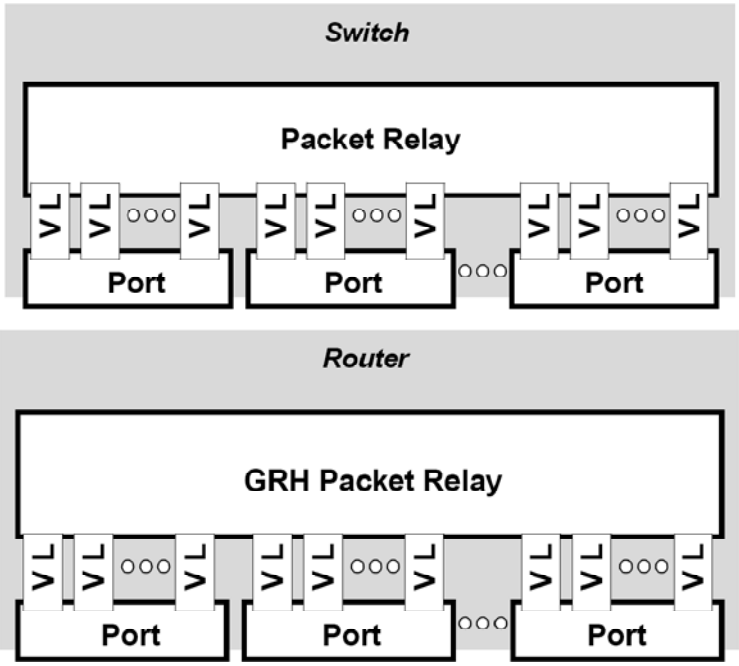
Links and connectors

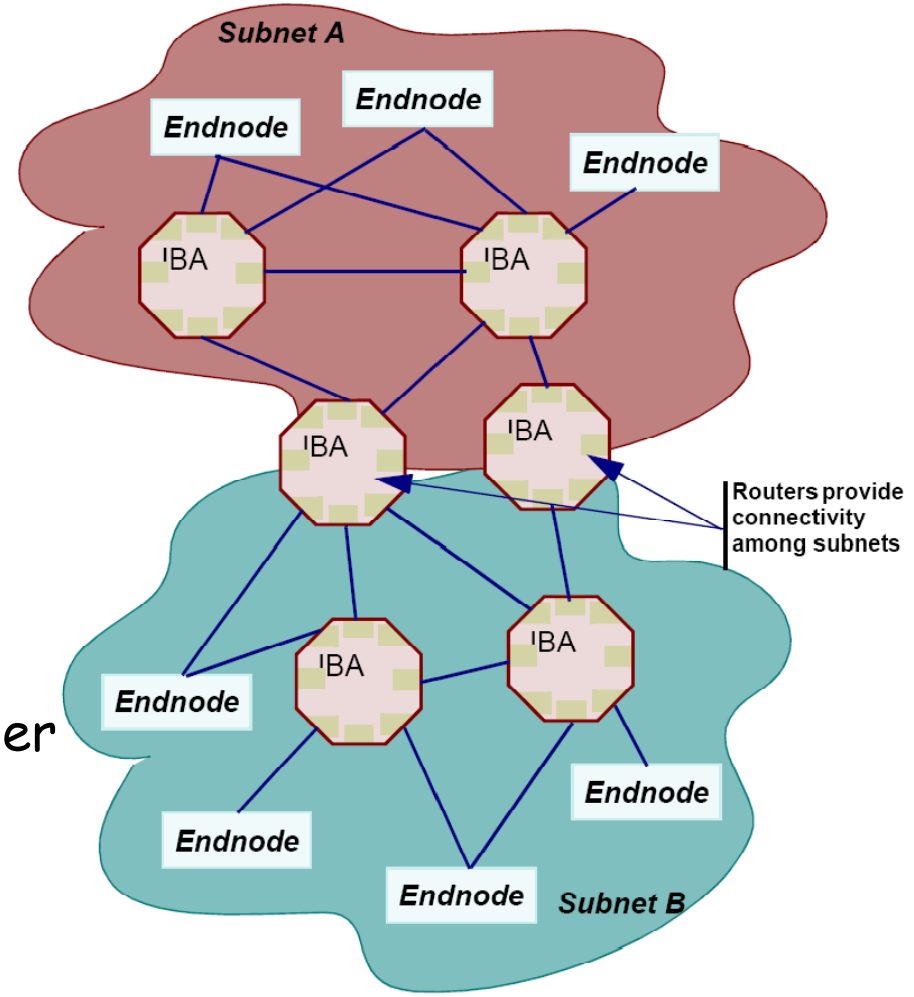
Components: Channel Adapters

- Used by processing and I/O units to connect to fabric
- Consume & generate IB packets
- Programmable DMA engines with protection features
- May have multiple ports
  - Independent buffering channeled through Virtual Lanes
- Host Channel Adapters (HCAs)



Components: Switches and Routers





- Relay packets from a link to another
- Switches: intra-subnet
- Routers: inter-subnet
- May support multicast

Components: Links & Repeaters

- Network Links
  - Copper, Optical, Printed Circuit wiring on Back Plane
  - Not directly addressable
- Traditional adapters built for copper cabling
  - Restricted by cable length (signal integrity)
- Intel Connects: Optical cables with Copper-to-optical conversion hubs (acquired by Emcore)
  - Up to 100m length
  - 550 picoseconds

    copper-to-optical conversion latency
- Available from other vendors (Luxtera)
- Repeaters (Vol. 2 of InfiniBand specification)

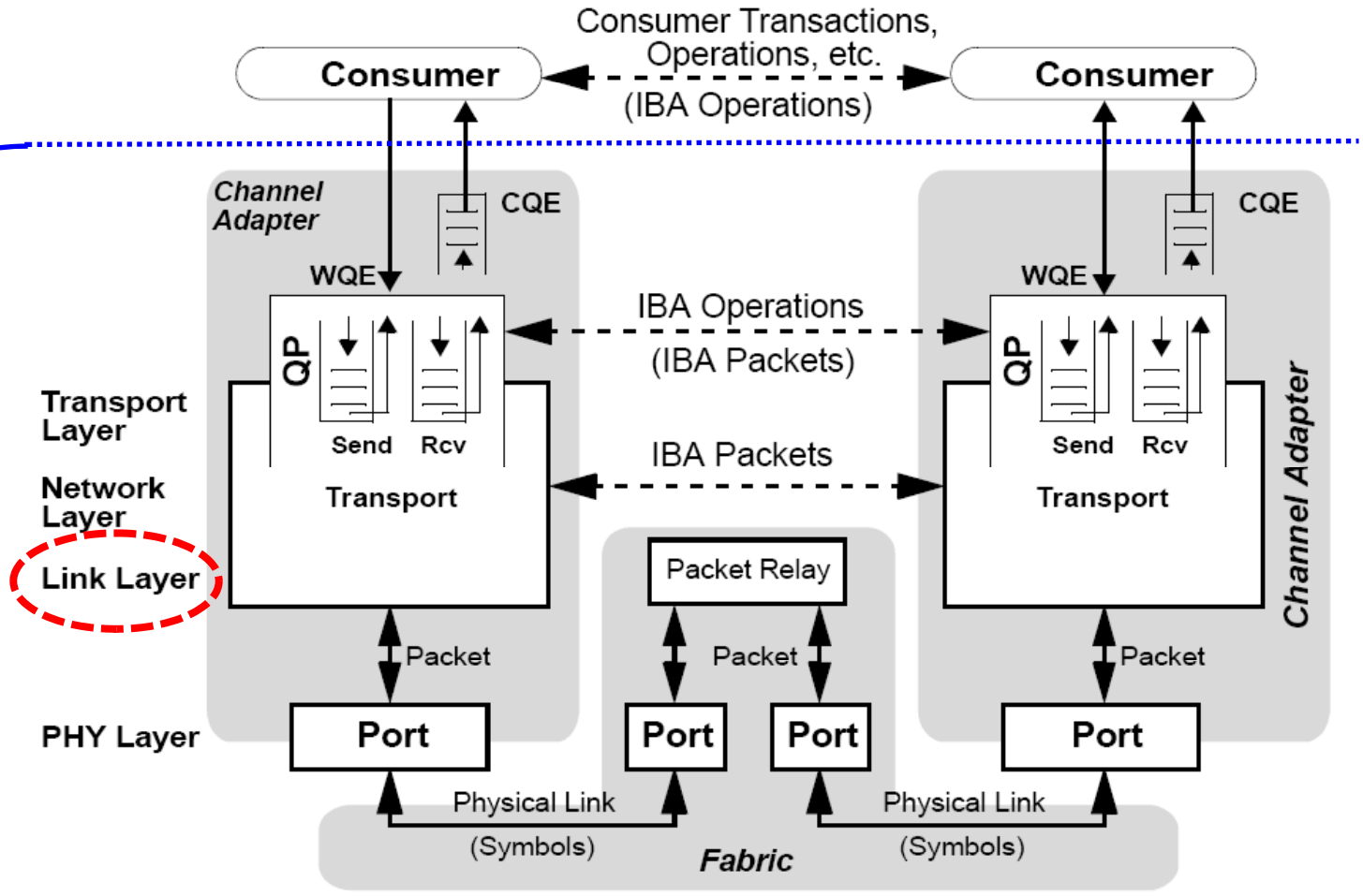
(Courtesy Intel)

IB and 10GE Overview

- InfiniBand
  - Architecture and Basic Hardware Components
  - Novel Features
    - Hardware Protocol Offload
      - Link, network and transport layer features
    - Communication Semantics
      - Memory registration and protection
      - Channel and memory semantics
  - IB Verbs Interface
  - Management and Services
    - Subnet Management
    - Hardware support for scalable network management

Hardware Protocol Offload

Complete Hardware Implementations Exist



Link Layer Capabilities

- CRC-based Data Integrity
- Buffering and Flow Control
- Virtual Lanes, Service Levels and QoS
- Switching and Multicast
- IB WAN Capability

CRC-based Data Integrity

- Two forms of CRC to achieve both early error detection and end-to-end reliability
  - Invariant CRC (ICRC) covers fields that do not change per link (per network hop)
    - E.g., routing headers (if there are no routers), transport headers, data payload
    - 32-bit CRC (compatible with Ethernet CRC)
    - End-to-end reliability (does not include I/O bus)
  - Variant CRC (VCRC) covers everything
    - Erroneous packets do not have to reach the destination before being discarded
    - Early error detection

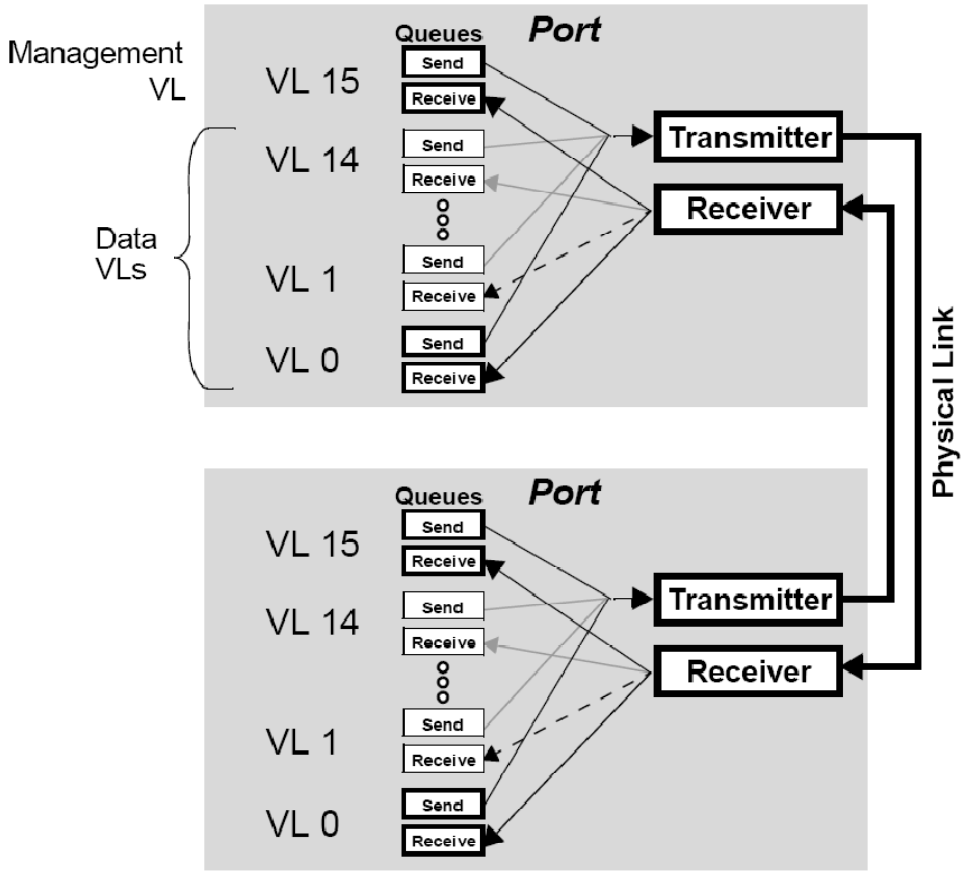
Buffering and Flow Control

- IB provides an absolute credit-based flow-control
  - Receiver guarantees that it has enough space allotted for N blocks of data
  - Occasional update of available credits by the receiver
- Has no relation to the number of messages, but only to the total amount of data being sent
  - One 1MB message is equivalent to 1024 1KB messages (except for rounding off at message boundaries)

Link Layer Capabilities

- CRC-based Data Integrity
- Buffering and Flow Control
- Virtual Lanes, Service Levels and QoS
- Switching and Multicast
- IB WAN Capability

Virtual Lanes



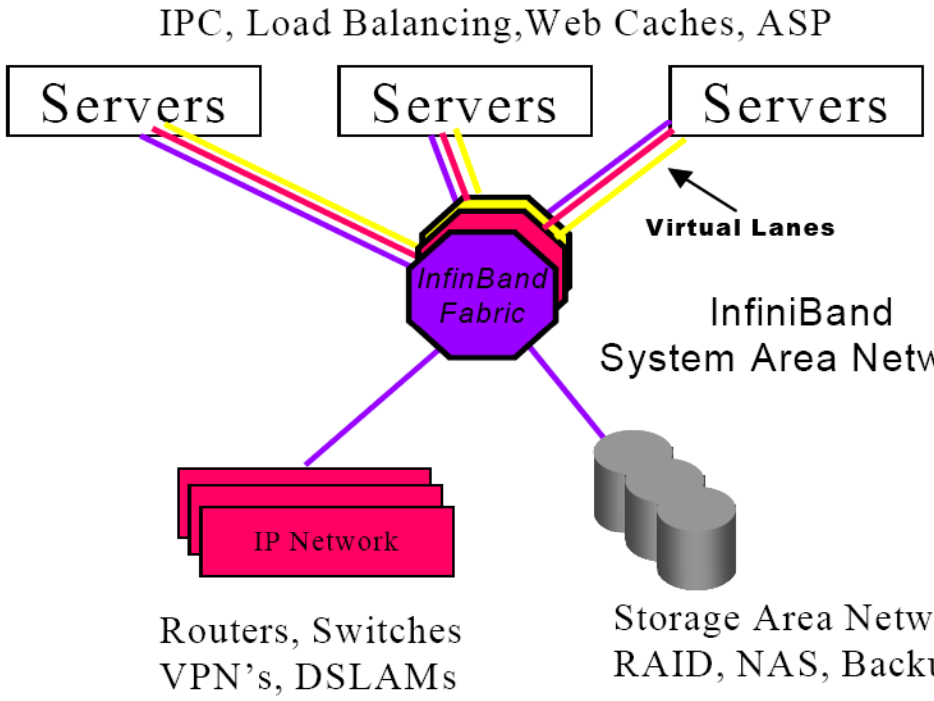
- Multiple virtual links within same physical link
  - Between 2 and 16
- Separate buffers and flow control
  - Avoids Head-of-Line Blocking
- VL15: reserved for management
- Each port supports one or more data VL

Service Levels and QoS

- Service Level (SL):
  - Packets may operate at one of 16 different SLs
  - Meaning not defined by IB
- SL to VL mapping:
  - SL determines which VL on the next link is to be used
  - Each port (switches, routers, end nodes) has a SL to VL mapping table configured by the subnet management
- Partitions:
  - Fabric administration (through Subnet Manager) may assign specific SLs to different partitions to isolate traffic flows

Traffic Segregation Benefits

- Segregation of Server, Network, and Storage Traffic - On the Same Physical Network



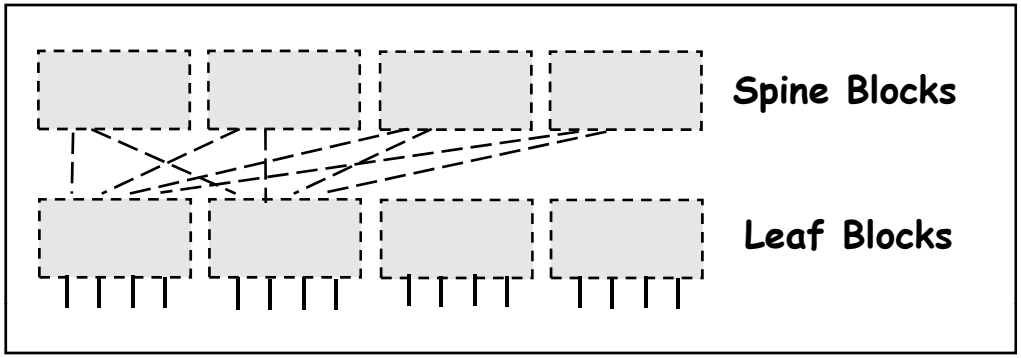
(Courtesy of Mellanox Technologies, Inc.)

- InfiniBand Virtual Lanes allow the multiplexing of multiple independent logical traffic flows on the same physical link.
- Providing the benefits of independent, separate networks while eliminating the cost and difficulties associated with maintaining two or more networks

Switching (Layer-2 Routing) and Multicast

- Each port has one or more associated LIDs (Local Identifiers)
  - Switches look up which port to forward a packet to based on its destination LID (DLID)
  - This information is maintained at the switch
- For multicast packets, the switch needs to maintain multiple output ports to forward the packet to
  - Packet is replicated to each appropriate output port
  - Ensures at-most once delivery & loop-free forwarding
  - There is an interface for a group management protocol
    - Create, join/leave, prune, delete group

Destination-based Switching/Routing



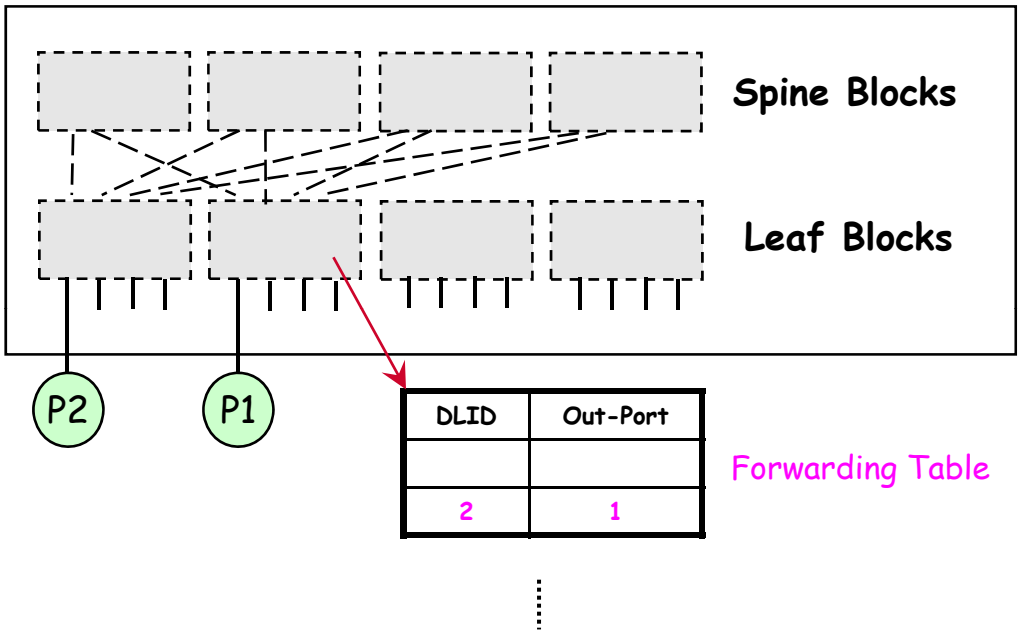
An Example IB Switch Block Diagram (Mellanox 144-Port)

Switching: IBA supports Virtual Cut Through (VCT)

Routing: Unspecified by IBA SPEC Up*/Down*, Shift are popular routing engines supported by OFED

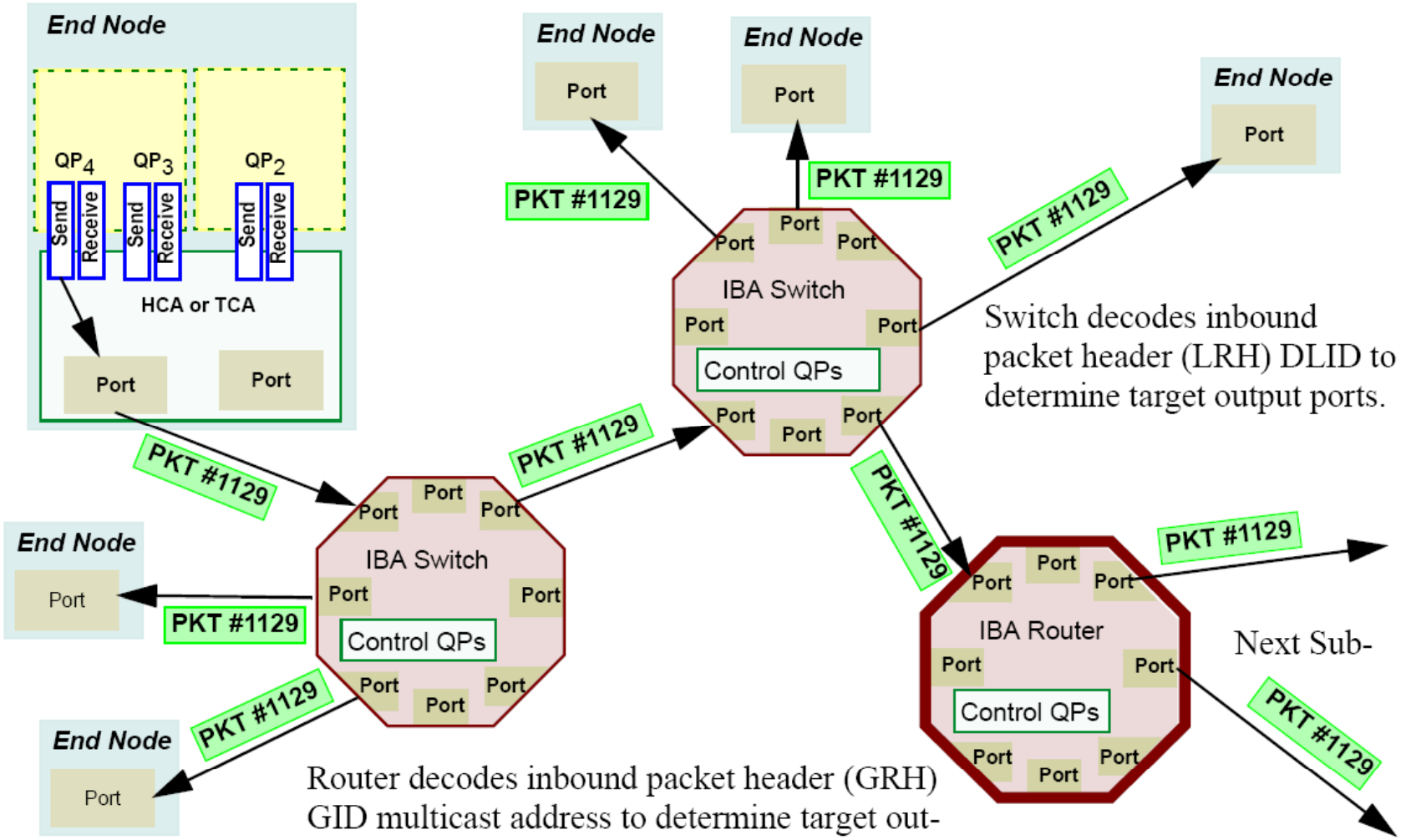
- Fat-Tree is a popular topology for IB Clusters
- Different over-subscription ratio may be used

IB Switching/Routing: An Example



- Someone has to setup these tables and give every port an LID
  - "Subnet Manager" does this work (more discussion on this later)
- Different routing algorithms may give different paths

IB Multicast Example

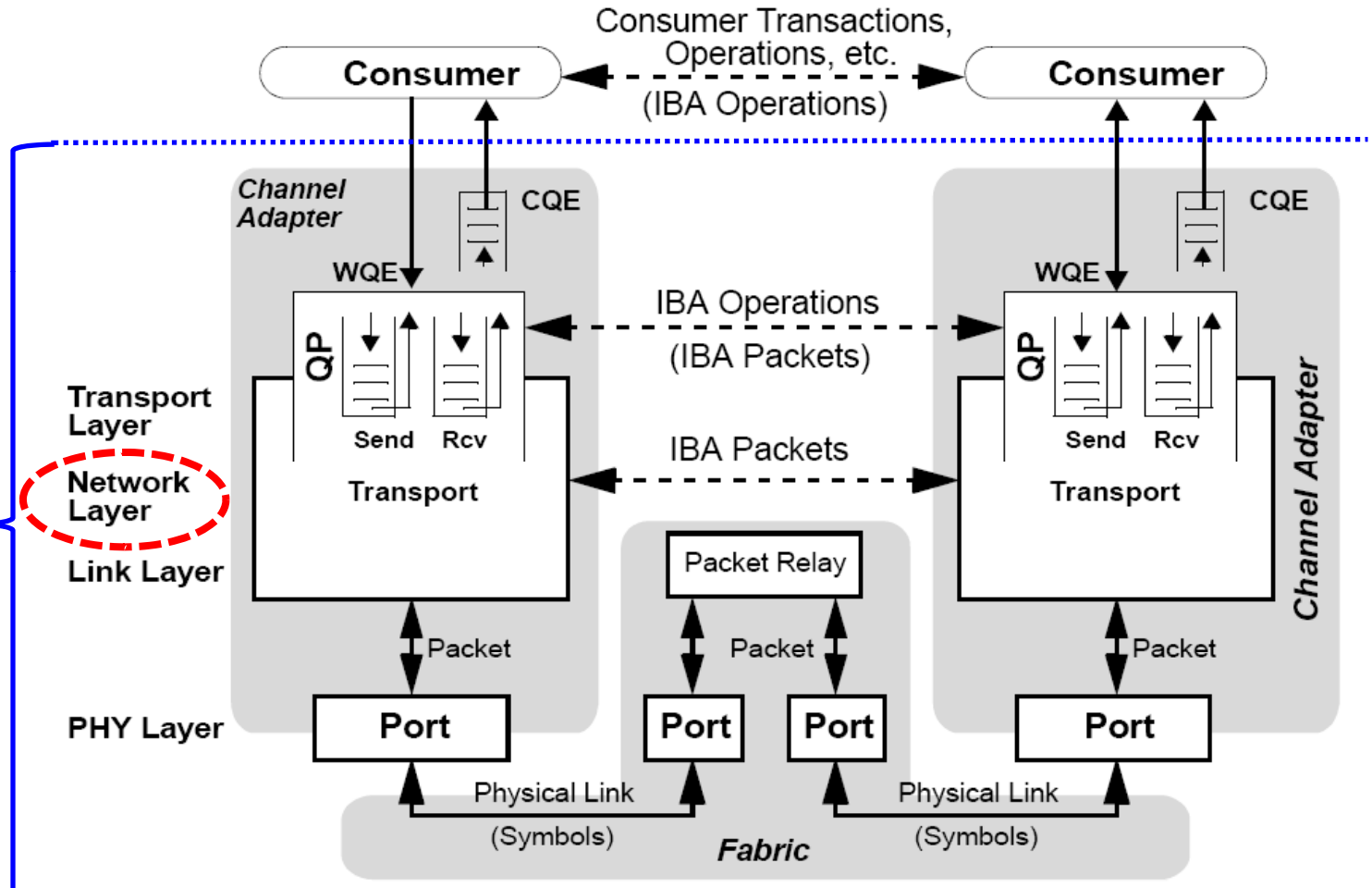


IB WAN Capability

- Getting increased attention for:
  - Remote Storage, Remote Visualization
  - Cluster Aggregation (Cluster-of-clusters)
- IB-Optical switches by multiple vendors
  - Obsidian Research Corporation: www.obsidianresearch.com
  - Network Equipment Technology (NET): www.net.com
  - Layer-1 changes from copper to optical; everything else stays the same
    - Low-latency copper-optical-copper conversion
- Large link-level buffers for flow-control
  - Data messages do not have to wait for round-trip hops
  - Important in the wide-area network

Hardware Protocol Offload

Complete Hardware Implementations Exist



IB Network Layer Capabilities

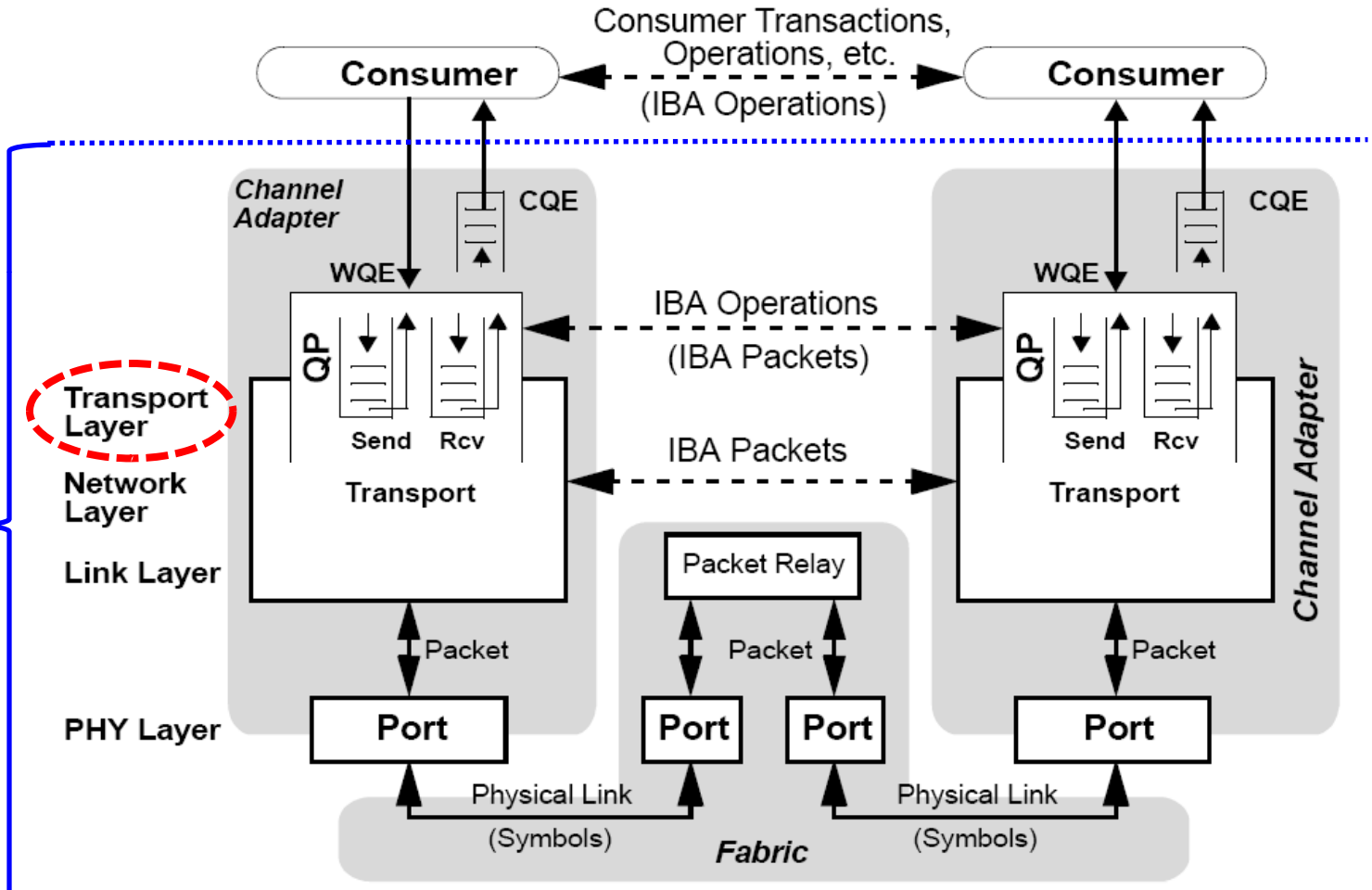
- Most capabilities are similar to that of the link layer, but as applied to IB routers
  - Routers can send packets across subnets (subnet are management domains, not administrative domains)
  - Subnet management packets are consumed by routers, not forwarded to the next subnet
- Several additional features as well
  - E.g., routing and flow labels

Routing and Flow Labels

- Routing follows the IPv6 specification
  - Easy interoperability with Wide-area translations
  - Link layer might still need to be translated to the appropriate layer-2 protocol (e.g., Ethernet, SONET)
- Flow Labels allow routers to specify which packets belong to the same connection
  - Switches can optimize communication by sending packets with the same label in order
  - Flow labels can change in the router, but packets belonging to one label, will always do so

Hardware Protocol Offload

Complete Hardware Implementations Exist



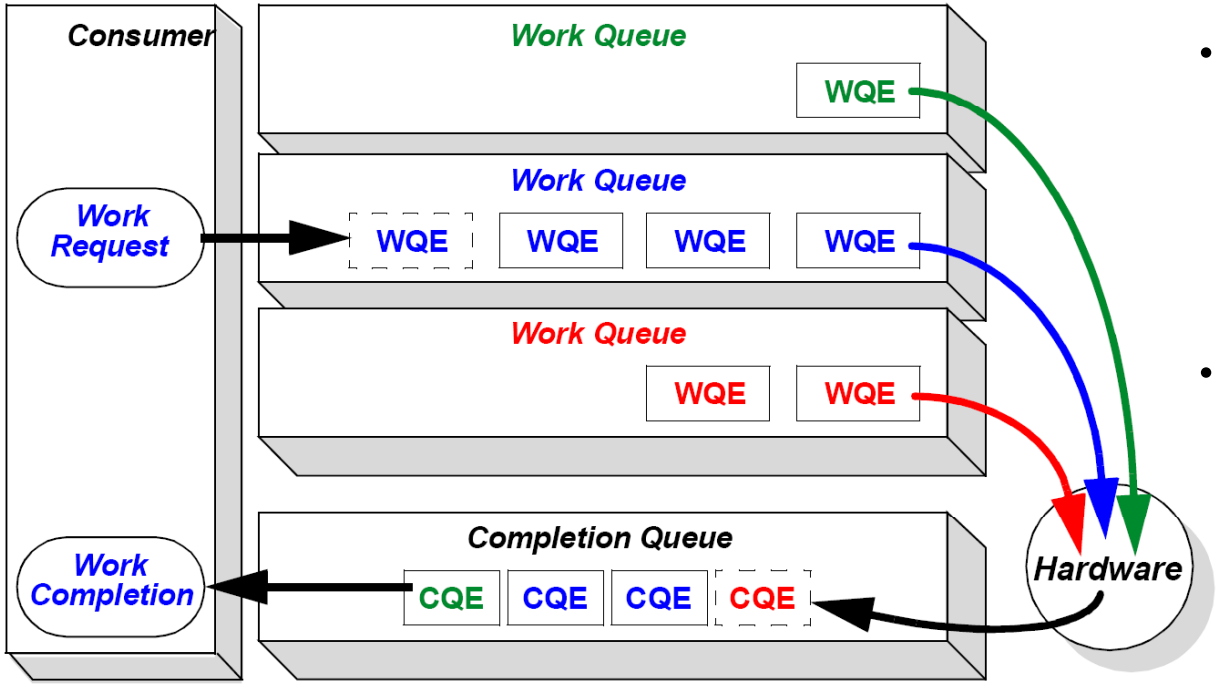
IB Transport Services

Service Type	Connection Oriented	Acknowledged	Transport
Reliable Connection	yes	Yes	IBA
Unreliable Connection	yes	no	IBA
Reliable Datagram	no	Yes	IBA
Unreliable Datagram	no	no	IBA
RAW Datagram	no	no	Raw

Interaction using QPs and CQs

- Interaction between the user and the network happens mainly through two entities:
  - Queue Pairs (QPs): Send queue + Receive queue
  - Completion Queues (CQs)
- Communication happens using data structures called Work Queue Requests (WQEs); called "Wookies"
- Completed WQEs are placed in the CQ with additional information
  - They are now called CQEs ("Cookies")

WQEs and CQEs



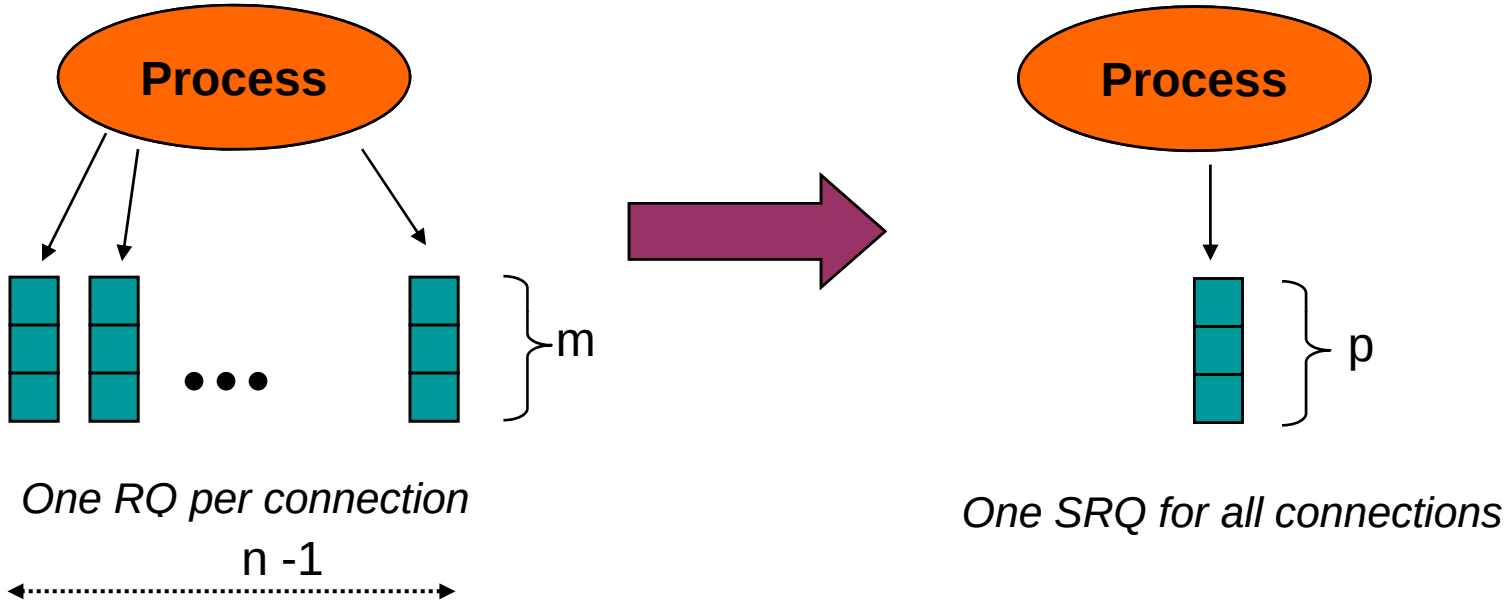
- Each transport service can have zero or more QPs associated with it
  - E.g., you can have 4 QPs based on RCs and one based on UD

- Send WQEs contain data about what buffer to send from, how much to send, etc.
- Receive WQEs contain data about what buffer to receive into, how much to receive, etc.
- CQEs contain data about which QP the completed WQE was posted on, how much data actually arrived

Trade-offs in Different Transport Types

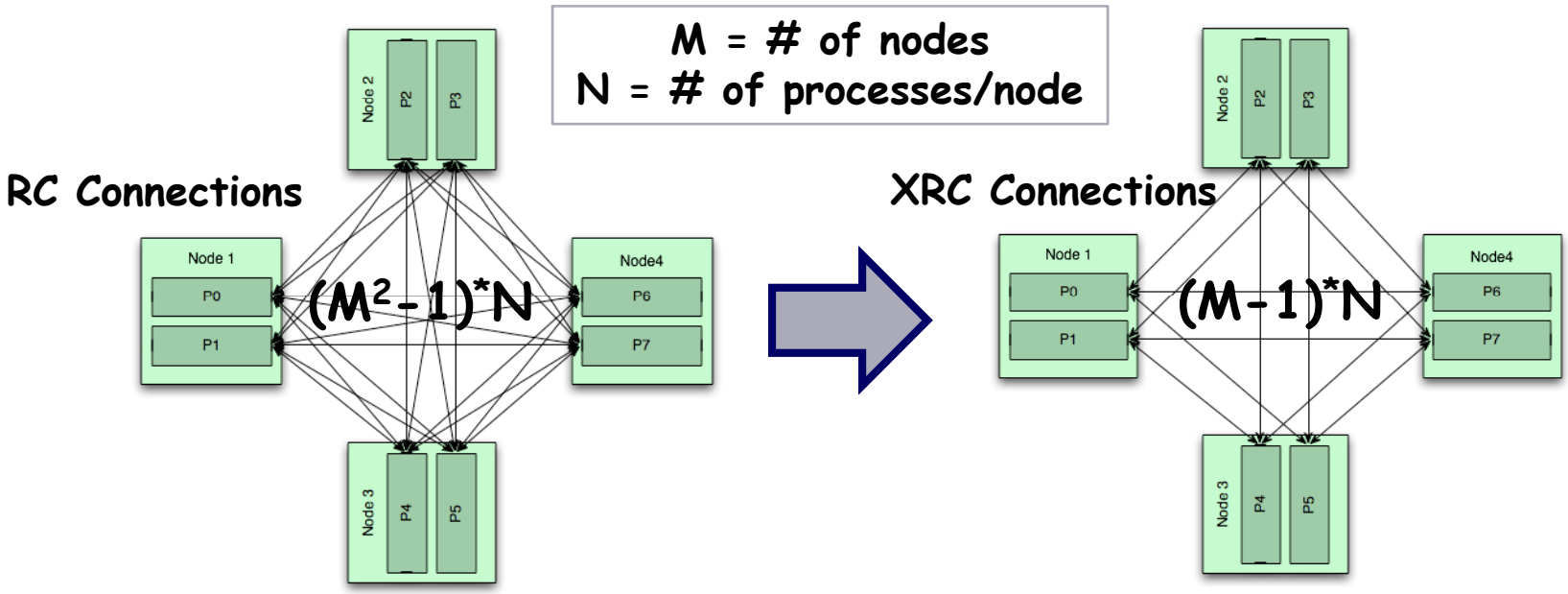
Attribute		Reliable Connection	Reliable Datagram	Unreliable Datagram	Unreliable Connection	Raw Datagram (both IPv6 & ethertype)
Scalability (M processes on N Processor nodes communicating with all processes on all nodes)		**M^2*N** QPs required on each processor node, per CA	**M** QPs required on each processor node, per CA.	**M** QPs required on each processor node, per CA.	**M^2*N** QPs required on each processor node, per CA.	**1** QP required on each end node, per CA.
Reliability	Corrupt data detected	Yes				
	Data delivery guarantee	Data delivered exactly once		No guarantees		
	Data order guaranteed	Yes, per connection	Yes, packets from any one source QP are ordered to multiple destination QPs.	No	Unordered and duplicate packets are detected.	No
	Data loss detected	Yes		No	Yes	No
	Error recovery	**Reliable**. Errors are detected at both the requestor and the responder. The requestor can transparently recover from errors (retransmission, alternate path, etc.) without any involvement of the client application. QP processing is halted only if the destination is inoperable or all fabric paths between the channel adapters have failed.		**Unreliable**. Packets with some types of errors may not be delivered. Neither source nor destination QPs are informed of dropped packets.	**Unreliable**. Packets with errors, including sequence errors, are detected and may be logged by the responder. The requestor is not informed.	**Unreliable**. Packets with errors are not delivered. The requestor and responder are not informed of dropped packets.

Shared Receive Queue (SRQ)



- SRQ is a hardware mechanism in IB by which a process can share receive resources (memory) across multiple connections
- A new feature, introduced in specification v1.2
- 0 < p << m*(n-1)

eXtended Reliable Connection (XRC)



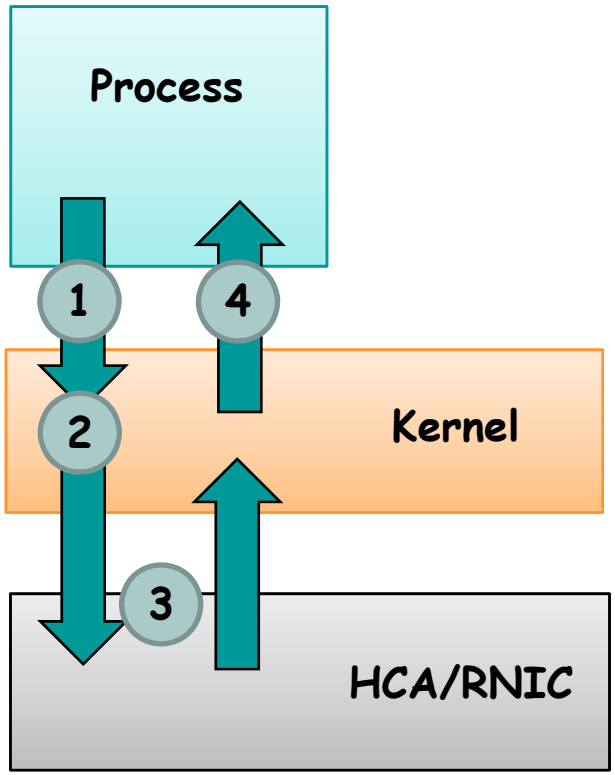
- Each QP takes at least one page of memory
  - Connections between all processes is very costly for RC
- New IB Transport added: eXtended Reliable Connection
  - Allows connections between nodes instead of processes

IB and 10GE Overview

- InfiniBand
  - Architecture and Basic Hardware Components
  - Novel Features
    - Hardware Protocol Offload
      - Link, network and transport layer features
    - Communication Semantics
      - Memory registration and protection
      - Channel and memory semantics
  - IB Verbs Interface
  - Management and Services
    - Subnet Management
    - Hardware support for scalable network management

Memory Registration

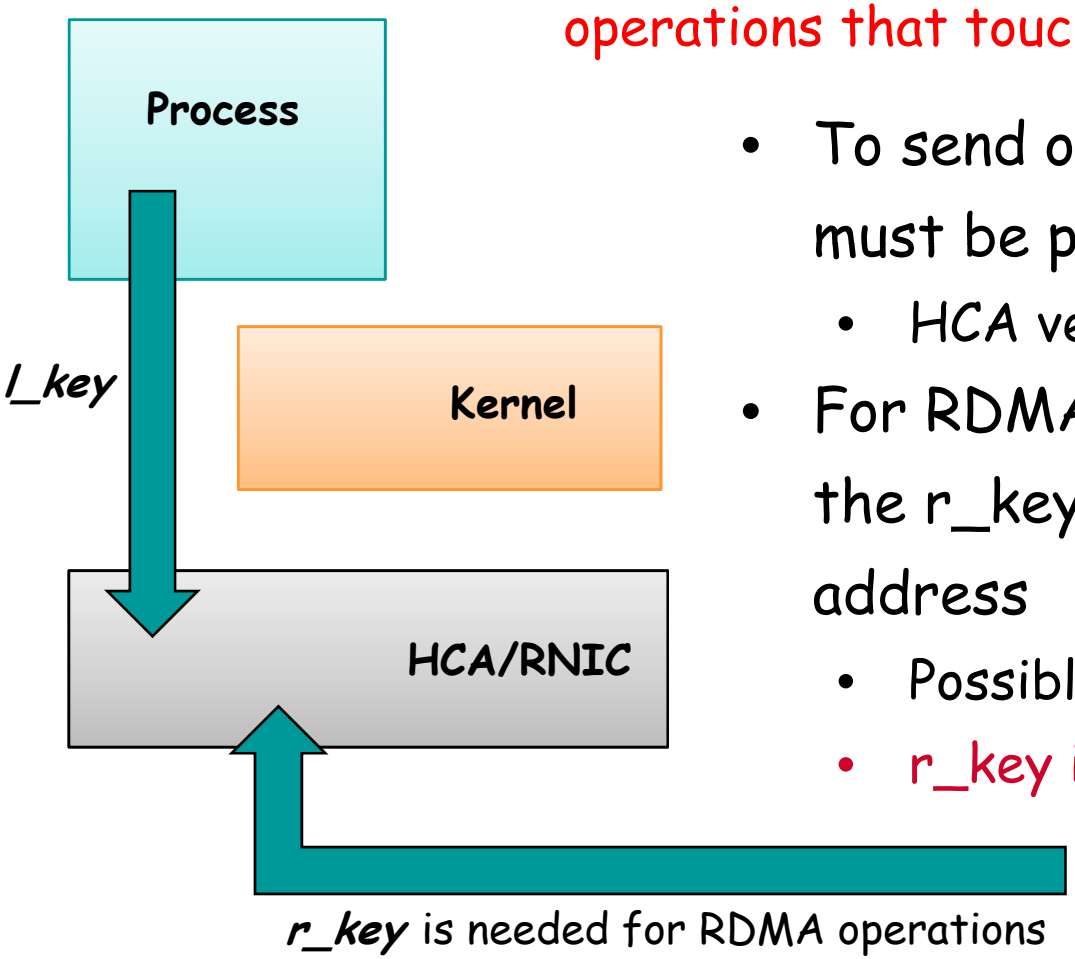
Before we do any communication:
All memory used for communication must be registered



1. Registration Request
   - Send virtual address and length
2. Kernel handles virtual->physical mapping and pins region into physical memory
   - Process cannot map memory that it does not own (security !)
3. HCA caches the virtual to physical mapping and issues a handle
   - Includes an *l_key* and *r_key*
4. Handle is returned to application

Memory Protections

For security, keys are required for all operations that touch buffers



- To send or receive data the *l_key* must be provided to the HCA
  - HCA verifies access to local memory
- For RDMA, the initiator must have the r_key for the remote virtual address
  - Possibly exchanged with a send/recv
  - r_key is not encrypted in IB

Communication in the Channel Semantics (Send-Receive Model)

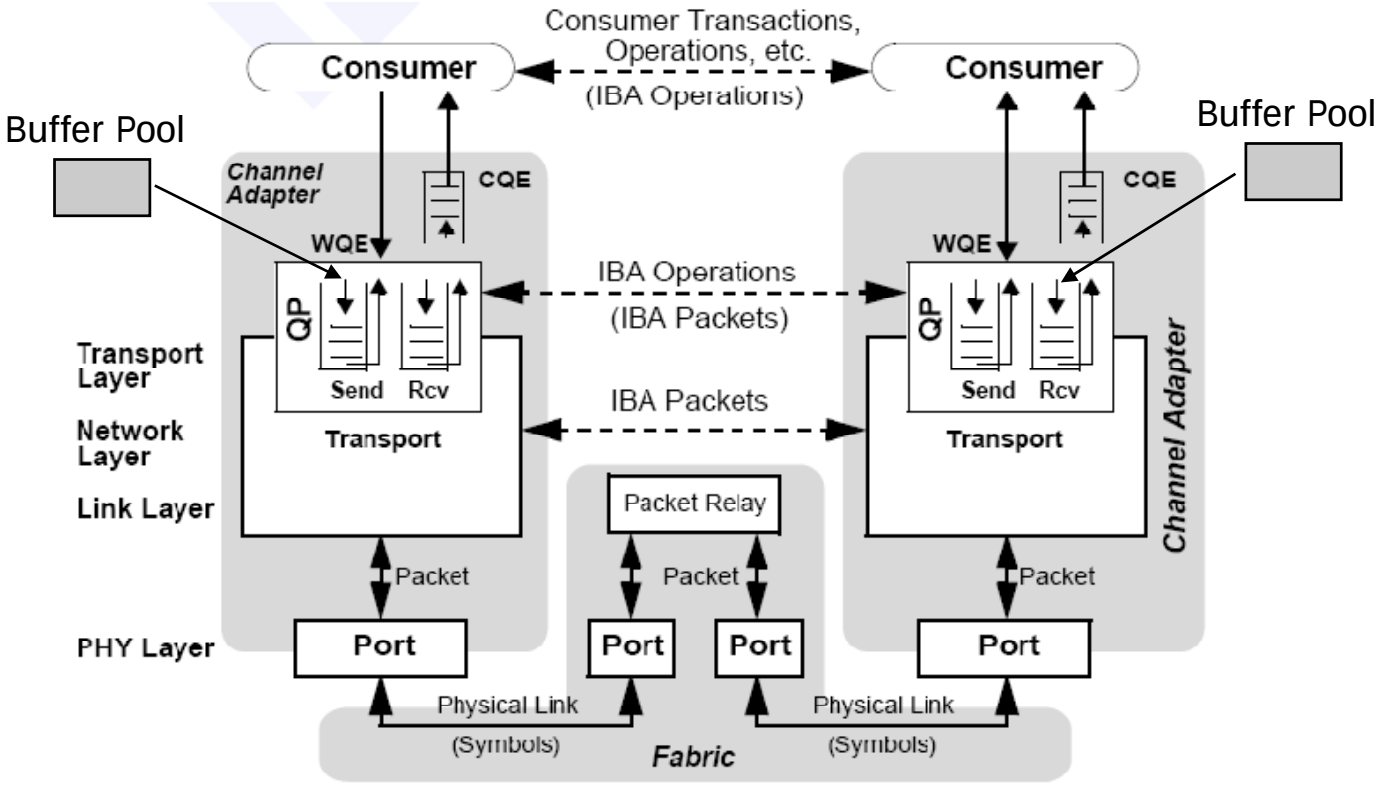
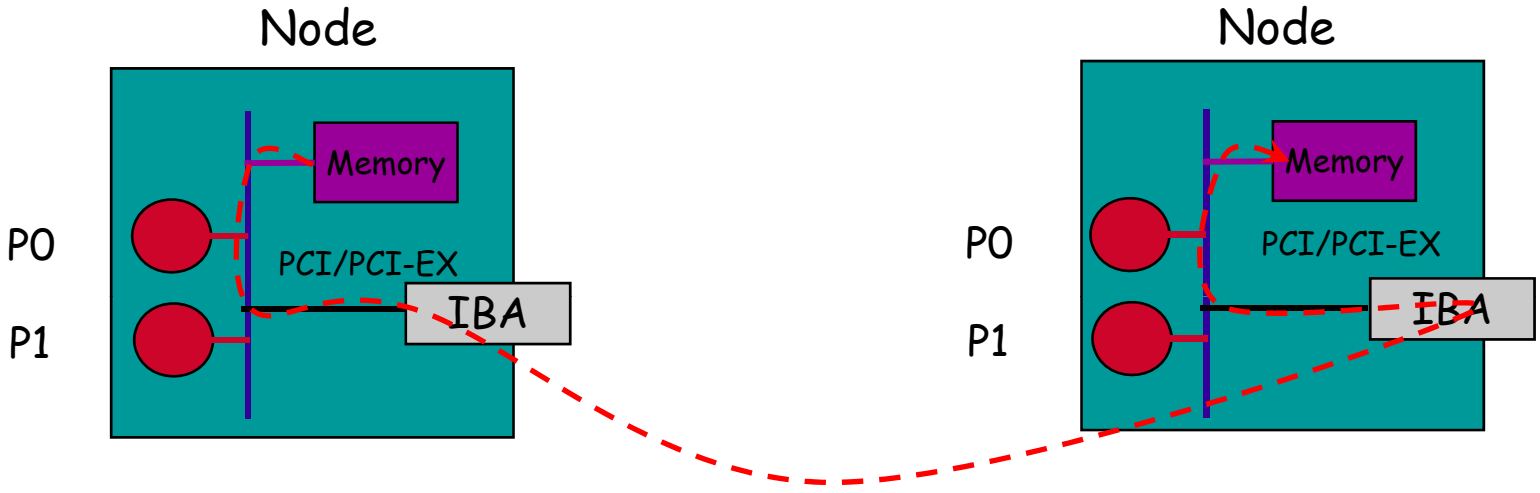


Figure 13 IBA Communication Stack

Communication in the Memory Semantics (RDMA Model)



- No involvement by the CPU at the receiver (RDMA Write/Put)
- No involvement by the CPU at the sender (RDMA Read/get)
- 1-2 μs latency (for short data)
- 1.5 – 2.6 GBps bandwidth (for large data)
- 3-5 μs for atomic operation

IB and 10GE Overview

- InfiniBand
  - Architecture and Basic Hardware Components
  - Novel Features
    - Hardware Protocol Offload
      - Link, network and transport layer features
    - Communication Semantics
      - Memory registration and protection
      - Channel and memory semantics
  - IB Verbs Interface
  - Management and Services
    - Subnet Management
    - Hardware support for scalable network management

IB Verbs Interface

- Different types of verbs
  - Transport Resource Management
    - HCA (Open, Close, Query, …)
    - Queue Pair (Create, Destroy, Modify, …)
    - Completion Queue (Create, Destroy, Modify, …)
    - Memory (Register, Deregister, …)
  - WQEs and CQEs
    - Post Send/Recv WQEs
    - Check CQ for CQEs
    - Request interrupt on completion of a WQE
  - Fabric Diagnostic Verbs
    - Check for devices on the network

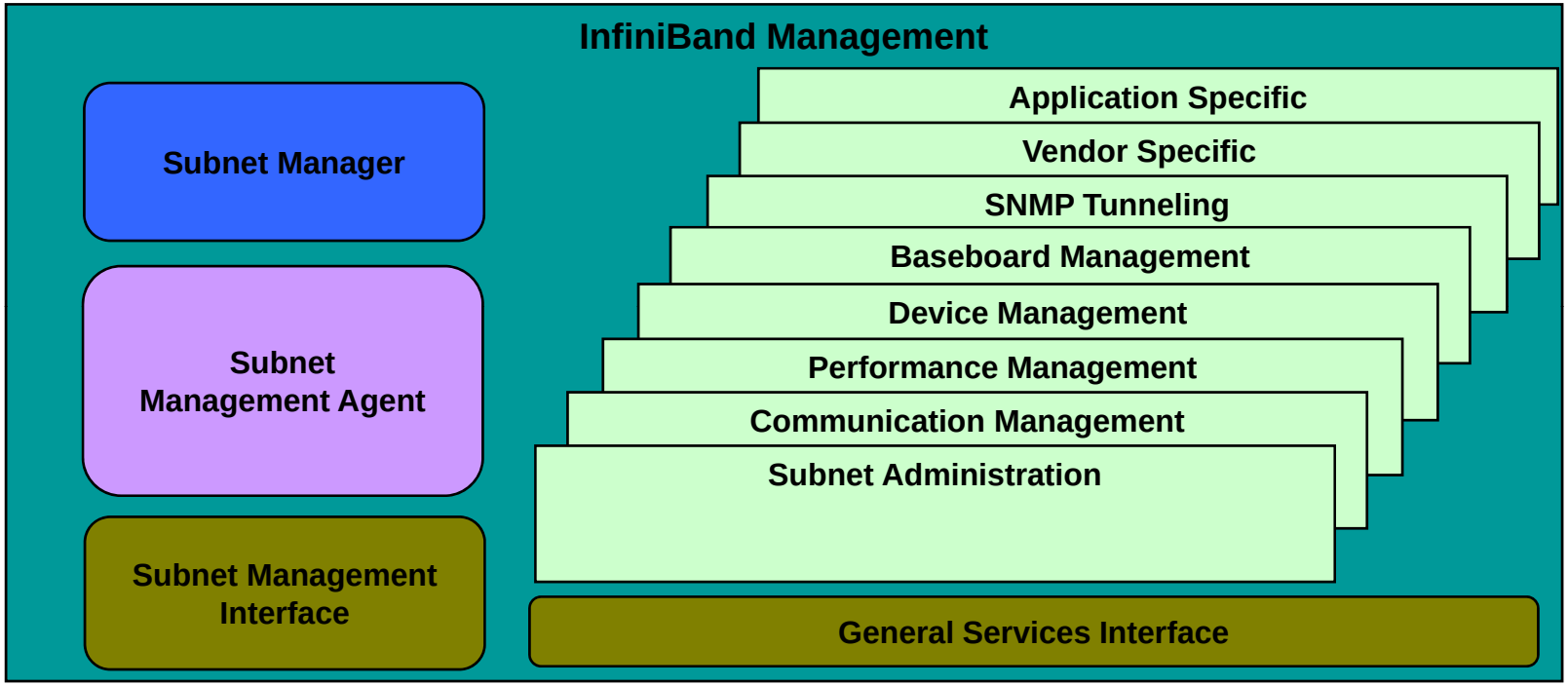
IB and 10GE Overview

- InfiniBand
  - Architecture and Basic Hardware Components
  - Novel Features
    - Hardware Protocol Offload
      - Link, network and transport layer features
    - Communication Semantics
      - Memory registration and protection
      - Channel and memory semantics
  - IB Verbs Interface
  - Management and Services
    - Subnet Management
    - Hardware support for scalable network management

Concepts in IB Management

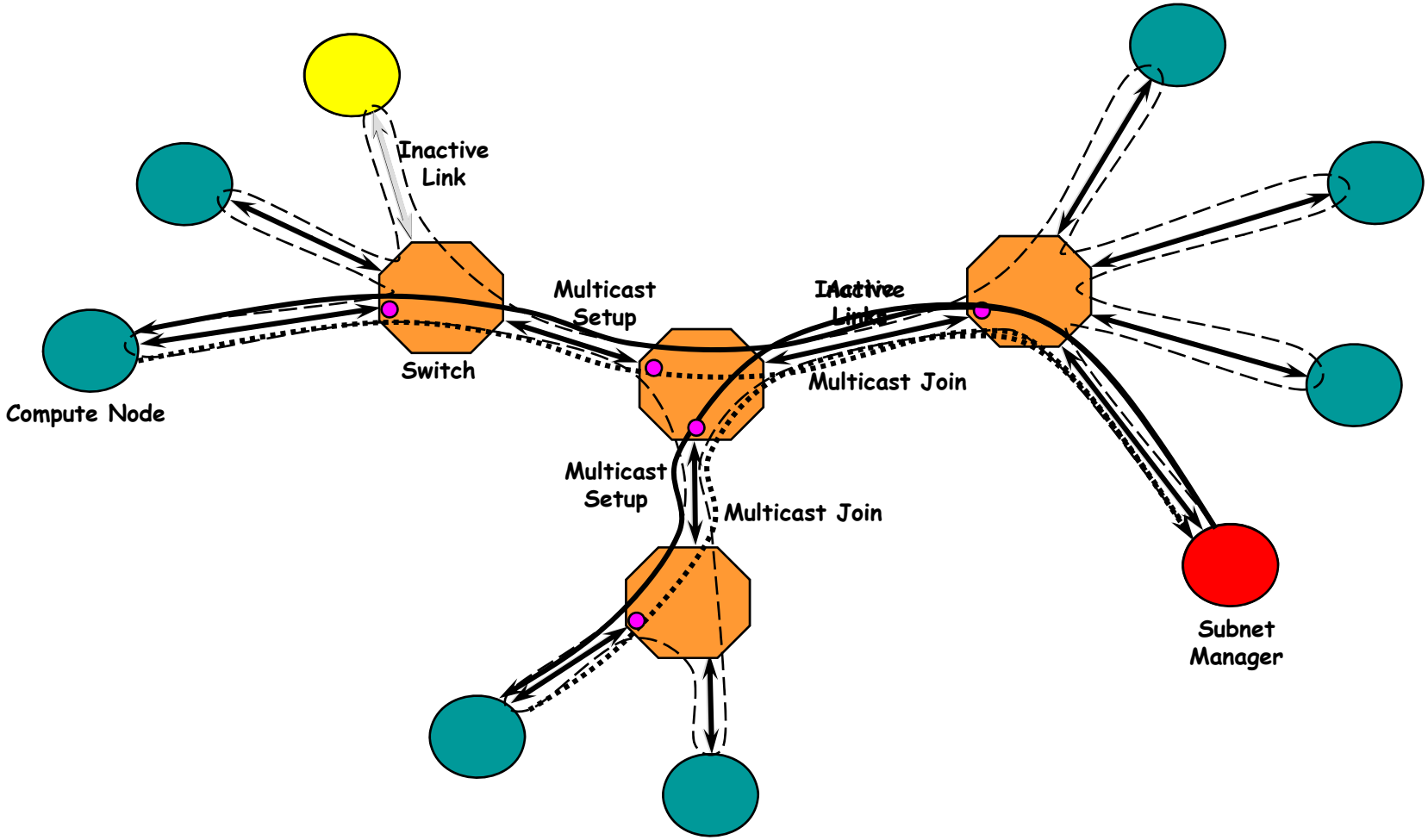
- Agents
  - Processes or hardware units running on each adapter, switch, router (everything on the network)
  - Provide capability to query and set parameters
- Managers
  - Make high-level decisions and implement it on the network fabric using the agents
- Messaging schemes
  - Used for interactions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

Management Model



- QP0 and QP1 are special QPs on each port
  - QP0 provides Subnet Management Interface (services)
  - QP1 provides General Services Interface (services)

Subnet Manager



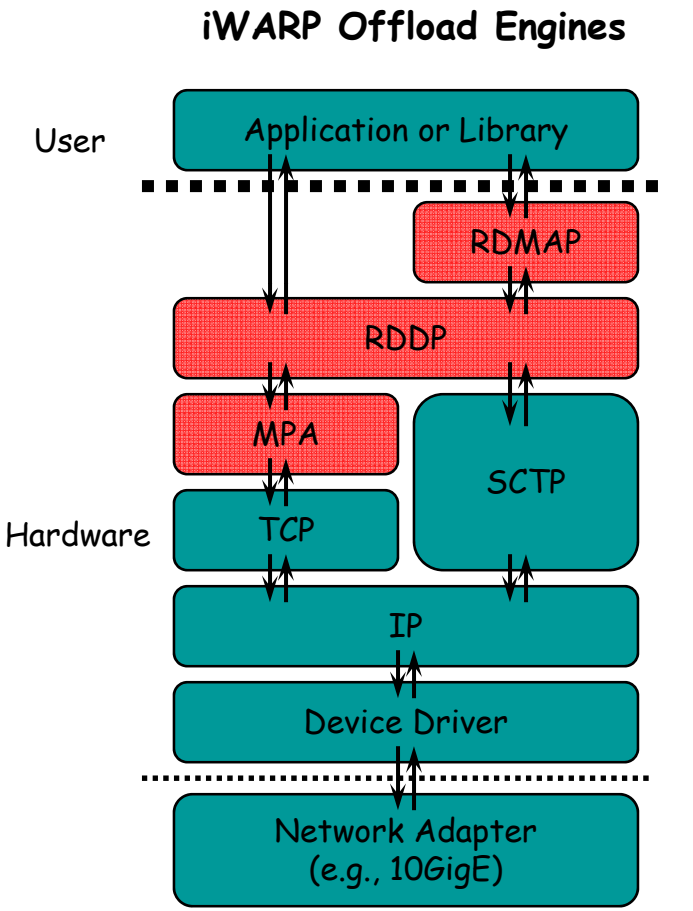
IB and 10GE Overview

- InfiniBand
  - […snip…]
- 10-Gigabit Ethernet Family
  - Architecture and Components
  - Out-of-Order Data Placement
  - Dynamic and Fine-grained Data Rate Control
  - Existing Implementations of 10GE/iWARP

IB and 10GE: Commonalities and Differences

	IB	iWARP/10GE
Hardware Acceleration	Supported	Supported (for TOE and iWARP)
RDMA	Supported	Supported (for iWARP)
Atomic Operations	Supported	Not supported
Multicast	Supported	Supported
Data Placement	Ordered	Out-of-order (for iWARP)
Data Rate-control	Static and Coarse-grained	Dynamic and Fine-grained (for TOE and iWARP)
QoS	Prioritization	Prioritization and Fixed Bandwidth QoS

iWARP Architecture and Components



(Courtesy iWARP Specification)

- RDMA Protocol (RDMAP)
  - Feature-rich interface
  - Security Management
- Remote Direct Data Placement (RDDP)
  - Data Placement and Delivery
  - Multi Stream Semantics
  - Connection Management
- Marker PDU Aligned (MPA)
  - Middle Box Fragmentation
  - Data Integrity (CRC)

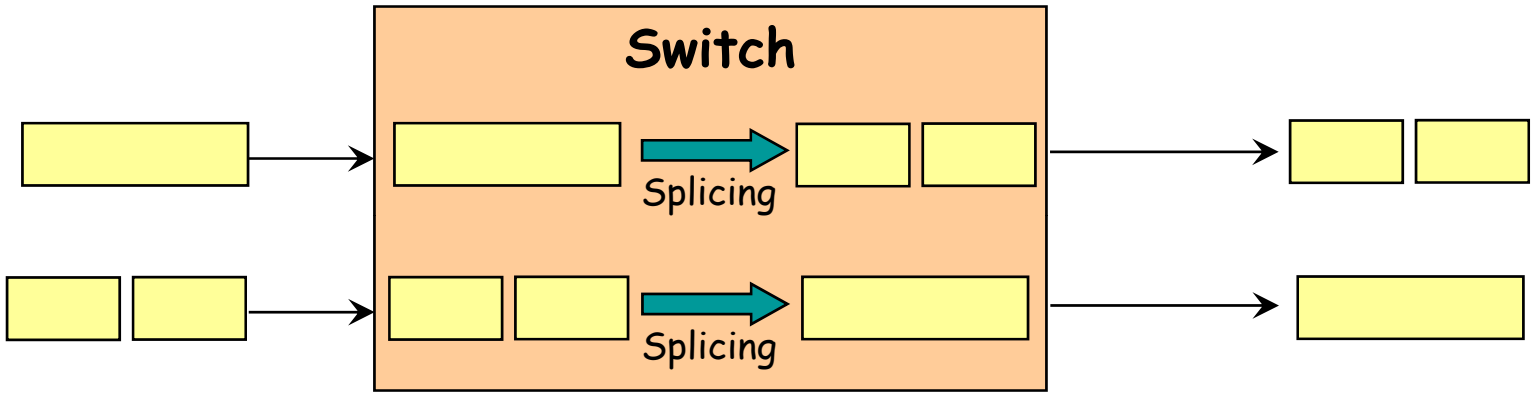
Decoupled Data Placement and Data Delivery

- Place data as it arrives, whether in-order or out-of-order
- If data is out-of-order, place it at the appropriate offset
- Issues from the application's perspective:
  - If the second half of the message has been placed, it does not mean that the first half of the message has arrived as well
  - If one message has been placed, it does not mean that that the previous messages have been placed

Protocol Stack Issues with Out-of-Order Data Placement

- The receiver network stack has to understand each frame of data
- If the frame is unchanged from what is sent by the sender, this is easy!
- Can we guarantee that the frame will be unchanged?
- Intermediate switch segmentations?

Switch Splicing

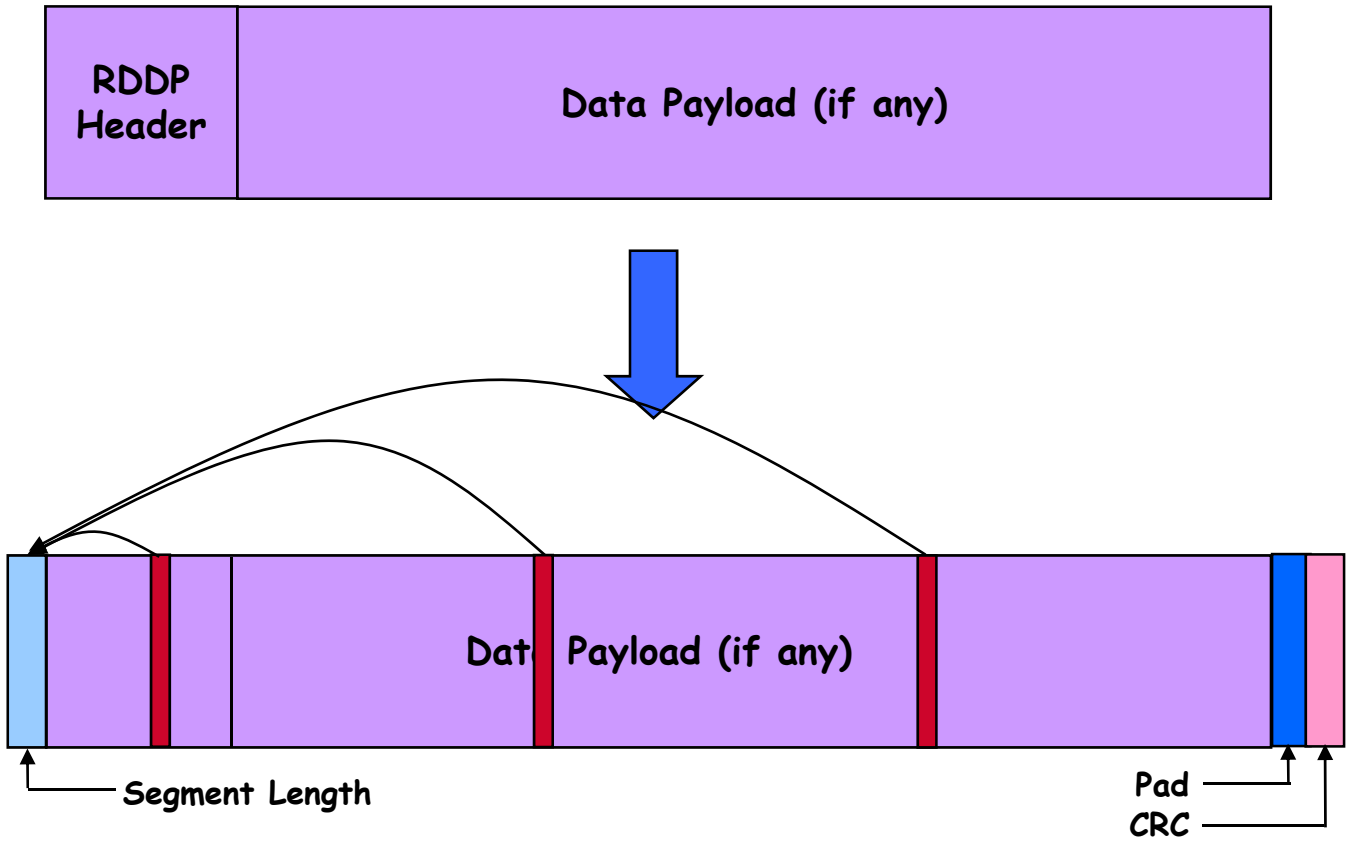


Intermediate Ethernet switches (e.g., those which support splicing) can segment a frame to multiple segments or coalesce multiple segments to a single segment

Marker PDU Aligned (MPA) Protocol

- Deterministic Approach to find segment boundaries
- Approach
  - Places strips of data at regular intervals (based on data sequence number)
  - Interval is set to be 512 bytes (small enough to ensure that each Ethernet frame has at least one)
  - Each strip points to the RDDP header
- Each segment independently has enough information about where it needs to be placed

MPA Frame Format



Dynamic and Fine-grained Rate Control

- Part of the Ethernet standard, not iWARP
  - Network vendors use a separate interface to support it
- Dynamic bandwidth allocation to flows based on interval between two packets in a flow
  - E.g., one stall for every packet sent on a 10-Gig network refers to a bandwidth allocation of 5Gbps
  - Complicated because of TCP windowing behavior
- Important for high-latency/high-bandwidth networks
  - Large windows exposed on the receiver side
  - Receiver overflow controlled through rate control

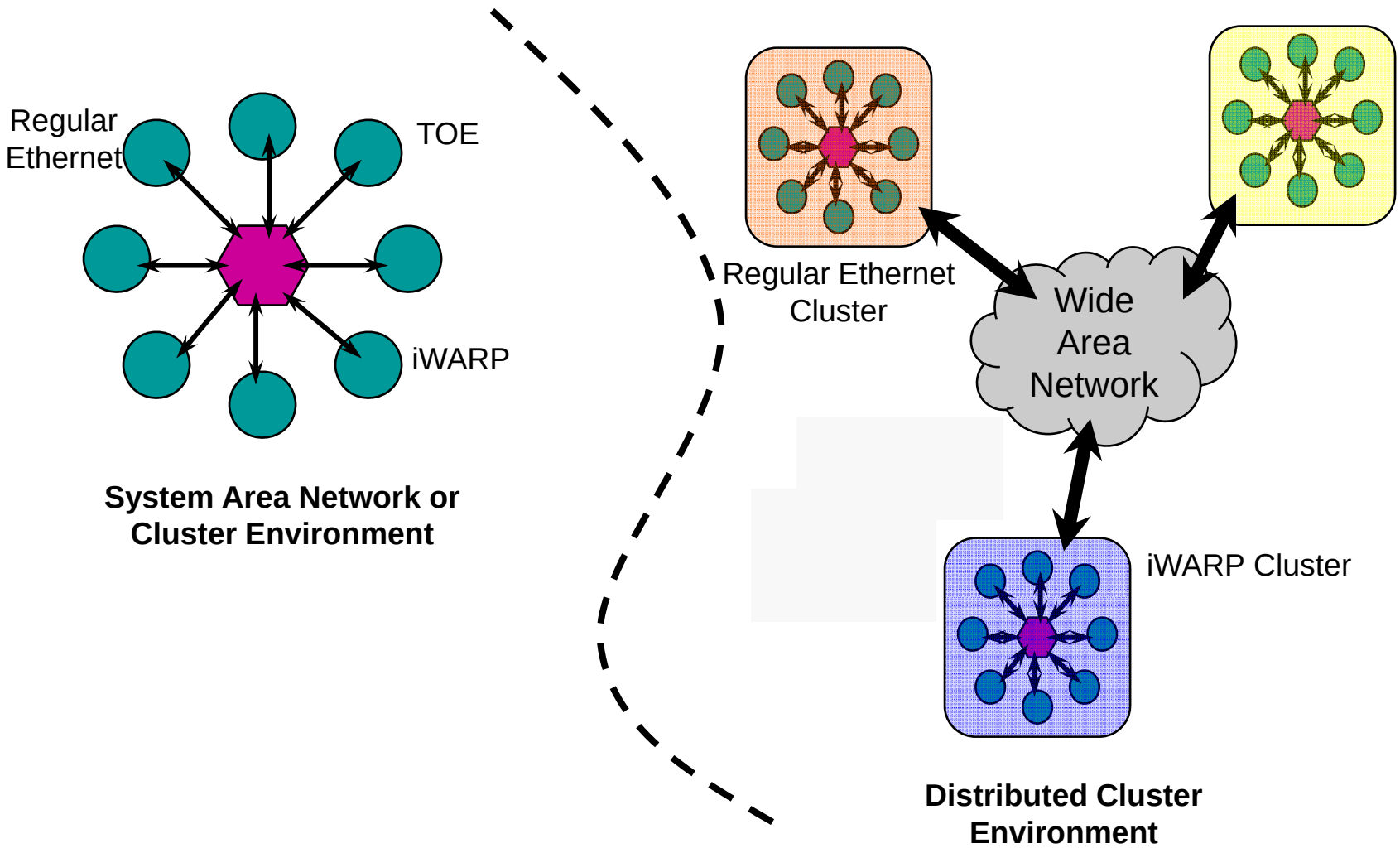
Prioritization vs. Fixed Bandwidth QoS

- Can allow for simple prioritization:
  - E.g., connection 1 performs better than connection 2
  - 8 classes provided (a connection can be in any class)
    - Similar to SLs in InfiniBand
  - Two priority classes for high-priority traffic
    - E.g., management traffic or your favorite application
- Or can allow for specific bandwidth requests:
  - E.g., can request for 3.62 Gbps bandwidth
  - Packet pacing and stalls used to achieve this
- Query functionality to find out "remaining bandwidth"

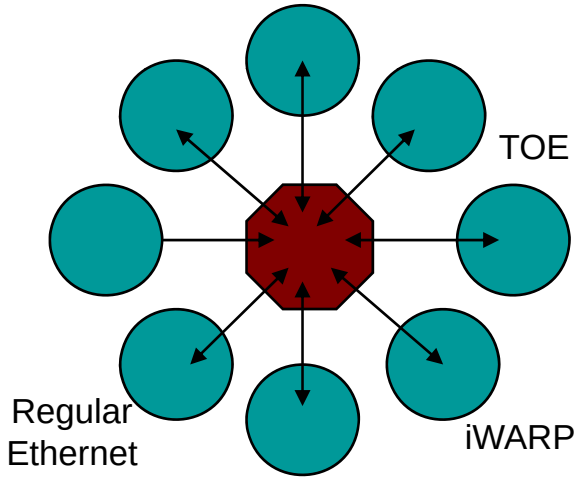
IB and 10GE Overview

- InfiniBand
  - […snip…]
- 10-Gigabit Ethernet Family
  - Architecture and Components
  - Out-of-Order Data Placement
  - Dynamic and Fine-grained Data Rate Control
  - Existing Implementations of 10GE/iWARP

Current Usage of Ethernet



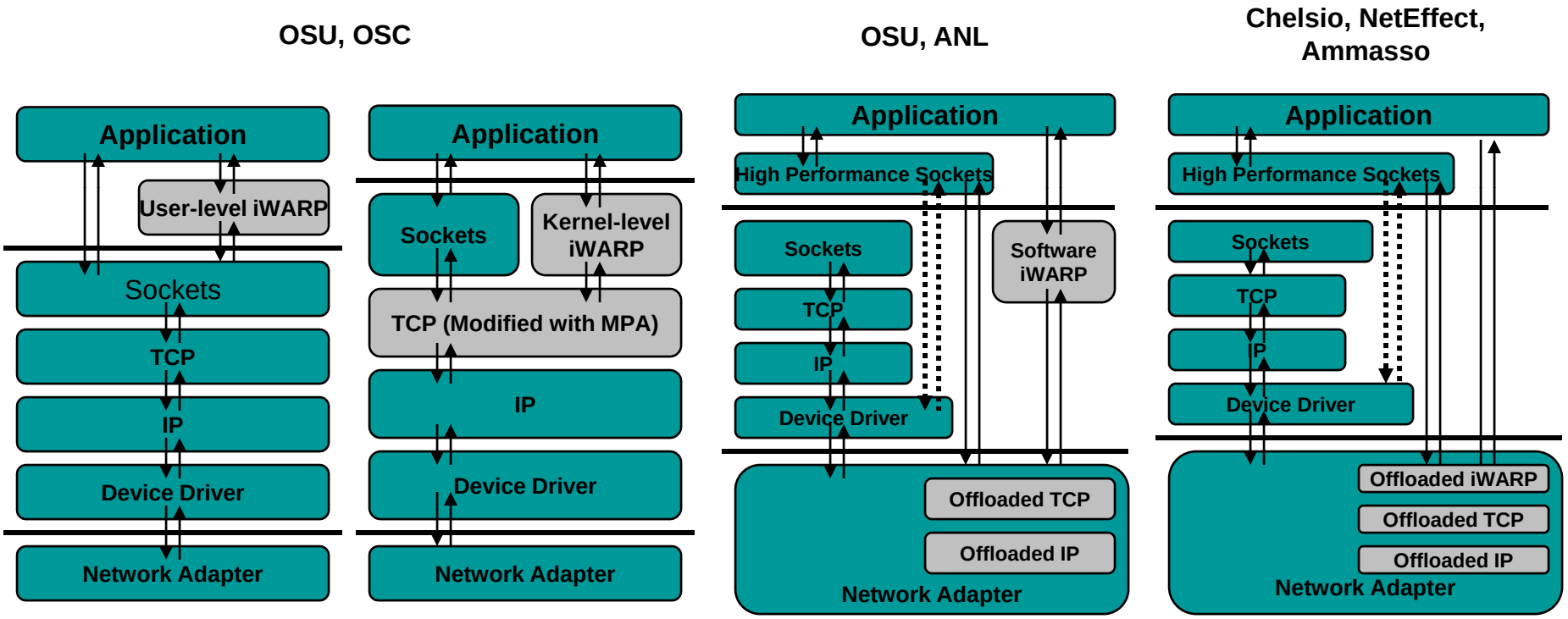
Software iWARP based Compatibility



Ethernet Environment

- Regular Ethernet adapters and TOEs are fully compatible
- Compatibility with iWARP required
- Software iWARP emulates the functionality of iWARP on the host
  - Fully compatible with hardware iWARP
  - Internally utilizes host TCP/IP stack

Different iWARP Implementations



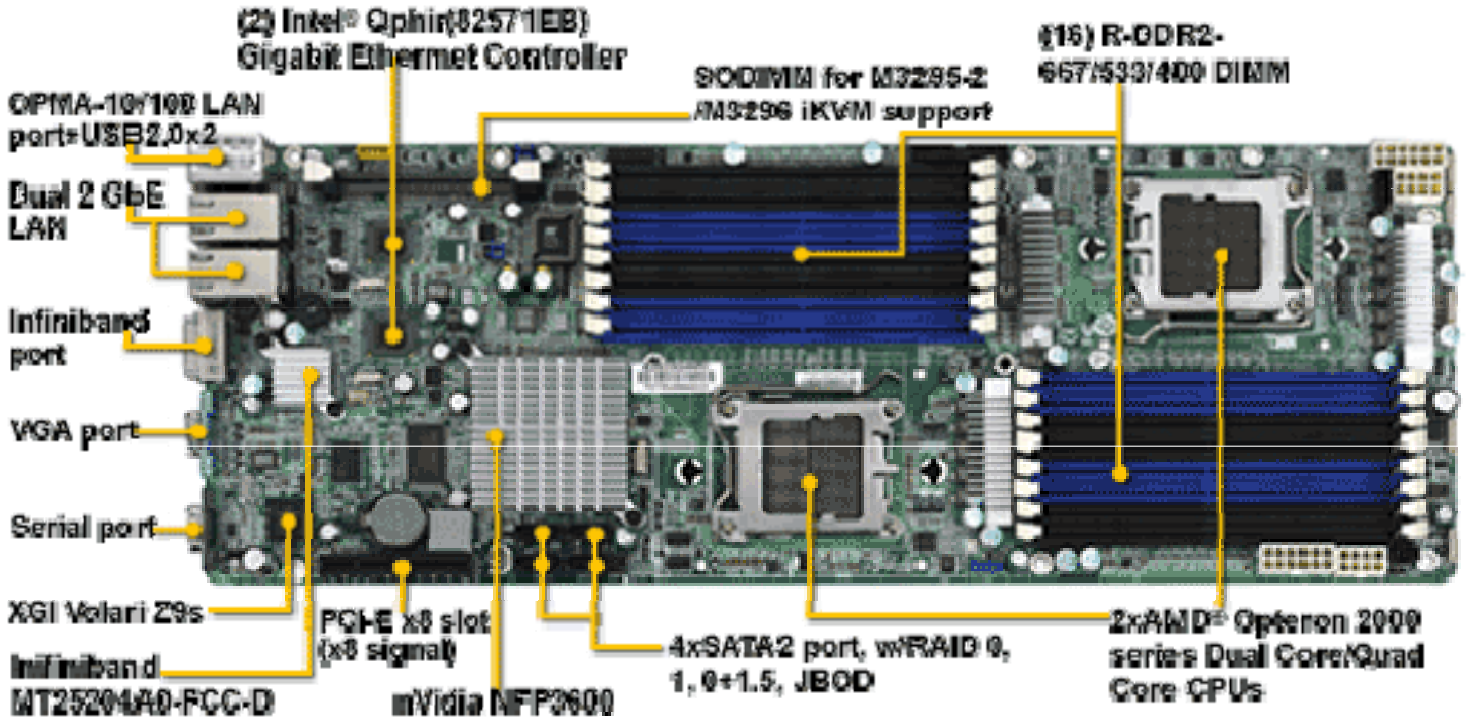
Presentation Overview

- Introduction
- Why InfiniBand and 10-Gigabit Ethernet?
- Overview of IB and 10GE and their Novel Features
- IB and 10GE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
  - MPI, SDP, File Systems, Data Center and Virtualization
- Conclusions and Final Q&A

IB Hardware Products

- Many IB vendors: Mellanox, Voltaire, Cisco, Qlogic
  - Aligned with many server vendors: Intel, IBM, SUN, Dell
  - And many integrators: Appro, Advanced Clustering, Microway, …
- Broadly two kinds of adapters
  - Offloading (Mellanox) and Onloading (Qlogic)
- Adapters with different interfaces:
  - Dual port 4X with PCI-X (64 bit/133 MHz), PCIe x8, PCIe 2.0 and HT
- MemFree Adapter
  - No memory on HCA → Uses System memory (through PCIe)
  - Good for LOM designs (Tyan S2935, Supermicro 6015T-INFB)
- Different speeds
  - SDR (8 Gbps), DDR (16 Gbps) and QDR (32 Gbps)
- Some 12X SDR adapters exist as well (24 Gbps each way)

Tyan Thunder S2935 Board



(Courtesy Tyan)

IB Hardware Products (contd.)

- Customized adapters to work with IB switches
  - Cray XD1 (formerly by Octigabay), Cray CX1
- Switches:
  - 4X SDR switch (8-288 ports)
    - 12X ports available for inter-switch connectivity
  - 4X DDR switch (mainly available in 8 to 288 port models)
  - 12X switches (small sizes available)
  - 3456-port "Magnum" switch from SUN → used at TACC
    - 72-port "nano magnum" switch with DDR speed
  - New 36-port InfiniScale IV QDR switch silicon by Mellanox
    - Will allow high-density switches to be built
- Switch Routers with Gateways
  - IB-to-FC; IB-to-IP

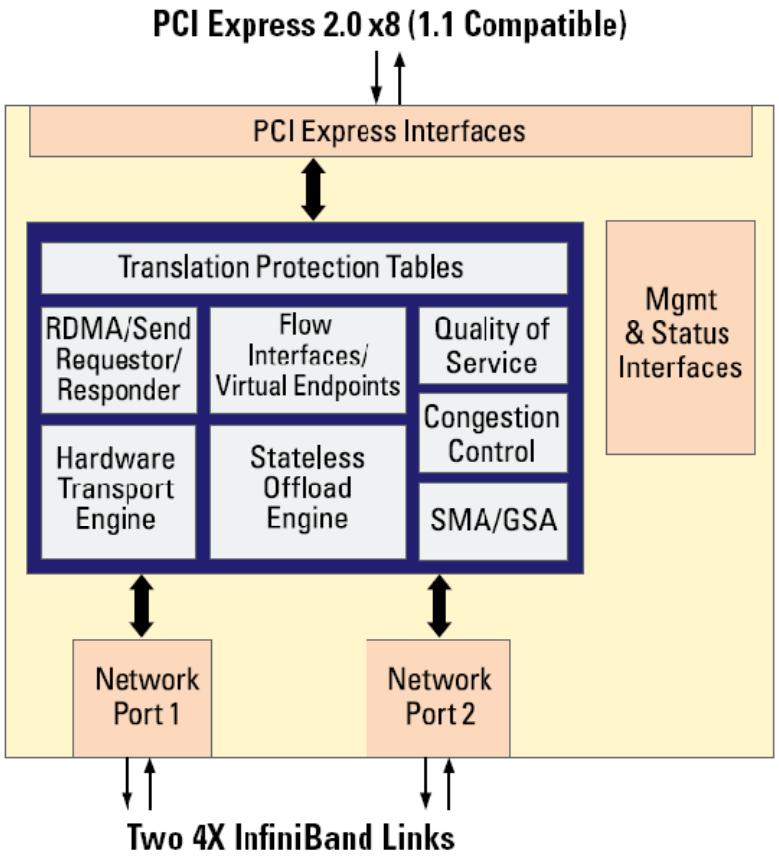
IB Software Products

- Low-level software stacks
  - VAPI (Verbs-Level API) from Mellanox
  - Modified and customized VAPI from other vendors
  - New initiative: Open Fabrics (formerly OpenIB)
    - http://www.openfabrics.org
    - Open-source code available with Linux distributions
    - Initially IB; later extended to incorporate iWARP
- High-level software stacks
  - MPI, SDP, IPoIB, SRP, iSER, DAPL, NFS, PVFS on various stacks (primarily VAPI and OpenFabrics)

10GE Products

- 10GE/iWARP adapters
  - Chelsio, NetEffect, NETXEN
- 10GE switches
  - Fulcrum Microsystems
    - Low latency switch based on 24-port silicon
    - FM4000 switch announced recently with layer 3 IP routing, layer 4 TCP and UDP support
  - Fujitsu, Woven Systems (144-port switch), Myricom (up to 512 ports), Quadrics (up to 96 ports), Force10, Cisco, Arastra

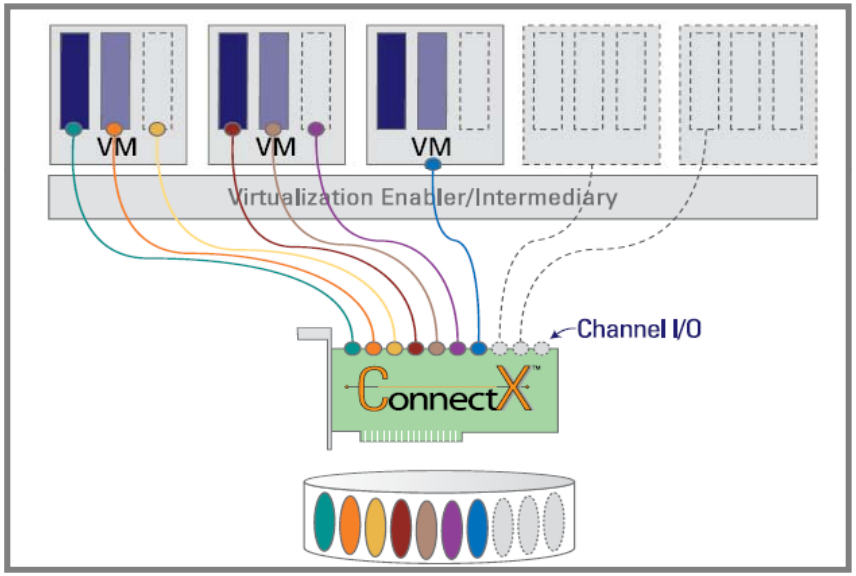
Mellanox ConnectX Architecture



(Courtesy Mellanox)

- Fourth Generation Silicon
  - DDR (Double Data Rate) with PCIe Gen1 or Gen2
  - QDR (Quad Data Rate) with PCIe Gen2
- Can configure each individual port to either IB or 10GE
- Hardware support for Virtualization
- Quality of Service
- Stateless Offloads

Direct Hardware Access in Virtual Machines



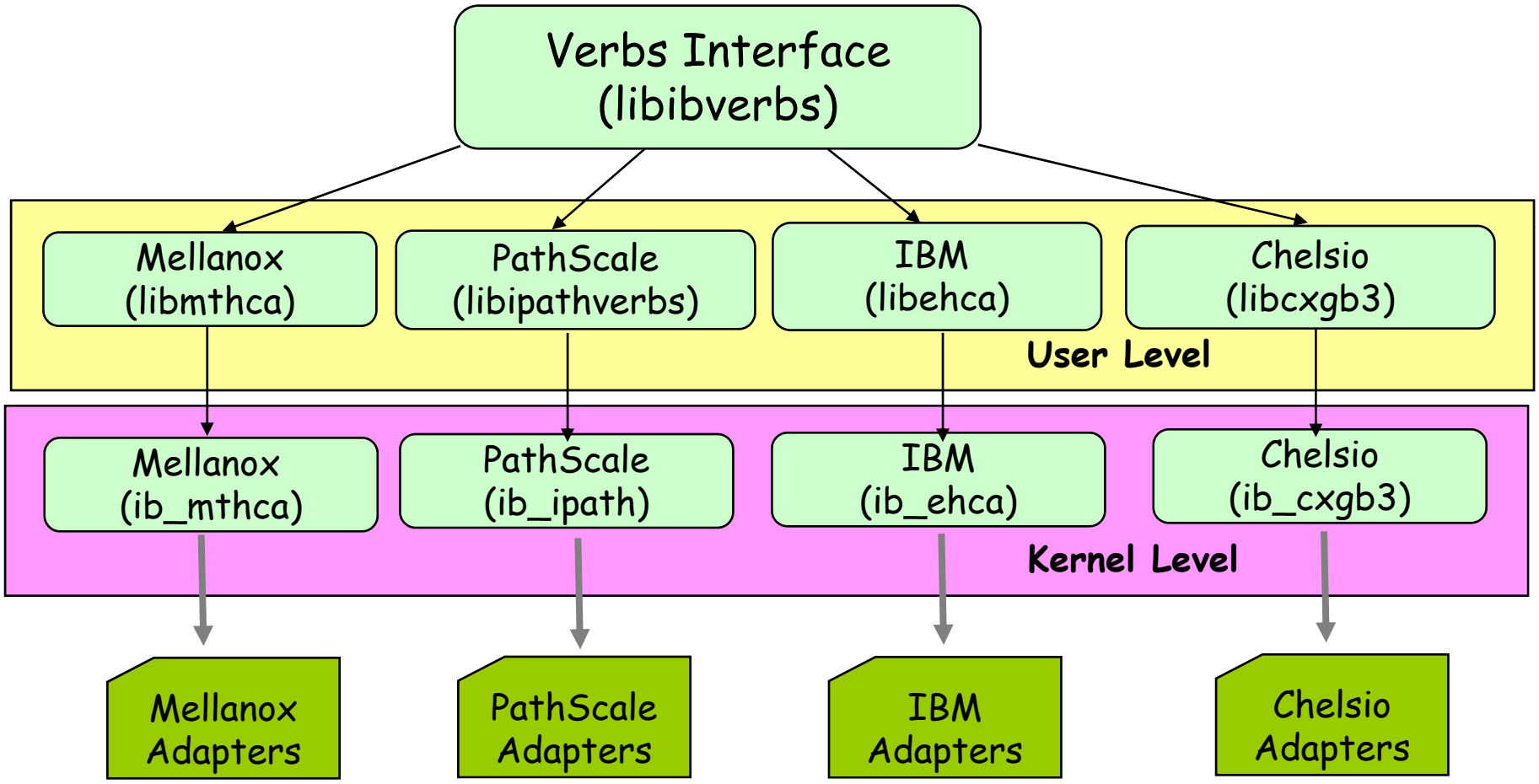
(Courtesy Mellanox)

- Hardware provides isolation for virtual machines
- Virtual Machines can directly access without going to privileged domain
- Dedicated end-to-end connections

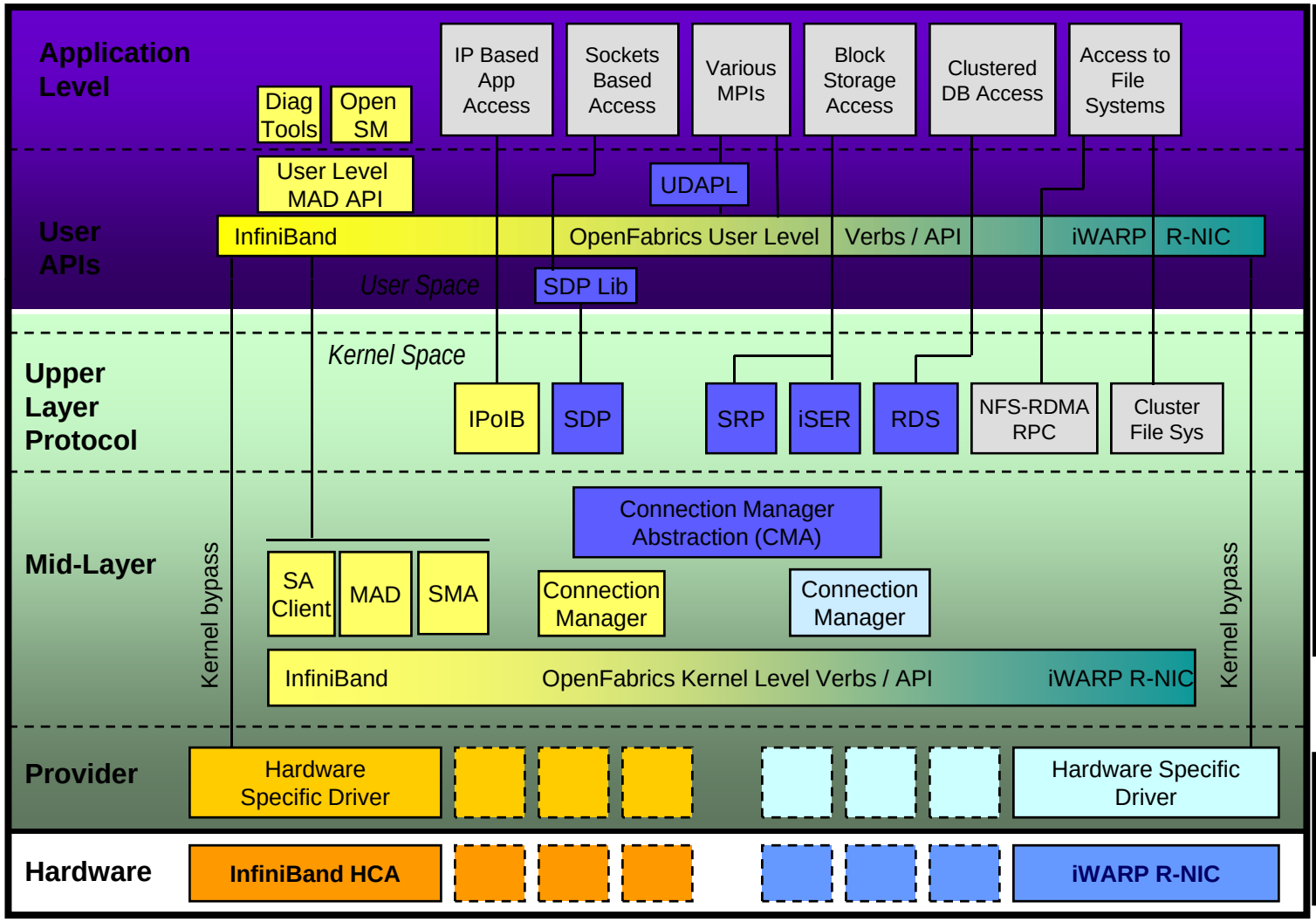
OpenFabrics

- www.openfabrics.org
- Open source organization (formerly OpenIB)
- Incorporates both IB and iWARP in a unified manner
- Focusing on effort for Open Source IBA and iWARP support for Linux and Windows
- Design of complete software stack with `best of breed' components
  - Gen1
  - Gen2 (current focus)
- Users can download the entire stack and run
  - Latest release is OFED 1.3.1
  - OFED 1.4 is being worked out

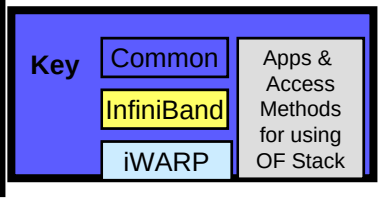
OpenFabrics Stack with Unified Verbs Interface



OpenFabrics Software Stack



SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA	Host Channel Adapter
R-NIC	RDMA NIC



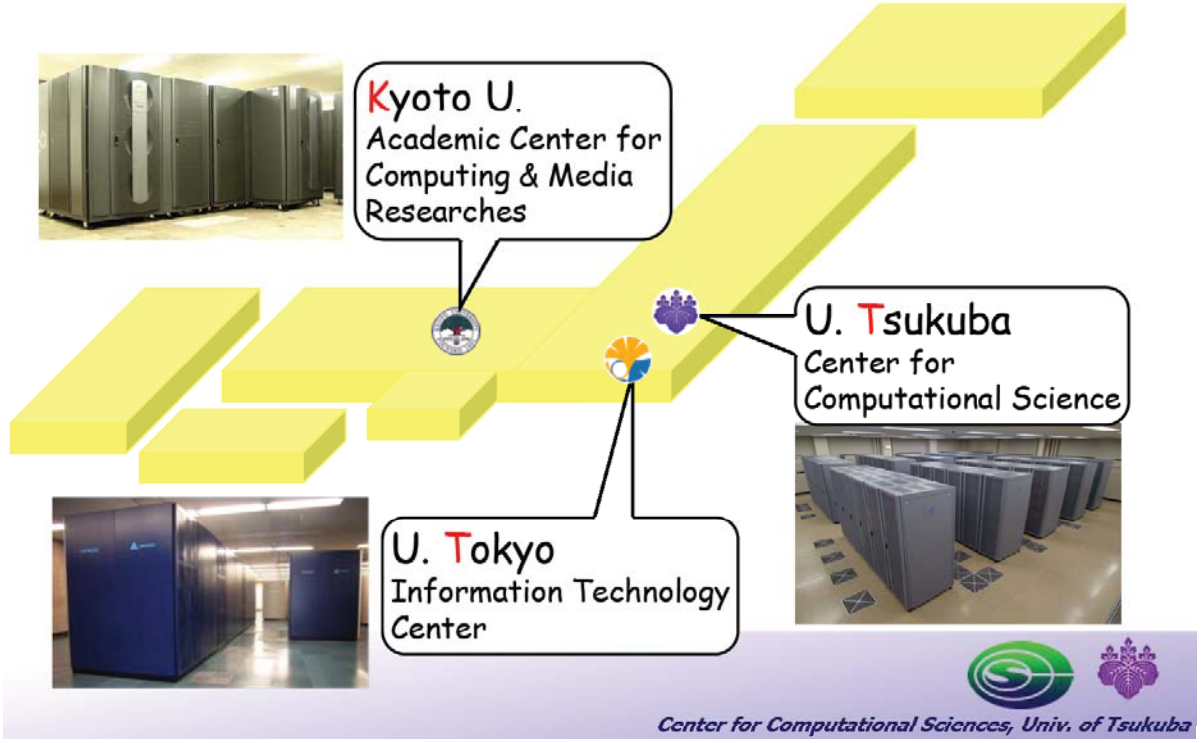
IB Installations

- 121 IB clusters (24.2%) in June '08 TOP500 list (www.top500.org)
- 12 IB clusters in TOP25
  - 122,400-cores (RoadRunner) at LANL (1st)
  - 62,976-cores (Ranger) at TACC (4th)
  - 14,336-cores at New Mexico (7th)
  - 14,384-cores at Tata CRL, India (8th)
  - 10,240-cores at TEP, France (10th)
  - 13,728-cores in Sweden (11th)
  - 8,320-cores in UK (18th)
  - 6,720-cores in Germany (19th)
  - 10,000-cores at CCS, Tsukuba, Japan (20th)
  - 9,600-cores at NCSA (23rd)
  - 12,344-cores at Tokyo Inst. of Technology (24th)
  - 13,824-cores at NASA/Columbia (25th)
- More are getting installed ....

10GE Installations

- Several Enterprise Computing Domains
  - Enterprise Datacenters (HP, Intel)
  - Animation firms (e.g., Universal Studios created "The Hulk" and many new movies using 10GE)
- Scientific Computing Installations
  - 640-core installation in University of Heidelberg, Germany
  - 512-core installation at Sandia National Laboratory (SNL) with Chelsio/iWARP and Woven Systems switch
  - 256-core installation at Ohio Supercomputer Center (OSC) with Ammasso/iWARP
  - 256-core installation at Argonne National Lab with Myri-10G
- Integrated Systems
  - BG/P uses 10GE for I/O (ranks 3, 6, 9, 13, 37 in the Top 50)

Dual IB/10GE Systems



(Courtesy Taisuke Boku, University of Tsukuba)

- Such systems are being integrated
- E.g., the T2K-Tsukuba system (300 TFlop System)
- Systems at three sites (Tsukuba, Tokyo, Kyoto)

- Internal connectivity: Quad-rail IB ConnectX network
- External connectivity: 10GE

Presentation Overview

- Introduction
- Why InfiniBand and 10-Gigabit Ethernet?
- Overview of IB and 10GE and their Novel Features
- IB and 10GE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
  - MPI, SDP, File Systems, Data Center and Virtualization
- Conclusions and Final Q&A

Sample Case Studies

- Message Passing Interface (MPI)
- SDP and IPoIB
- File Systems (Lustre, NFS-RDMA)
- Datacenter
- Virtualization

Message Passing Interface (MPI)

- De-facto message passing standard
  - Point-to-point communication
  - Collective communication (broadcast, multicast, reduction, barrier)
  - MPI-1 and MPI-2 available; MPI-3 under discussion
- Has been implemented for various past commodity networks (Myrinet, Quadrics)
- How can it be designed and efficiently implemented for InfiniBand and iWARP?

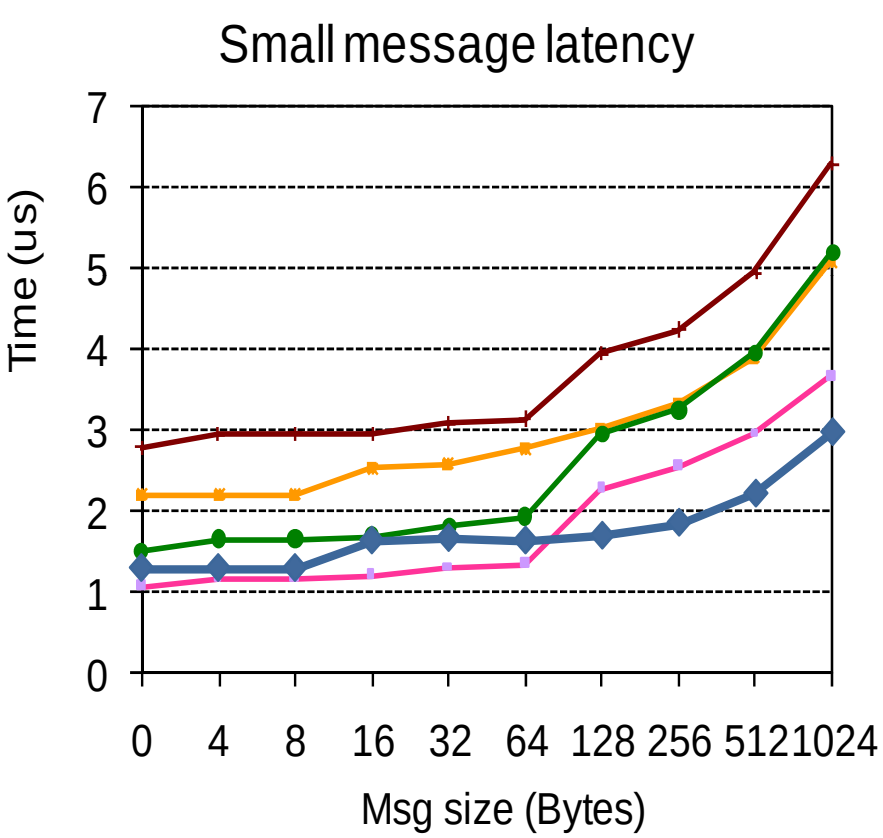
MVAPICH/MVAPICH2 Software

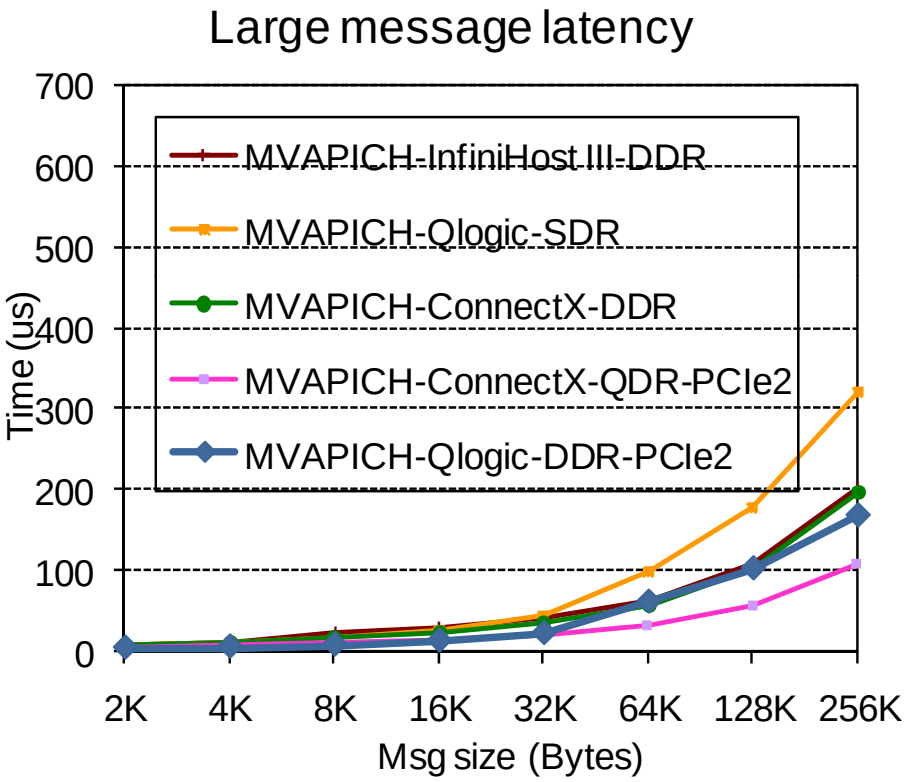
- High Performance MPI Library for IB and 10GE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
  - Used by more than 760 organizations in 42 countries
  - More than 23,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 4th ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, 10GE and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - Also supports uDAPL device to work with any network supporting uDAPL
  - http://mvapich.cse.ohio-state.edu/

MPICH2 Software Stack

- High-performance and Widely Portable MPI
  - Supports MPI-1, MPI-2 and MPI-2.1
  - Supports multiple networks (TCP, IB, iWARP, Myrinet)
  - Commercial support by many vendors
    - IBM (integrated stack distributed by Argonne)
    - Microsoft, Intel (in process of integrating their stack)
  - Used by many derivative implementations
    - E.g., MVAPICH2, IBM, Intel, Microsoft, SiCortex, Cray, Myricom
    - MPICH2 and its derivatives support many Top500 systems (estimated at more than 90%)
  - Available with many software distributions
  - Integrated with the ROMIO MPI-IO implementation and the MPE profiling library
  - http://www.mcs.anl.gov/research/projects/mpich2

One-way Latency: MPI over IB

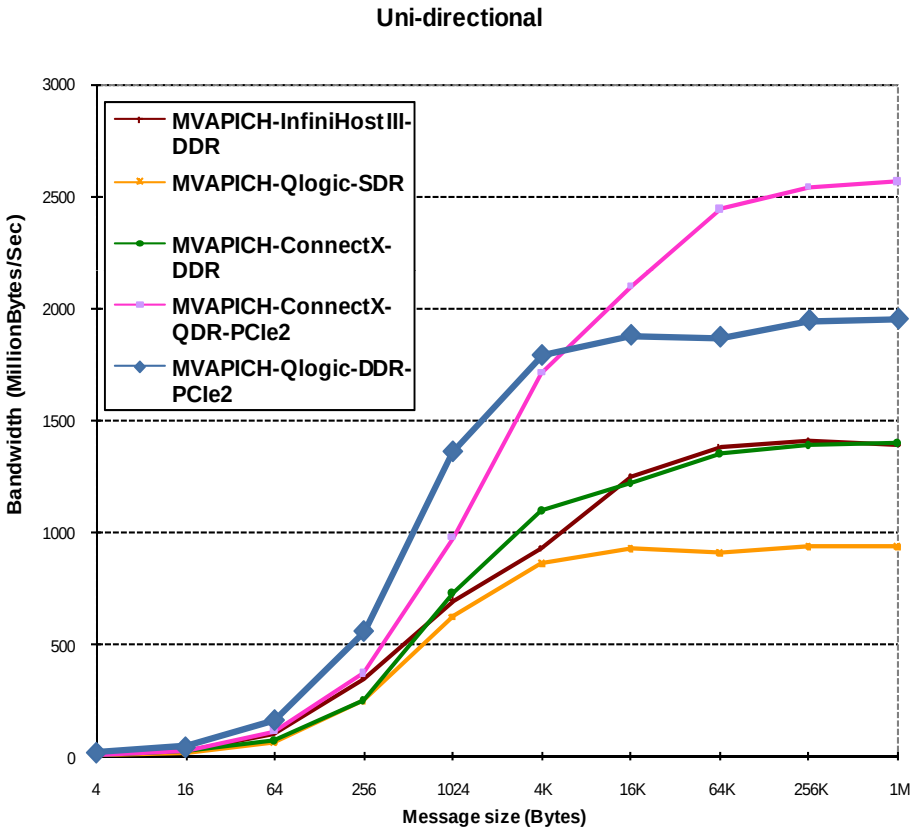




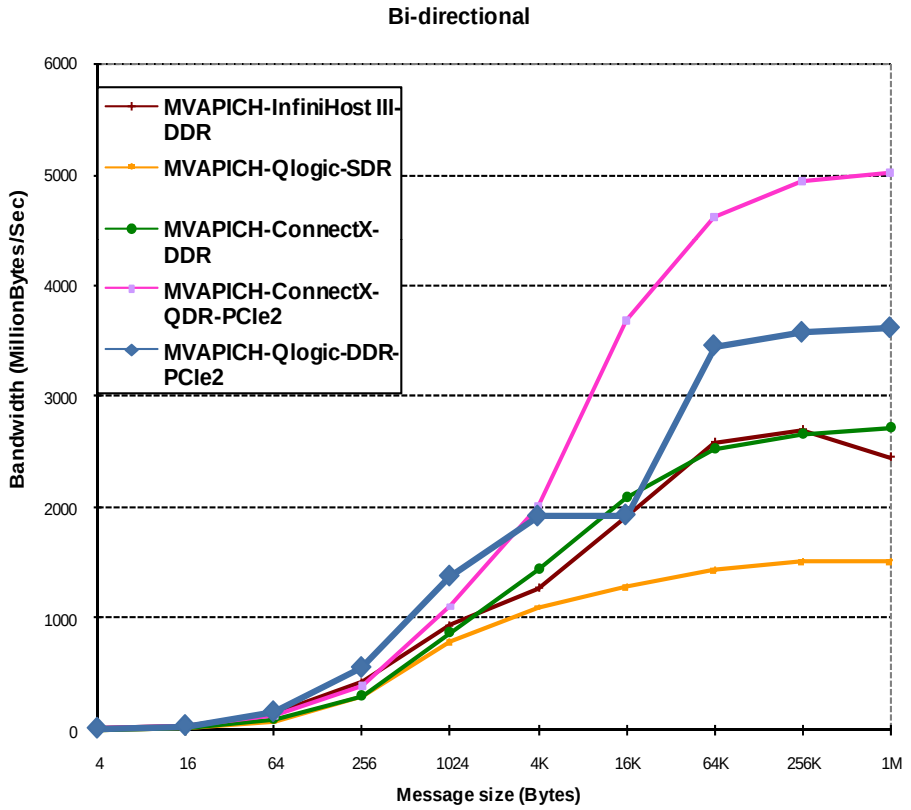
InfiniHost III and ConnectX-DDR: 2.33 GHz Quad-core (Clovertown) Intel with IB switch

ConnectX-QDR-PCIe2: 2.83 GHz Quad-core (Harpertown) Intel with back-to-back

Bandwidth: MPI over IB

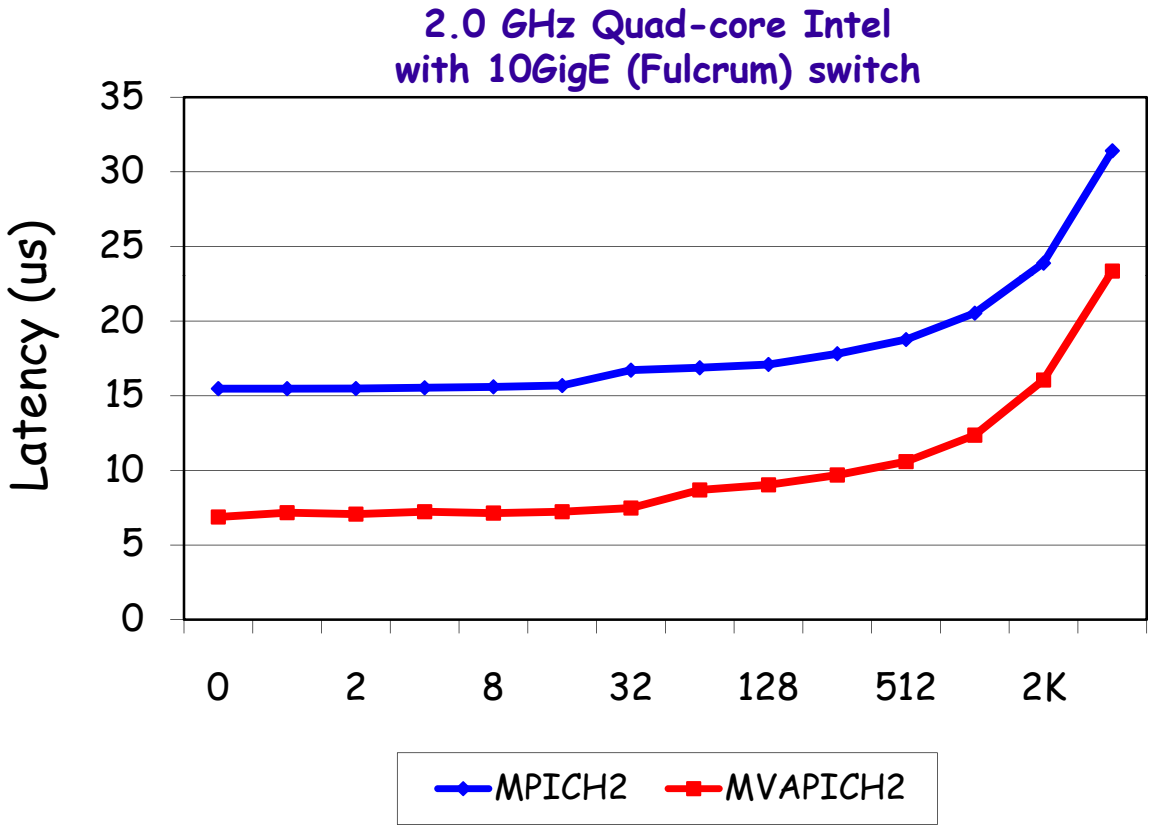


InfiniHost III and ConnectX-DDR: 2.33 GHz Quad-core (Clovertown) Intel with IB switch



ConnectX-QDR-PCIe2: 2.83 GHz Quad-core (Harpertown) Intel with back-to-back

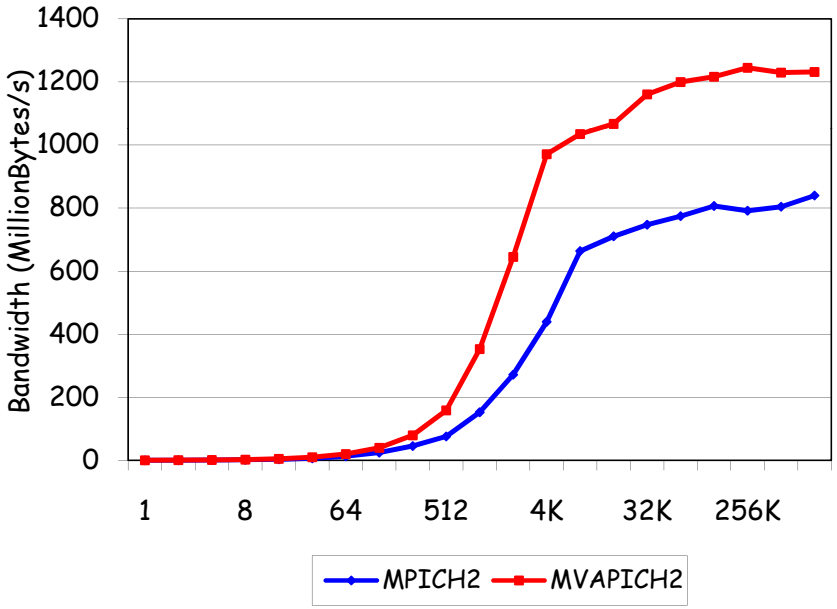
One-way Latency: MPI over iWARP (Chelsio)



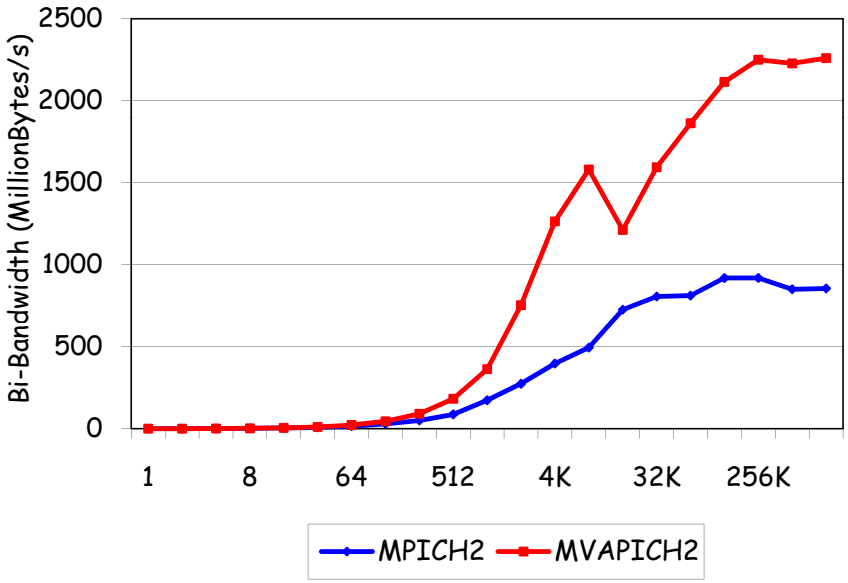
MVAPICH2 gives a latency of about 6.88us as compared to 15.47 for MPICH2

Bandwidth: MPI over iWARP (Chelsio)

2.0 GHz Quad-core Intel
with 10GigE (Fulcrum)
switch



Peak bandwidth of
about 1231 MillionBytes/s

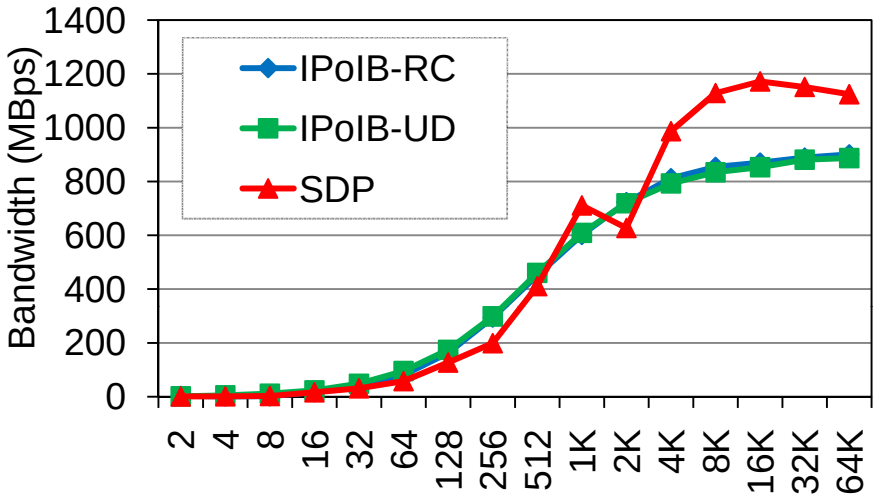


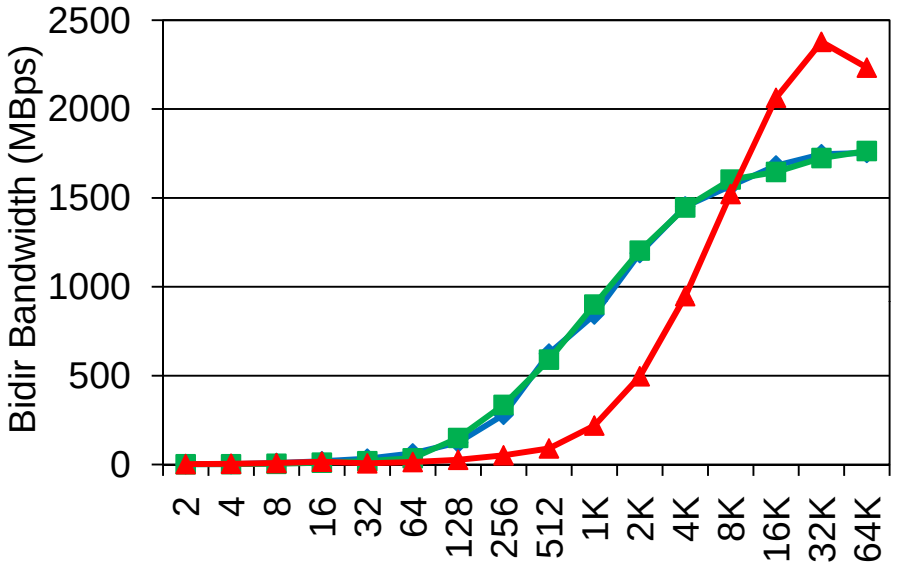
Peak bi-dir bandwidth of
about 2380 MillionBytes/s

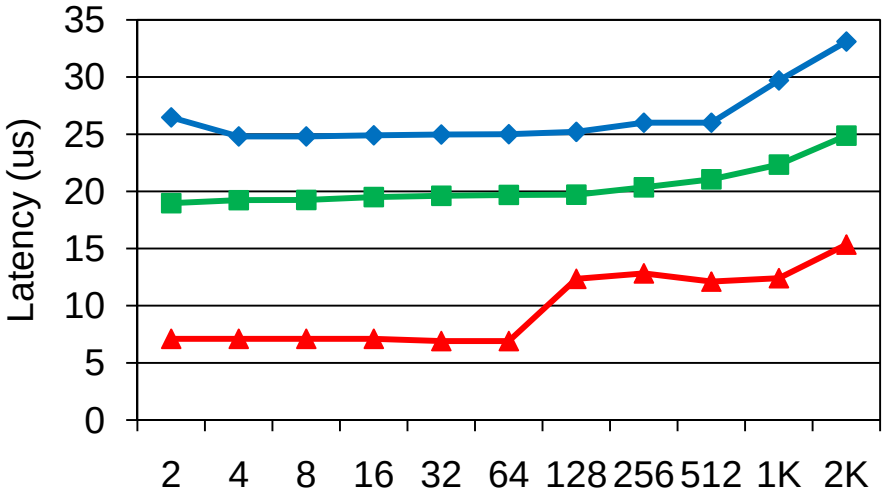
Sample Case Studies

- Message Passing Interface (MPI)
- SDP and IPoIB
- File Systems (Lustre, NFS-RDMA)
- Datacenter
- Virtualization

SDP vs. IPoIB (IB DDR)

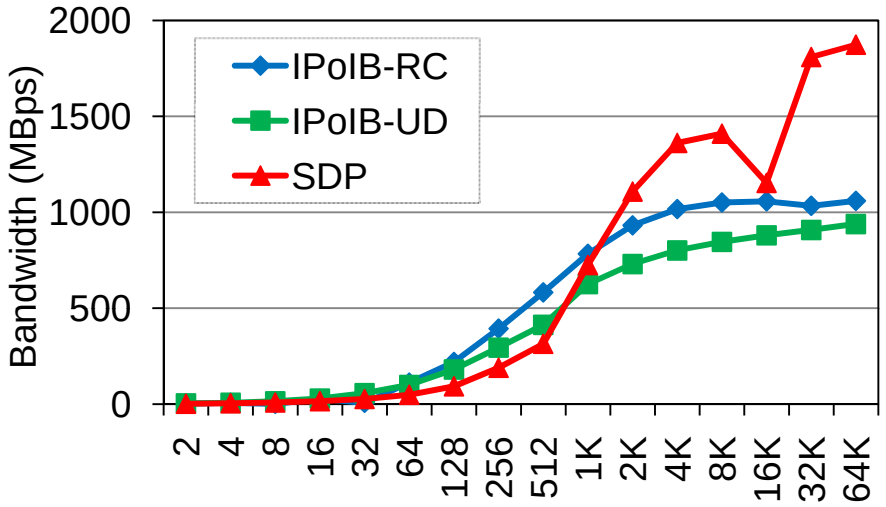


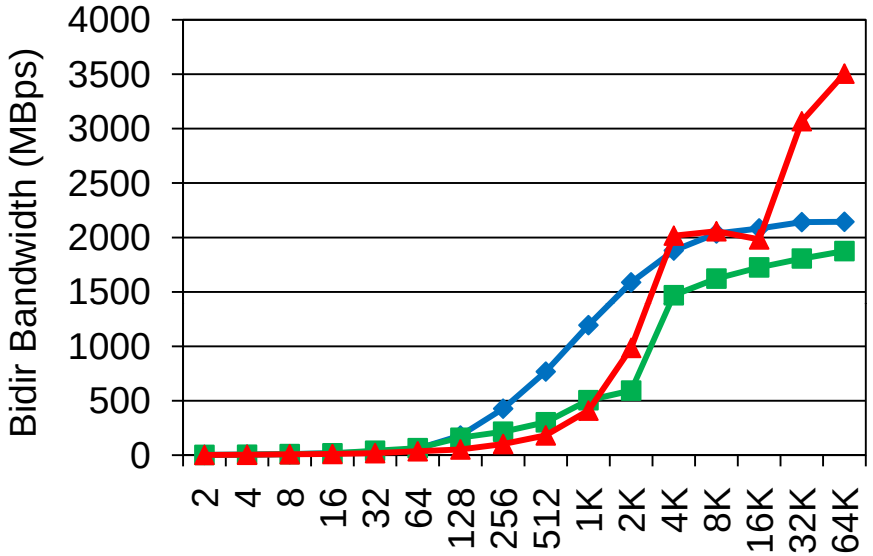


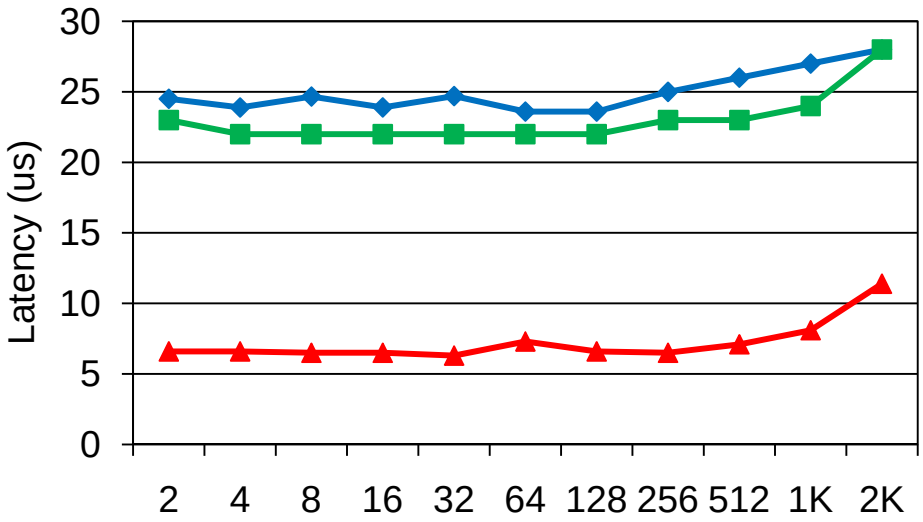


SDP enables high bandwidth
(up to **9.3** Gbps),
low latency (7 µs)

SDP vs. IPoIB (IB QDR)





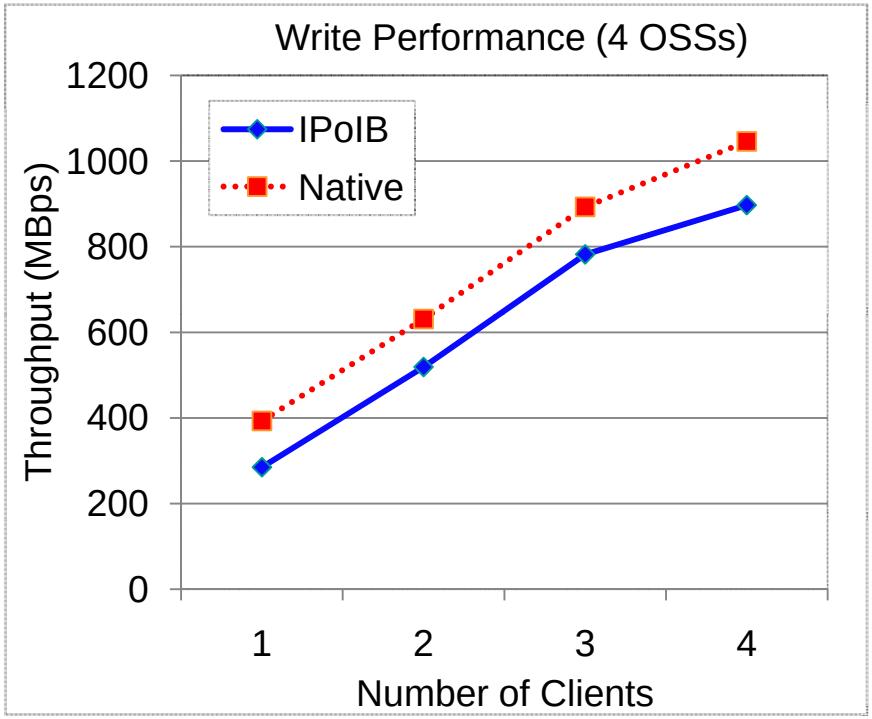


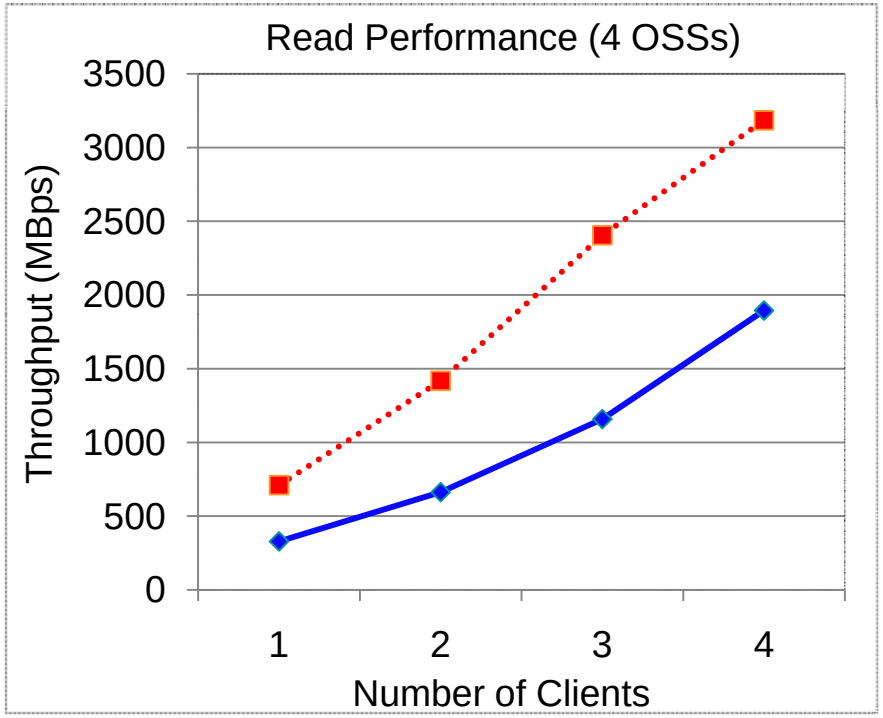
SDP enables high bandwidth
(up to **15** Gbps),
low latency (6.6 µs)

Sample Case Studies

- Message Passing Interface (MPI)
- SDP and IPoIB
- File Systems (Lustre, NFS-RDMA)
- Datacenter
- Virtualization

Lustre Performance



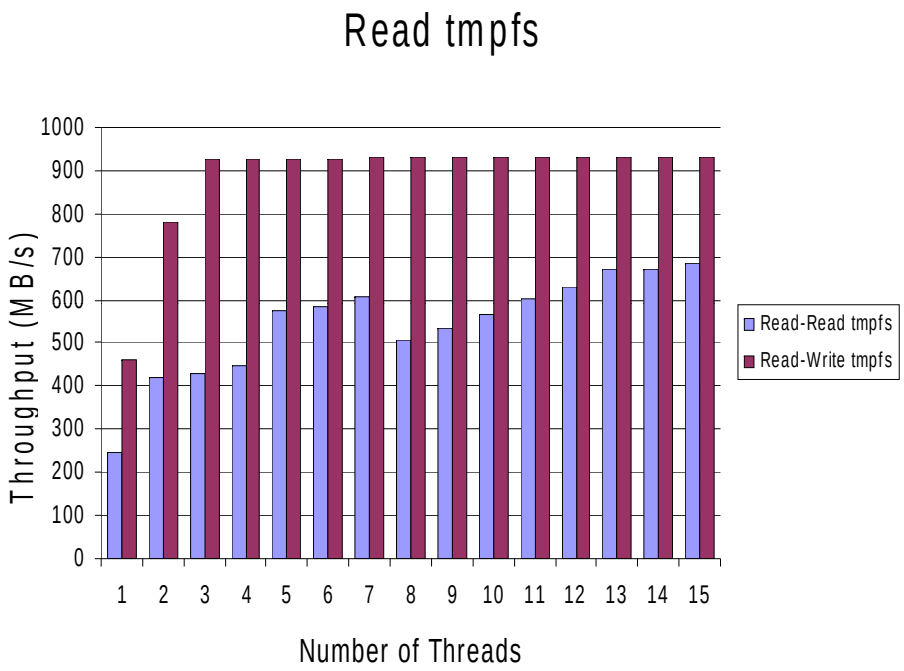


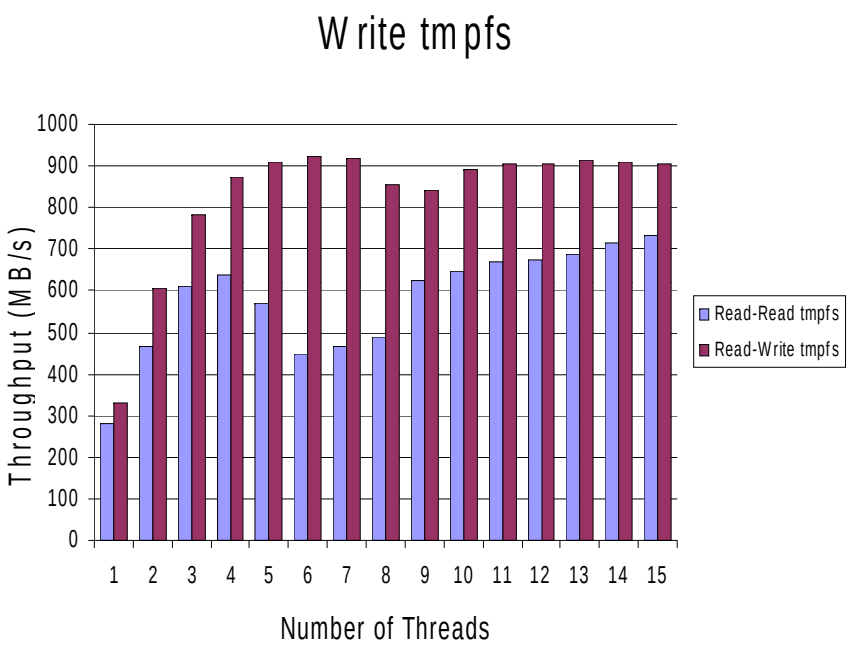
- Lustre over Native IB
  - Write: 1.38X faster than IPoIB; Read: 2.16X faster than IPoIB
- Memory copies in IPoIB and Native IB
  - Reduced throughput and high overhead; I/O servers are saturated

CPU Utilization

- 4 OSS nodes, IOzone record size 1MB
- Offers potential for greater scalability

NFS/RDMA Performance





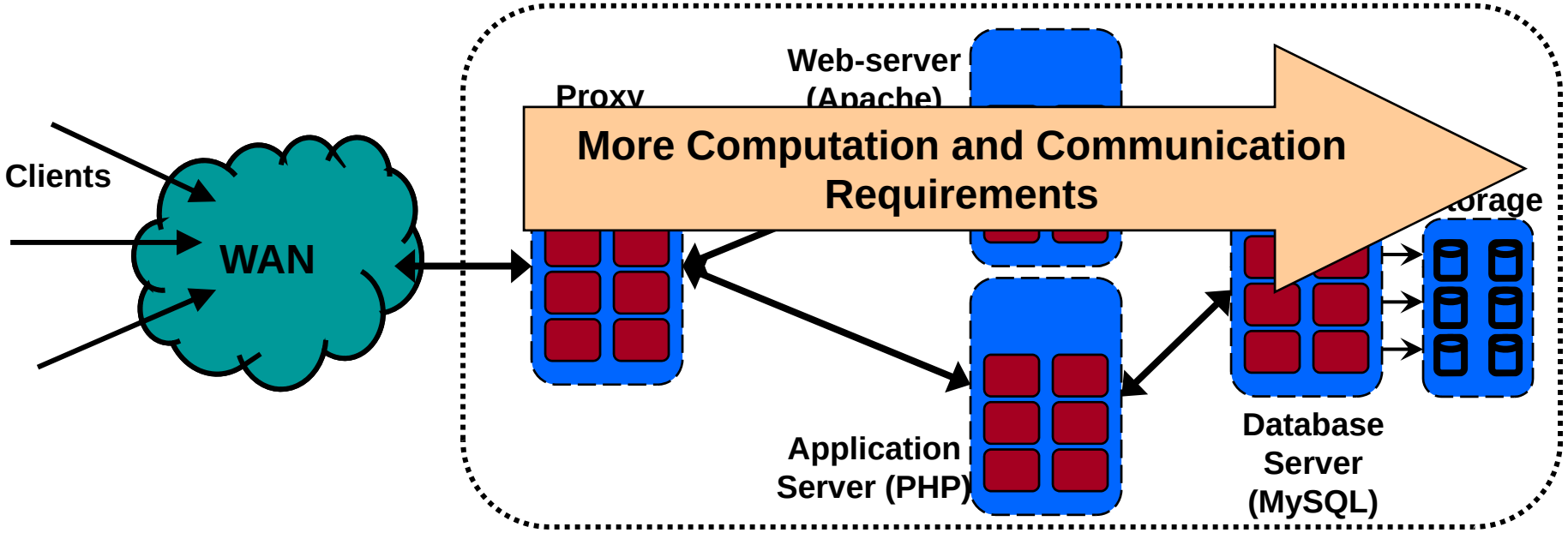
- IOzone Read Bandwidth up to 913 MB/s (Sun x2200's with x8 PCIe)
- Read-Write design by OSU, available with the latest OpenSolaris
- NFS/RDMA will also be added in OFED 1.4

R. Noronha, L. Chai, T. Talpey and D. K. Panda, "Designing NFS With RDMA For Security, Performance and Scalability", ICPP '07

Sample Case Studies

- Message Passing Interface (MPI)
- SDP and IPoIB
- File Systems (Lustre, NFS-RDMA)
- Datacenter
- Virtualization

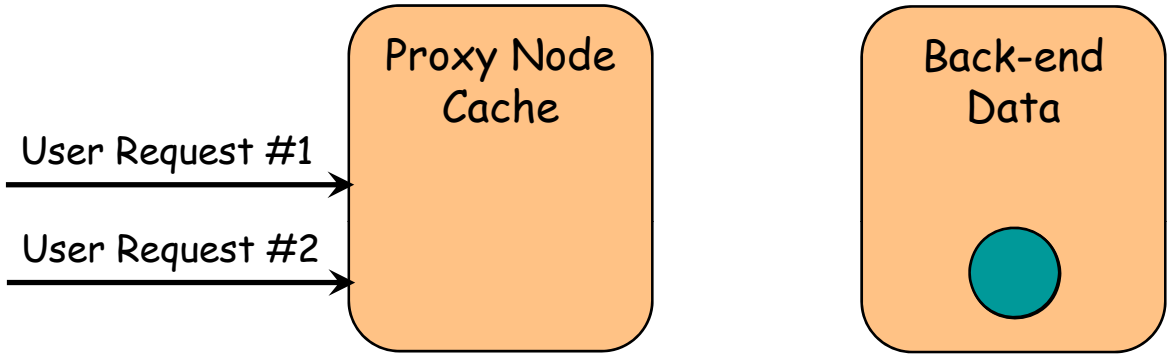
Typical Multi-Tier Datacenter



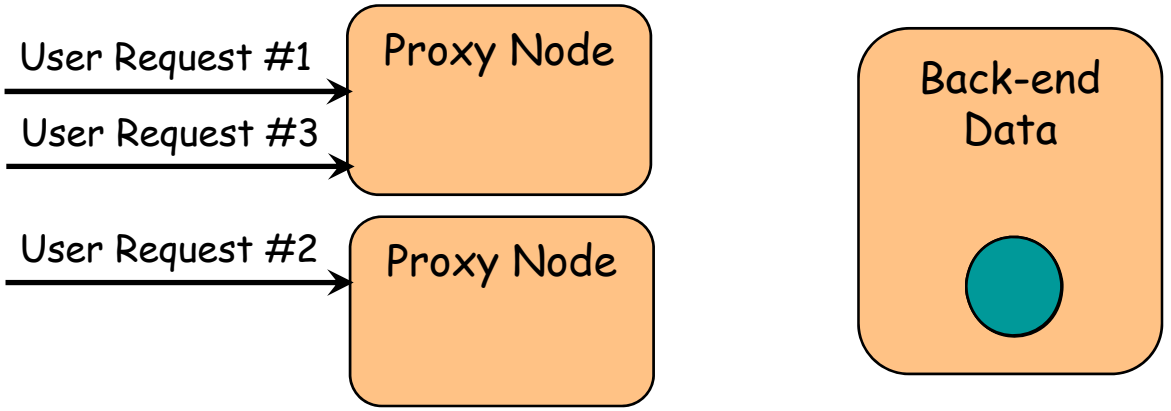
- Requests are received from clients over the WAN
- Proxy nodes perform caching, load balancing, resource monitoring, etc.
  - If not cached, request forwarded to the next tier → Application Server
- Application server performs the business logic (CGI, Java servlets)
  - Retrieves appropriate data from the database to process the requests

Cache Coherency and Consistency with Dynamic Data

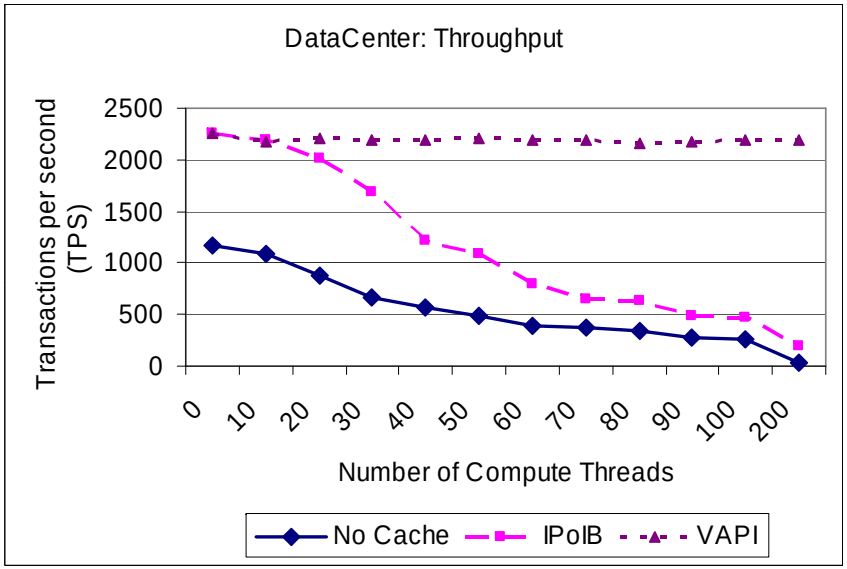
Example of Strong Cache Coherency: Never Send Stale Data

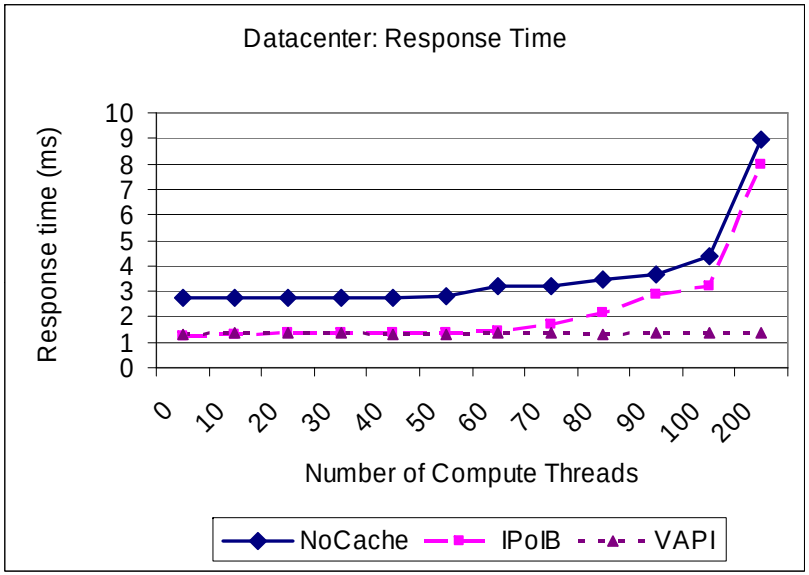


Example of Strong Cache Consistency:
Always Follow Increasing Time Line of Events



Strong Cache Coherency with RDMA





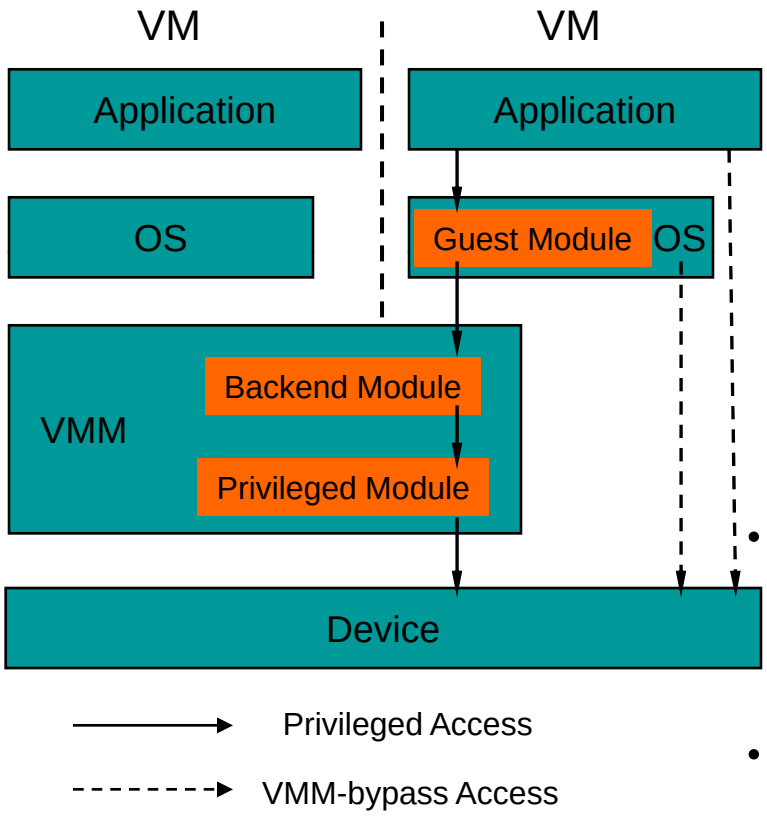
RDMA can sustain performance even with heavy load on the back-end

S. Narravula, P. Balaji, K. Vaidyanathan, S. Krishnamoorthy, J. Wu, and D. K. Panda, "Supporting Strong Cache Coherency for Active Caches in Multi-Tier Data-Centers over InfiniBand", SAN '04

Sample Case Studies

- Message Passing Interface (MPI)
- SDP and IPoIB
- File Systems (Lustre, NFS-RDMA)
- Datacenter
- Virtualization

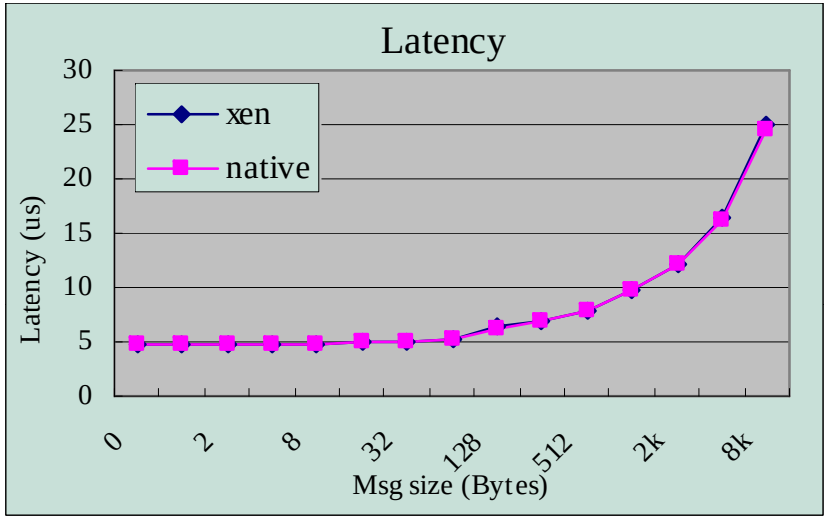
From OS-bypass to VMM-bypass

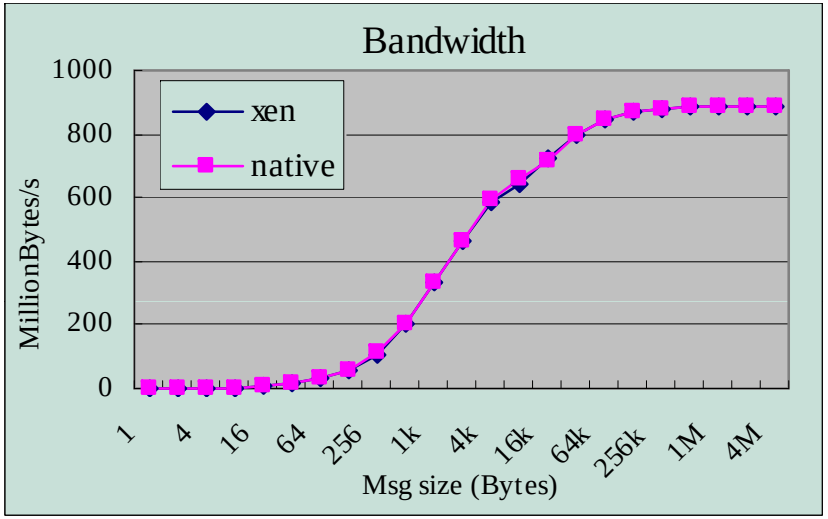


- Guest modules in guest VMs handle setup and management operations (privileged access)
  - Guest modules communicate with VMM backend modules to get jobs done
  - Original privileged module can be reused
- Once setup, devices are accessed directly from guest VMs (VMM-bypass)
  - Either from OS kernel or applications
- Backend and privileged modules can also reside in a special VM

J. Liu, W. Huang, B. Abali, D. K. Panda. "High Performance VMM-Bypass I/O in Virtual Machines", USENIX '06

MPI Latency and Bandwidth





- Only VMM Bypass operations are used
- Xen-IB performs similar to native InfiniBand
- Numbers taken with MVAPICH

– J. Liu, W. Huang, B. Abali, D. K. Panda. "High Performance VMM-Bypass I/O in Virtual Machines", USENIX '06

– W. Huang, J. Liu, B. Abali, D. K. Panda. "A Case for High Performance Computing with Virtual Machines", ICS '06

Summary of Design Performance Results

- Current generation IB adapters, 10GE/iWARP adapters and software environments are already delivering competitive performance
- IB and 10GE/iWARP hardware, firmware, and software are going through rapid changes
- Significant performance improvement is expected in near future

Presentation Overview

- Introduction
- Why InfiniBand and 10-Gigabit Ethernet?
- Overview of IB and 10GE and their Novel Features
- IB and 10GE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
  - MPI, SDP, File Systems, Data Center and Virtualization
- Conclusions and Final Q&A

Concluding Remarks

- Presented network architectures & trends in Clusters
- Presented background and details of IB and 10GE
  - Highlighted differences with other technologies
  - Gave an overview of IB and 10GE hardware/software products
  - Discussed sample performance numbers in designing various high-end systems with IB and 10GE
- IB and 10GE are emerging as new architectures leading to a new generation of networked computing systems, opening many research issues needing novel solutions

Funding Acknowledgments

Our research is supported by the following organizations

- Current Funding support by

- Current Equipment support by

Personnel Acknowledgments

Current Students

- L. Chai (Ph.D.)
- T. Gangadharappa (M. S.)
- K. Gopalakrishnan (M. S.)
- M. Koop (Ph.D.)
- P. Lai (Ph. D.)
- G. Marsh (Ph. D.)
- X. Ouyang (Ph.D.)
- G. Santhanaraman (Ph.D.)
- J. Sridhar (M. S.)
- H. Subramoni (M. S.)

Current Programmer

- J. Perkins

Past Students

- P. Balaji (Ph.D.)
- D. Buntinas (Ph.D.)
- S. Bhagvat (M.S.)
- B. Chandrasekharan (M.S.)
- W. Jiang (M.S.)
- W. Huang (Ph.D.)
- S. Kini (M.S.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- J. Liu (Ph.D.)
- A. Mamidala (Ph.D.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- S. Sur (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- R. Noronha (Ph.D.)
- S. Sur (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)

Web Pointers

http://www.cse.ohio-state.edu/~panda

http://www.mcs.anl.gov/~balaji

http://www.cse.ohio-state.edu/~koop

http://nowlab.cse.ohio-state.edu

MVAPICH Web Page

http://mvapich.cse.ohio-state.edu

panda@cse.ohio-state.edu

balaji@mcs.anl.gov

koop@cse.ohio-state.edu