

# InfiniBand and High-Speed Ethernet for Dummies

A Tutorial at ISC '13

by

**Dhabaleswar K. (DK) Panda**

The Ohio State University

E-mail: [panda@cse.ohio-state.edu](mailto:panda@cse.ohio-state.edu)

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

**Hari Subramoni**

The Ohio State University

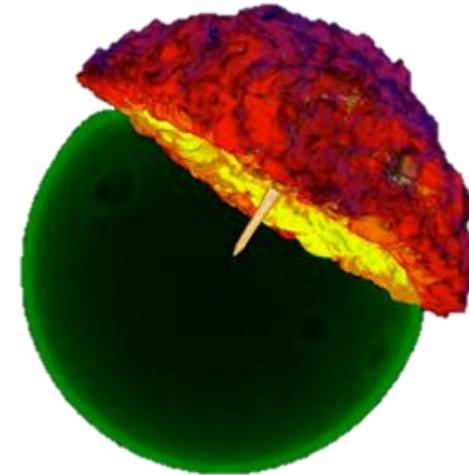
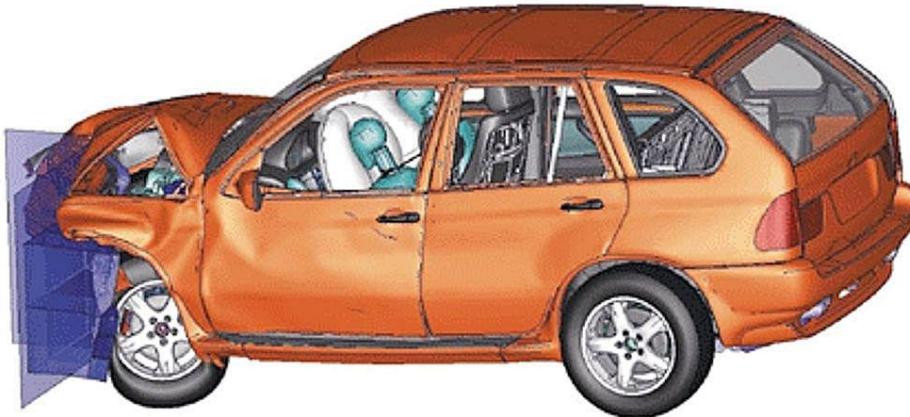
E-mail: [subramon@cse.ohio-state.edu](mailto:subramon@cse.ohio-state.edu)

<http://www.cse.ohio-state.edu/~subramon>

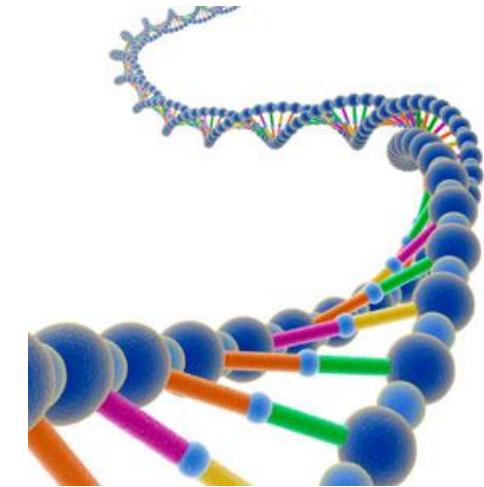
# Presentation Overview

- **Introduction**
- Why InfiniBand and High-speed Ethernet?
- Overview of IB, HSE, their Convergence and Features
- IB and HSE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
- Conclusions and Final Q&A

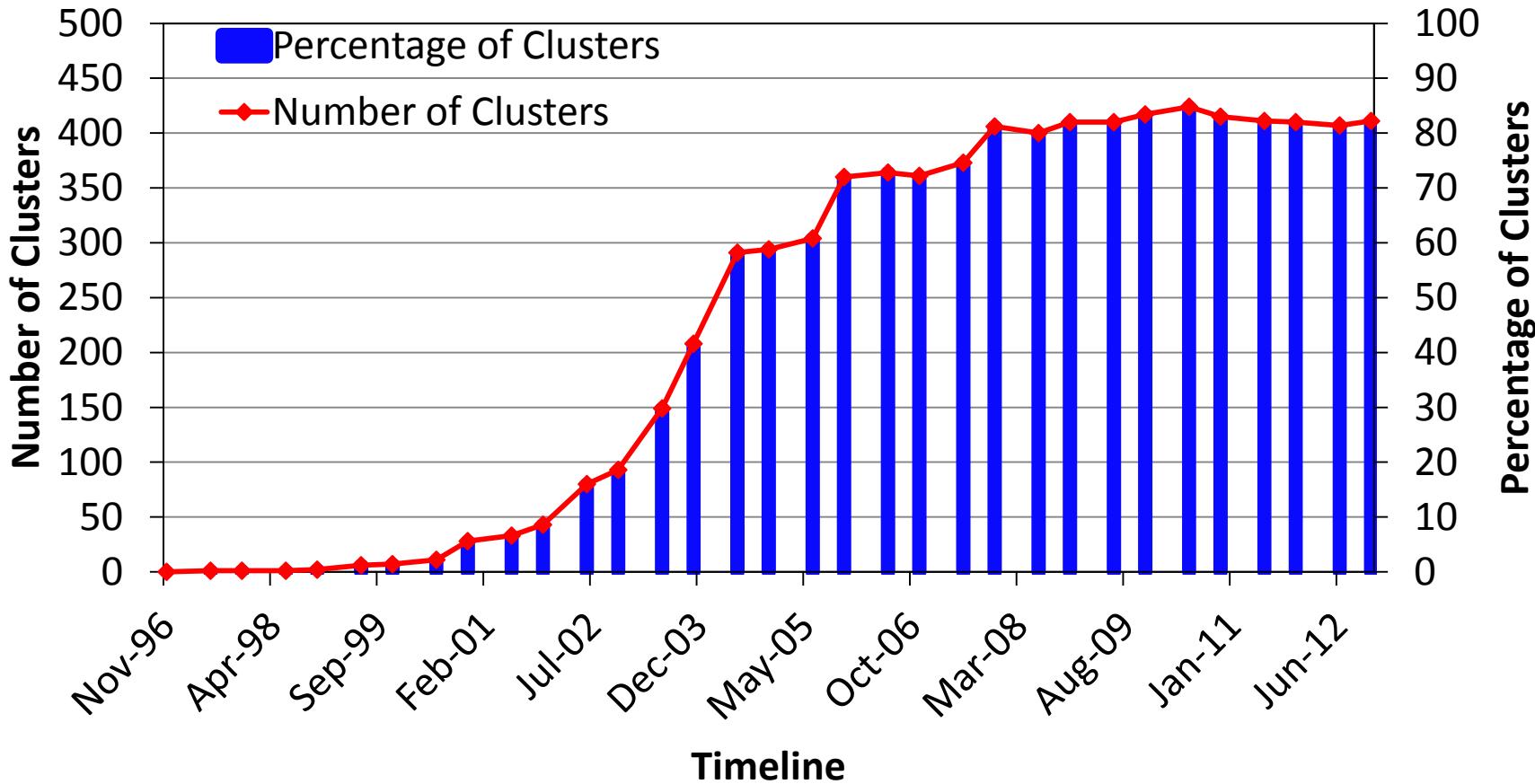
# Current and Next Generation Applications and Computing Systems



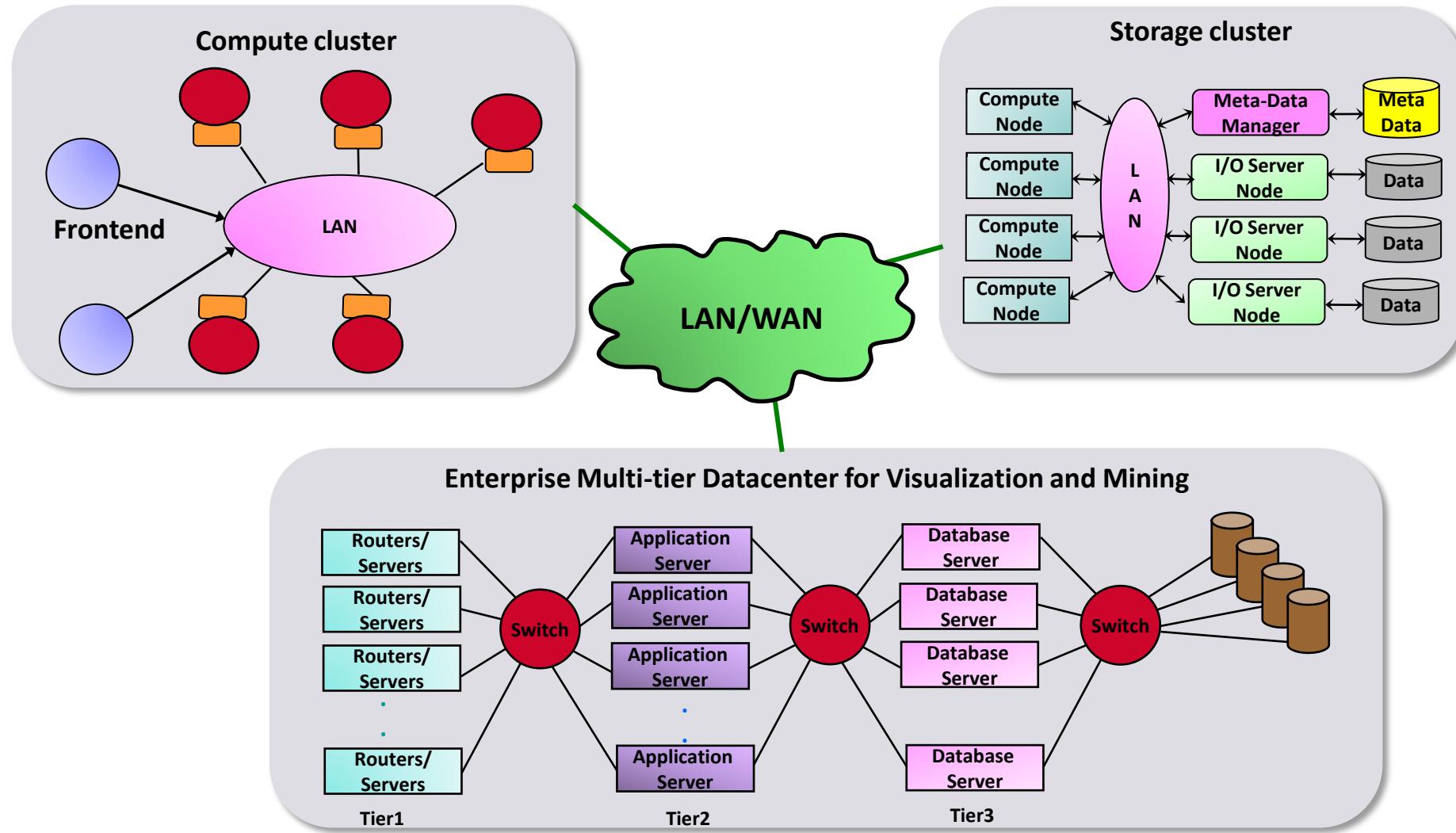
- Growth of High Performance Computing
  - Growth in processor performance
    - Chip density doubles every 18 months
  - Growth in commodity networking
    - Increase in speed/features + reducing cost
- Clusters: popular choice for HPC
  - Scalability, Modularity and Upgradeability



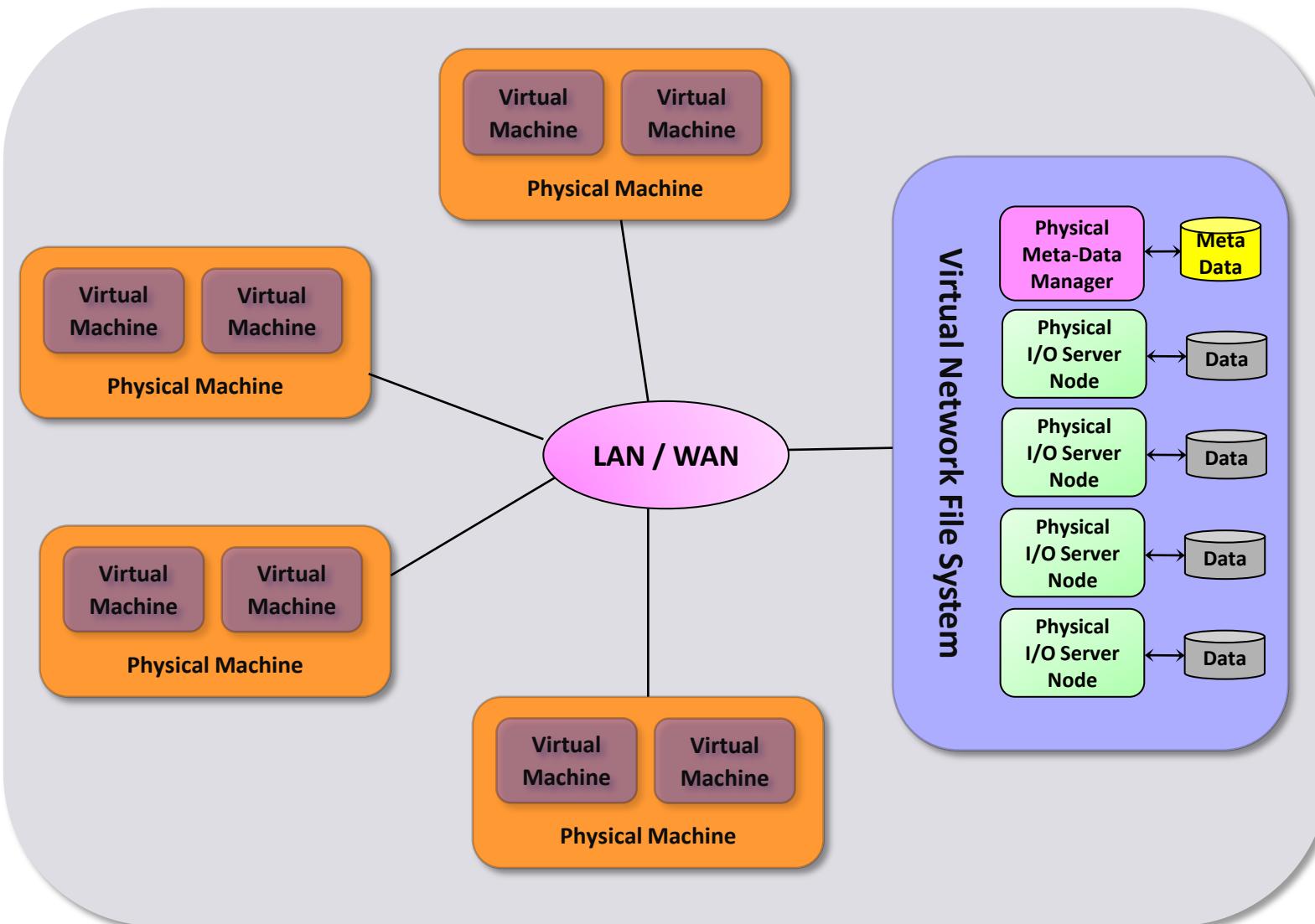
# Trends for Commodity Computing Clusters in the Top 500 List (<http://www.top500.org>)



# Integrated High-End Computing Environments

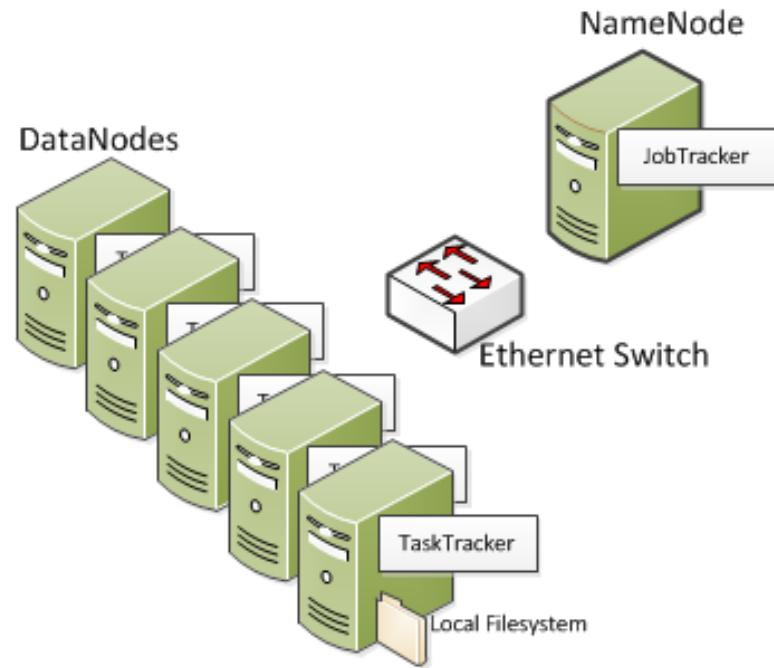


# Cloud Computing Environments



# Big Data Analytics with Hadoop

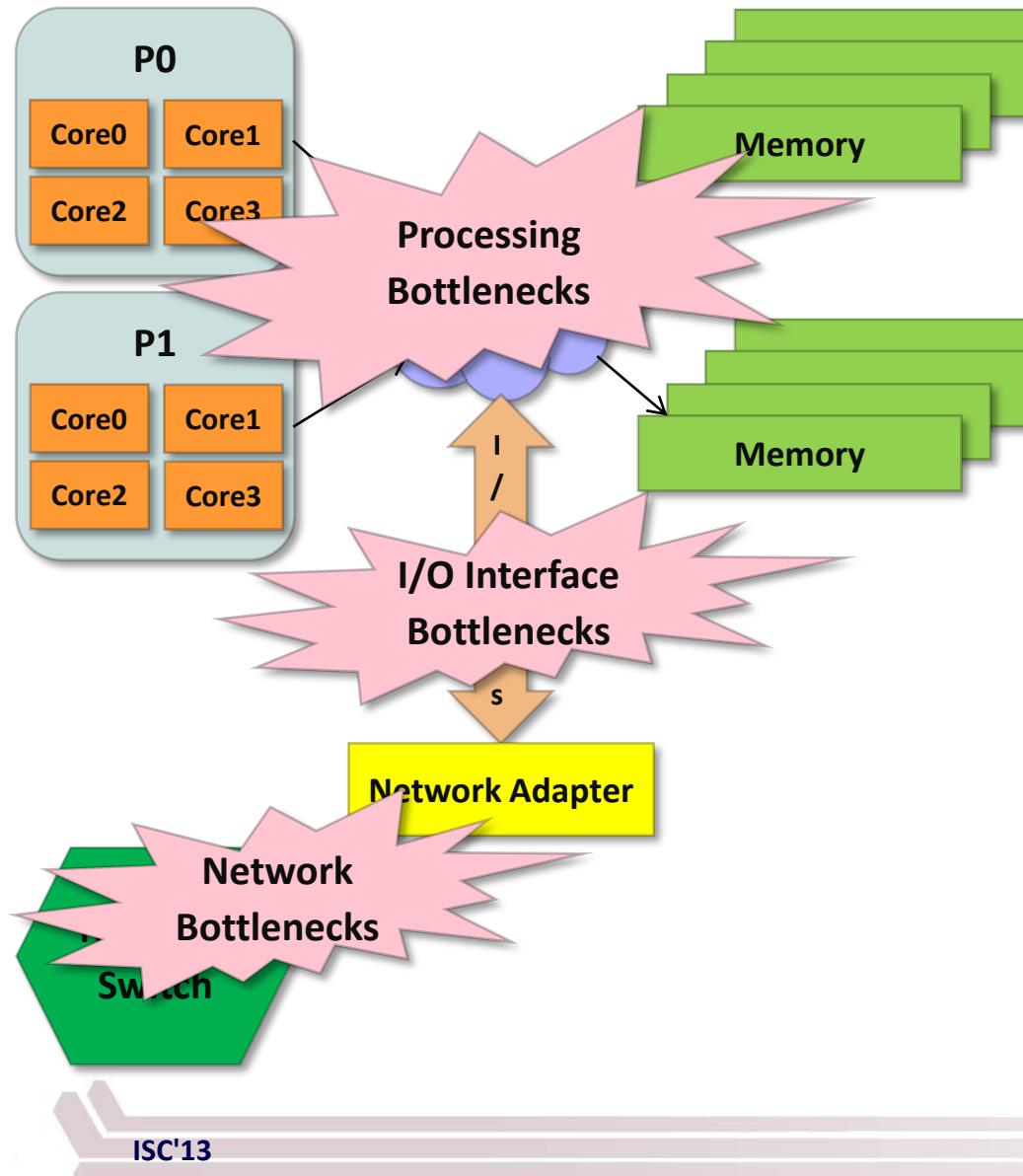
- Underlying Hadoop Distributed File System (HDFS)
- Fault-tolerance by replicating data blocks
- NameNode: stores information on data blocks
- DataNodes: store blocks and host Map-reduce computation
- JobTracker: track jobs and detect failure
- MapReduce (Distributed Computation)
- HBase (Database component)
- Model scales but high amount of communication during intermediate phases



# Networking and I/O Requirements

- Good System Area Networks with excellent performance (low latency, high bandwidth and low CPU utilization) 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

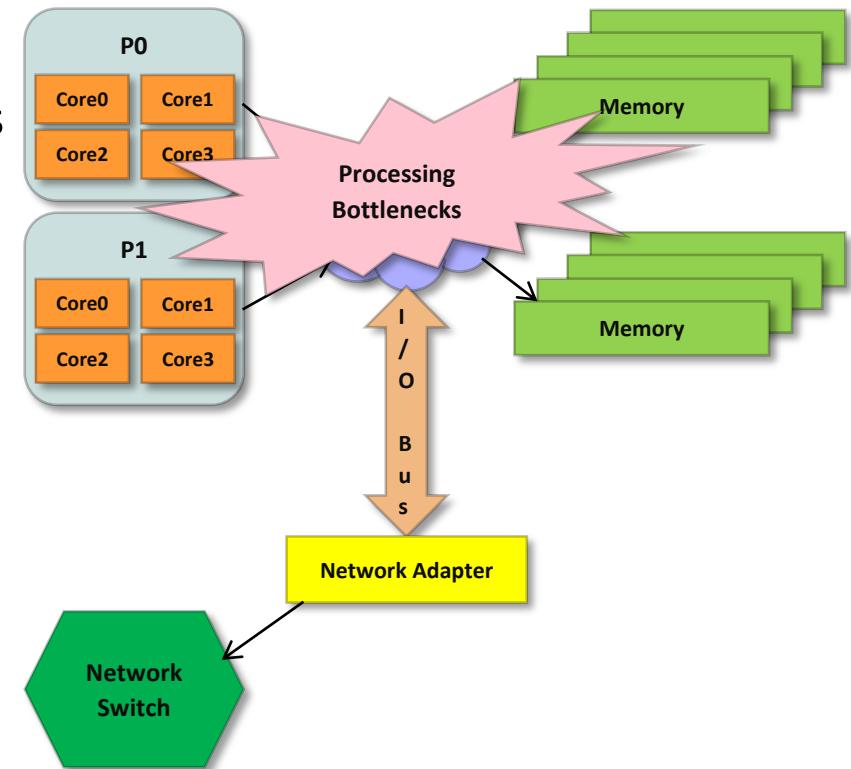
# Major Components in Computing Systems



- Hardware components
  - Processing cores and memory subsystem
  - I/O bus or links
  - Network adapters/switches
- Software components
  - Communication stack
- *Bottlenecks can artificially limit the network performance the user perceives*

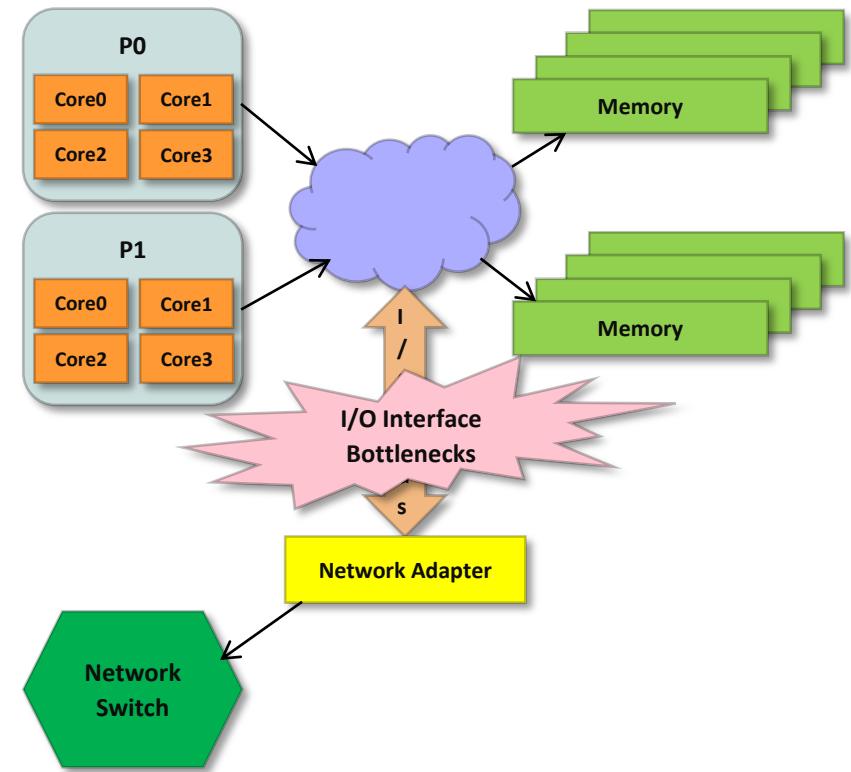
# Processing Bottlenecks in Traditional Protocols

- Ex: TCP/IP, UDP/IP
- Generic architecture for all networks
- Host processor 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 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

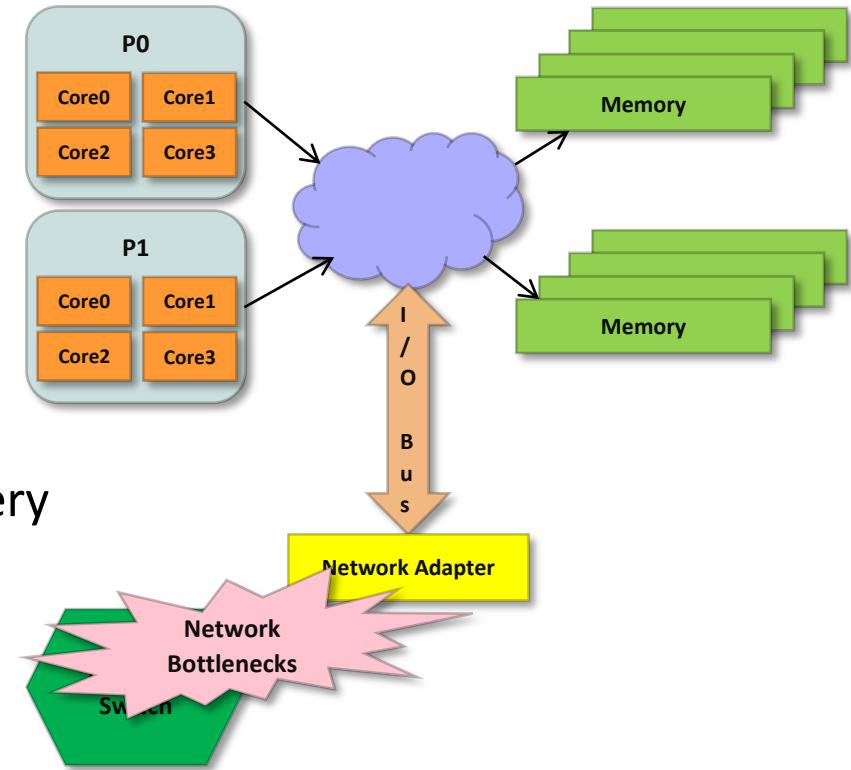
- Traditionally relied on bus-based technologies (last mile bottleneck)
  - E.g., PCI, PCI-X
  - One bit per wire
  - Performance increase through:
    - Increasing clock speed
    - Increasing bus width
  - Not scalable:
    - Cross talk between bits
    - Skew between wires
    - Signal integrity makes it difficult to increase bus width significantly, especially for high clock 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)

# 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

# Motivation for InfiniBand and High-speed Ethernet

- Industry Networking 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, and 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

# Presentation Overview

- Introduction
- **Why InfiniBand and High-speed Ethernet?**
- Overview of IB, HSE, their Convergence and Features
- IB and HSE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
- Conclusions and Final Q&A

## 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
  - Several annexes released after that (RDMA\_CM - Sep'06, iSER – Sep'06, XRC – Mar'09, RoCE – Apr'10)
- <http://www.infinibandta.org>

## High-speed Ethernet Consortium (10GE/40GE/100GE)

- 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>
- 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

# Tackling Communication Bottlenecks with IB and HSE

- **Network speed bottlenecks**
- Protocol processing bottlenecks
- I/O interface bottlenecks

# Network Bottleneck Alleviation: InfiniBand (“Infinite Bandwidth”) and High-speed Ethernet (10/40/100 GE)

- Bit serial differential signaling
  - Independent pairs of wires to transmit independent data (called a lane)
  - Scalable to any number of lanes
  - Easy to increase clock speed of lanes (since each lane consists only of a pair of wires)
- Theoretically, no perceived limit on the bandwidth

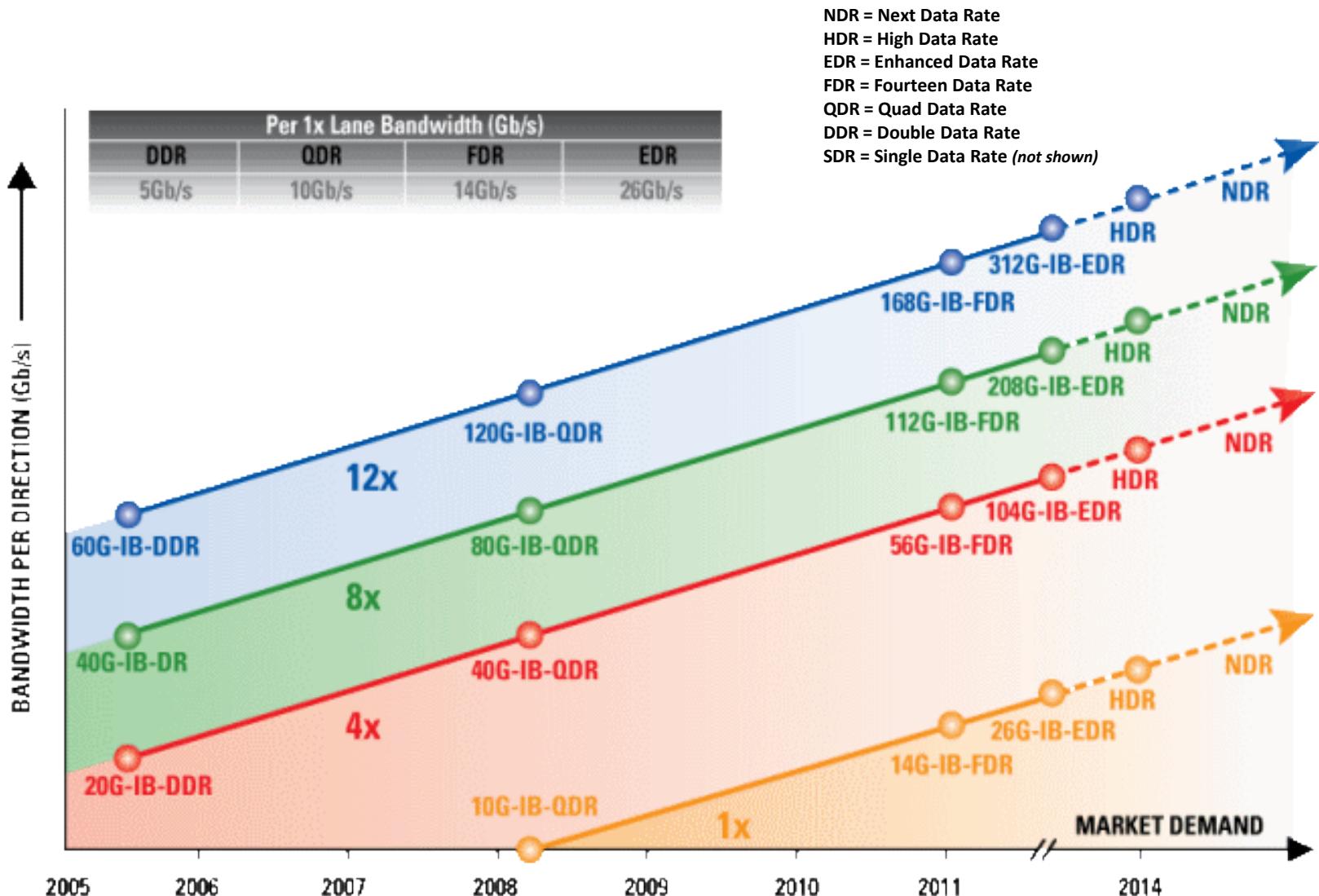


# Network Speed Acceleration with IB and HSE

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 -)	54.6 Gbit/sec (4X FDR)
InfiniBand (2012 -)	2 x 54.6 Gbit/sec (4X Dual-FDR)
InfiniBand (2013?)	100 Gbit/sec (4X EDR)

*50 times in the last 12 years*

# InfiniBand Link Speed Standardization Roadmap



# Tackling Communication Bottlenecks with IB and HSE

- Network speed bottlenecks
- **Protocol processing bottlenecks**
- I/O interface bottlenecks

# Capabilities of 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

- 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
- Virtual Interface Architecture (VIA)
  - Standardized by Intel, Compaq, Microsoft
  - Precursor to IB

# IB Hardware Acceleration

- Some IB models have multiple hardware accelerators
  - E.g., Mellanox IB adapters
- Protocol Offload Engines
  - Completely implement ISO/OSI layers 2-4 (link layer, network layer and transport layer) in hardware
- Additional hardware supported features also present
  - RDMA, Multicast, QoS, Fault Tolerance, and many more

# 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 (a.k.a. Virtual MTU)
  - Host processor “thinks” that the adapter supports large Jumbo frames, but the adapter splits it into regular sized (1500-byte) frames
  - Supported by most HSE products because of its backward compatibility → considered “regular” Ethernet

# 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, usually referred to as the entire network adapter
- Internet Wide-Area RDMA Protocol (iWARP)
  - Standardized by IETF and the RDMA Consortium
  - Supports acceleration features (like IB) for Ethernet
- <http://www.ietf.org> & <http://www.rdmaconsortium.org>

# Converged (Enhanced) Ethernet (CEE or CE)

- Also known as “Datacenter Ethernet” or “Lossless Ethernet”
  - Combines a number of optional Ethernet standards into one umbrella as mandatory requirements
- Sample enhancements include:
  - Priority-based flow-control: Link-level flow control for each Class of Service (CoS)
  - Enhanced Transmission Selection (ETS): Bandwidth assignment to each CoS
  - Datacenter Bridging Exchange Protocols (DBX): Congestion notification, Priority classes
  - End-to-end Congestion notification: Per flow congestion control to supplement per link flow control

# Tackling Communication Bottlenecks with IB and HSE

- Network speed bottlenecks
- Protocol processing bottlenecks
- **I/O interface bottlenecks**

# Interplay with I/O Technologies

- InfiniBand initially intended to replace I/O bus technologies with networking-like technology
  - That is, bit serial differential signaling
  - With enhancements in I/O technologies that use a similar architecture (HyperTransport, PCI Express), this has become mostly irrelevant now
- Both IB and HSE today come as network adapters that plug into existing I/O technologies

# Trends in I/O Interfaces with Servers

- Recent trends in I/O interfaces show that they are nearly matching head-to-head with network speeds (though they still lag a little bit)

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)

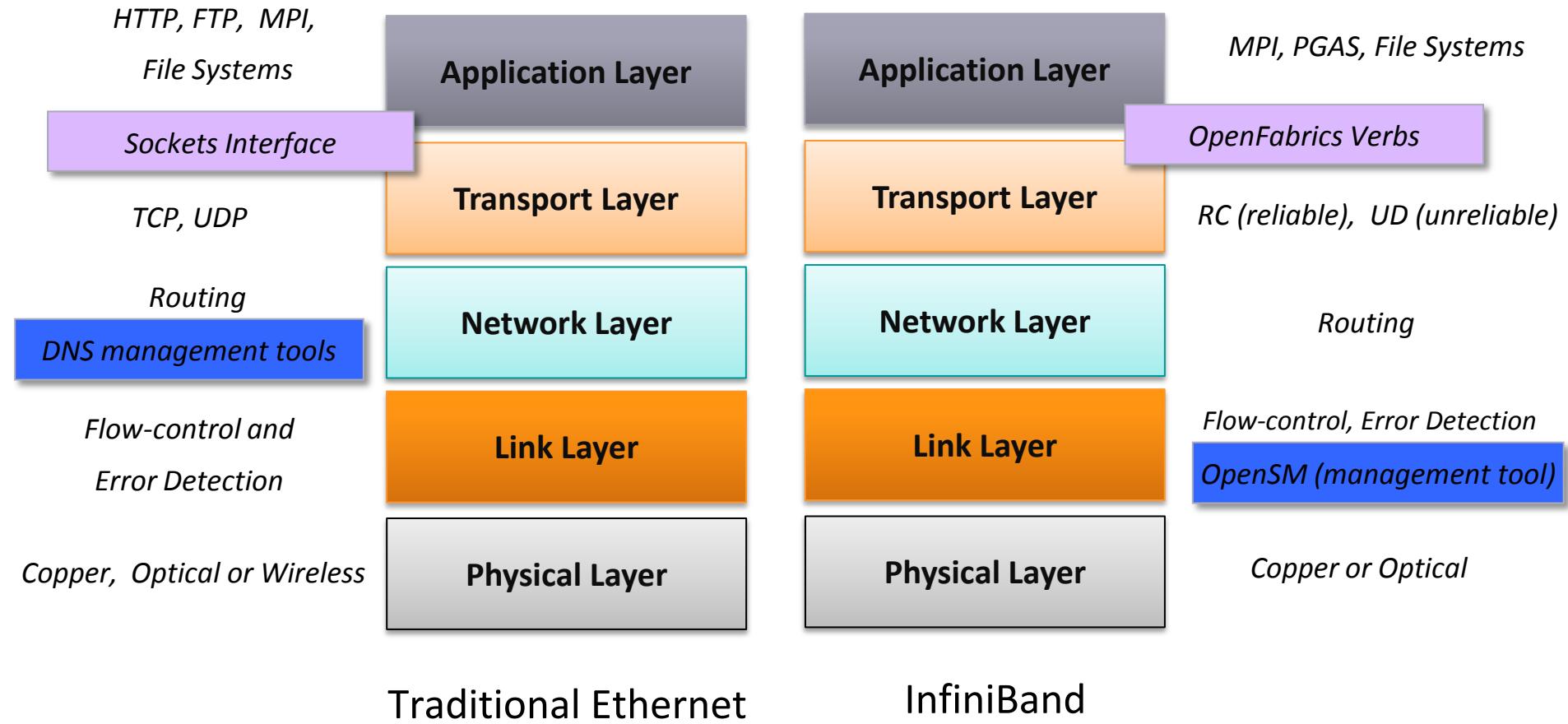
# Presentation Overview

- Introduction
- Why InfiniBand and High-speed Ethernet?
- **Overview of IB, HSE, their Convergence and Features**
- IB and HSE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
- Conclusions and Final Q&A

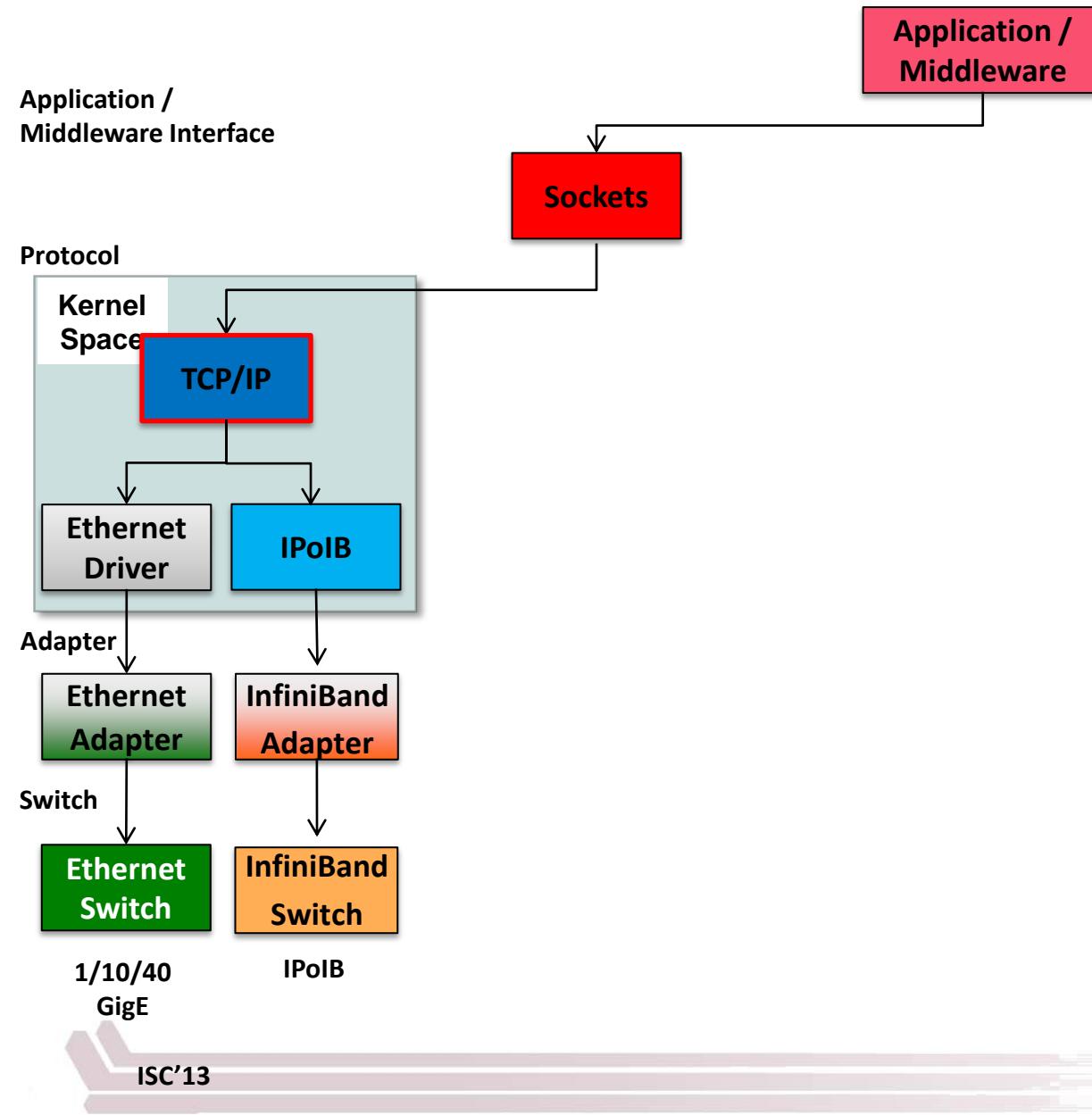
# IB, HSE and their Convergence

- InfiniBand
  - Architecture and Basic Hardware Components
  - Communication Model and Semantics
  - Novel Features
  - Subnet Management and Services
- High-speed Ethernet Family
  - Internet Wide Area RDMA Protocol (iWARP)
  - Alternate vendor-specific protocol stacks
- InfiniBand/Ethernet Convergence Technologies
  - Virtual Protocol Interconnect (VPI)
  - (InfiniBand) RDMA over Converged (Enhanced) Ethernet (RoCE)

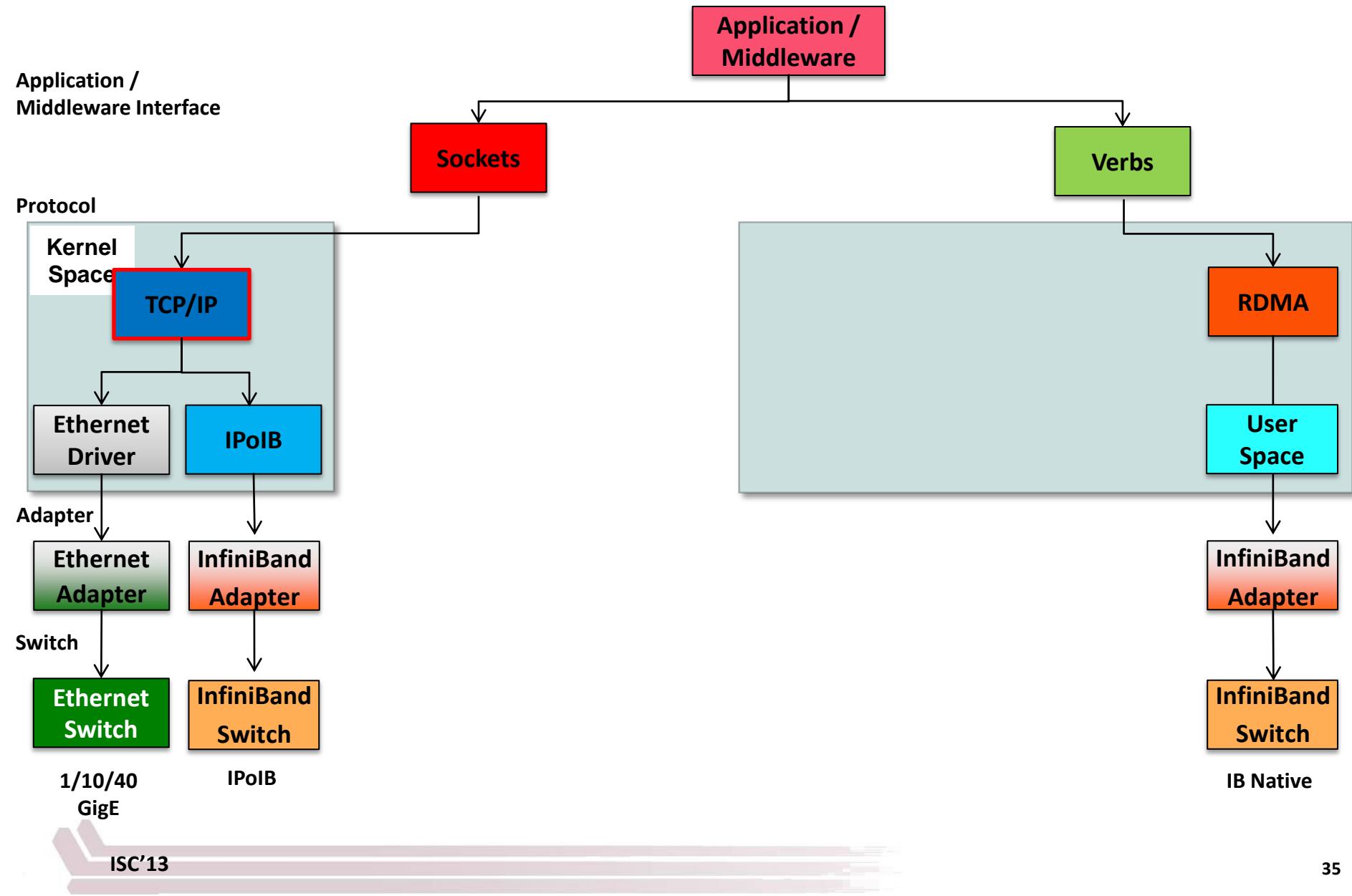
# Comparing InfiniBand with Traditional Networking Stack



# TCP/IP Stack and IPoIB



# TCP/IP, IPoIB and Native IB Verbs

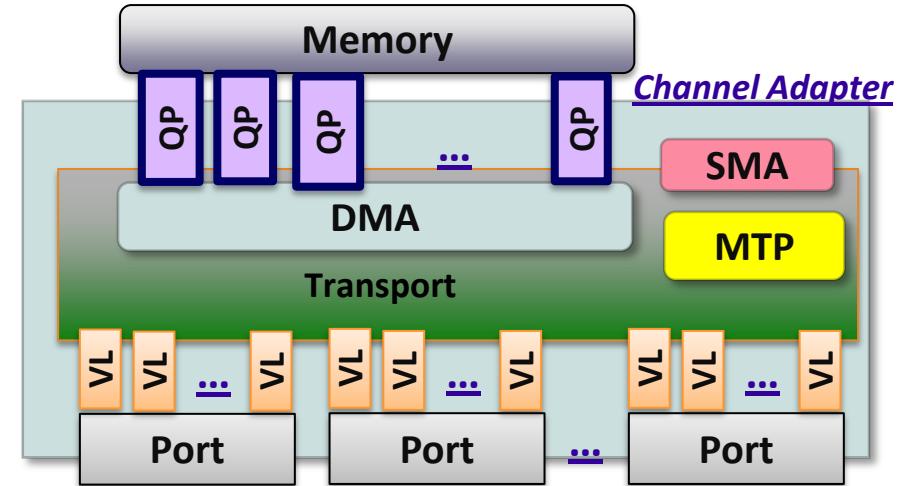


# IB Overview

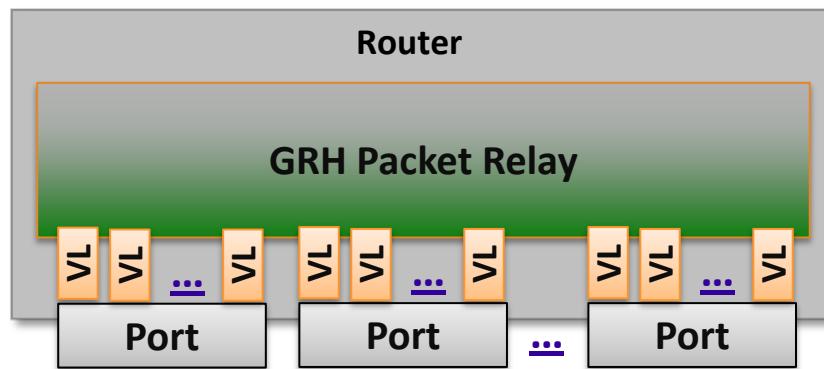
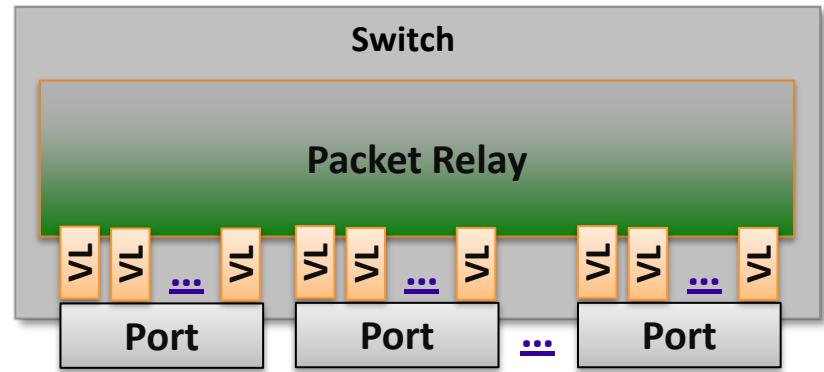
- **InfiniBand**
  - **Architecture and Basic Hardware Components**
  - Communication Model and Semantics
    - Communication Model
    - Memory registration and protection
    - Channel and memory semantics
  - Novel Features
    - Hardware Protocol Offload
      - Link, network and transport layer features
    - Subnet Management and Services
    - Sockets Direct Protocol (SDP) stack
    - RSocket Protocol Stack

# Components: Channel Adapters

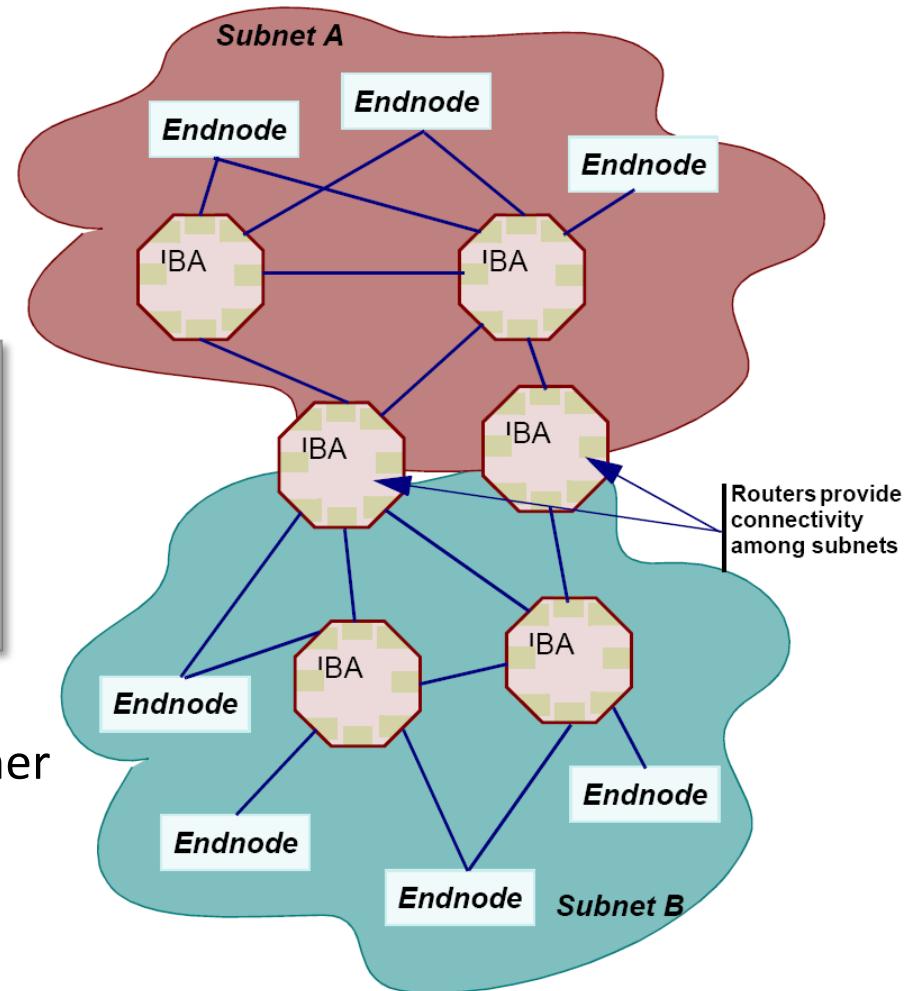
- Used by processing and I/O units to connect to fabric
- Consume & generate IB packets
- Programmable DMA engines with protection features
- May have multiple ports
  - Independent buffering channeled through Virtual Lanes
- Host Channel Adapters (HCAs)



# 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)
  - For example, QDR copper cables are restricted to 7m
- 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)



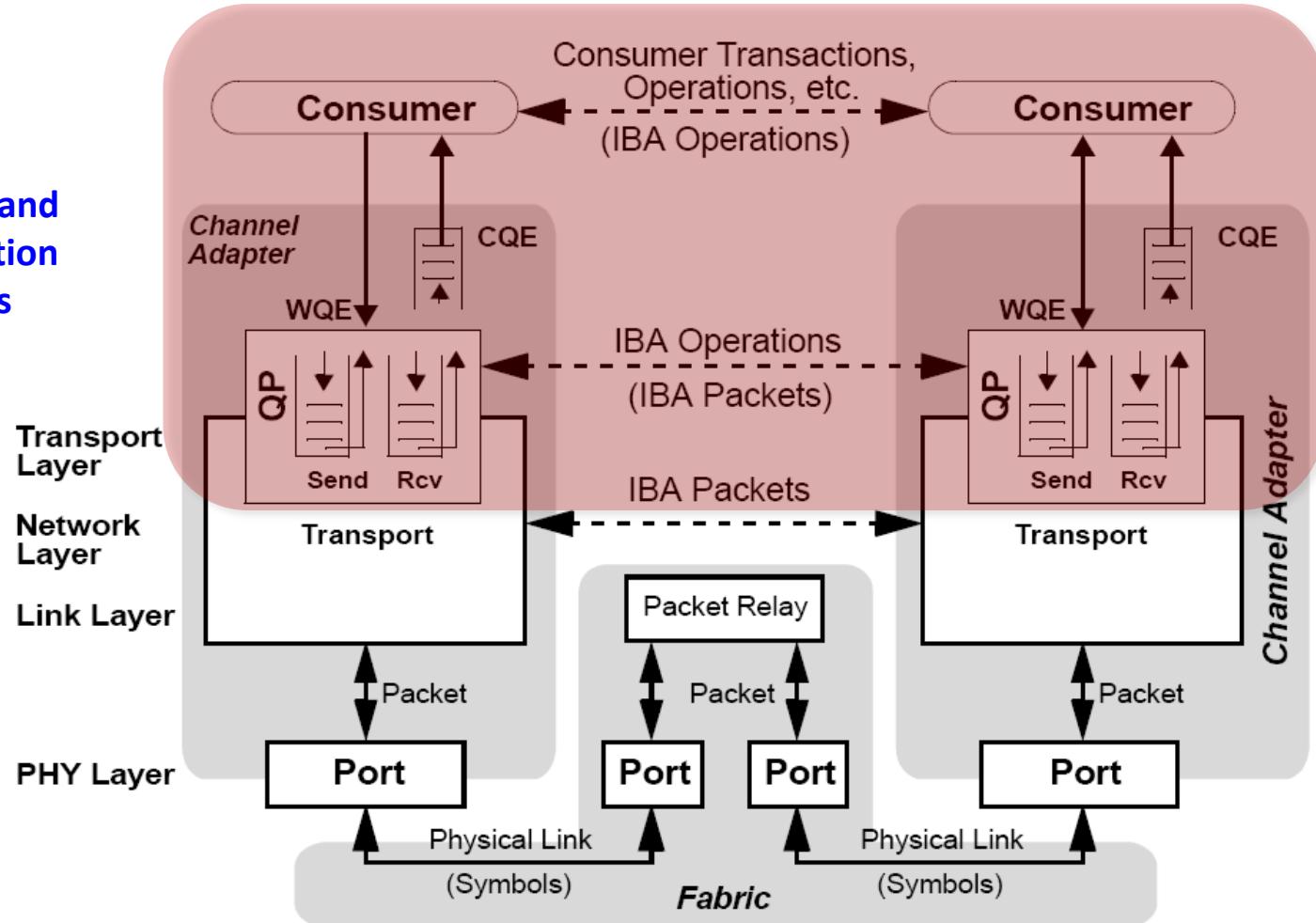
(Courtesy Intel)

# IB Overview

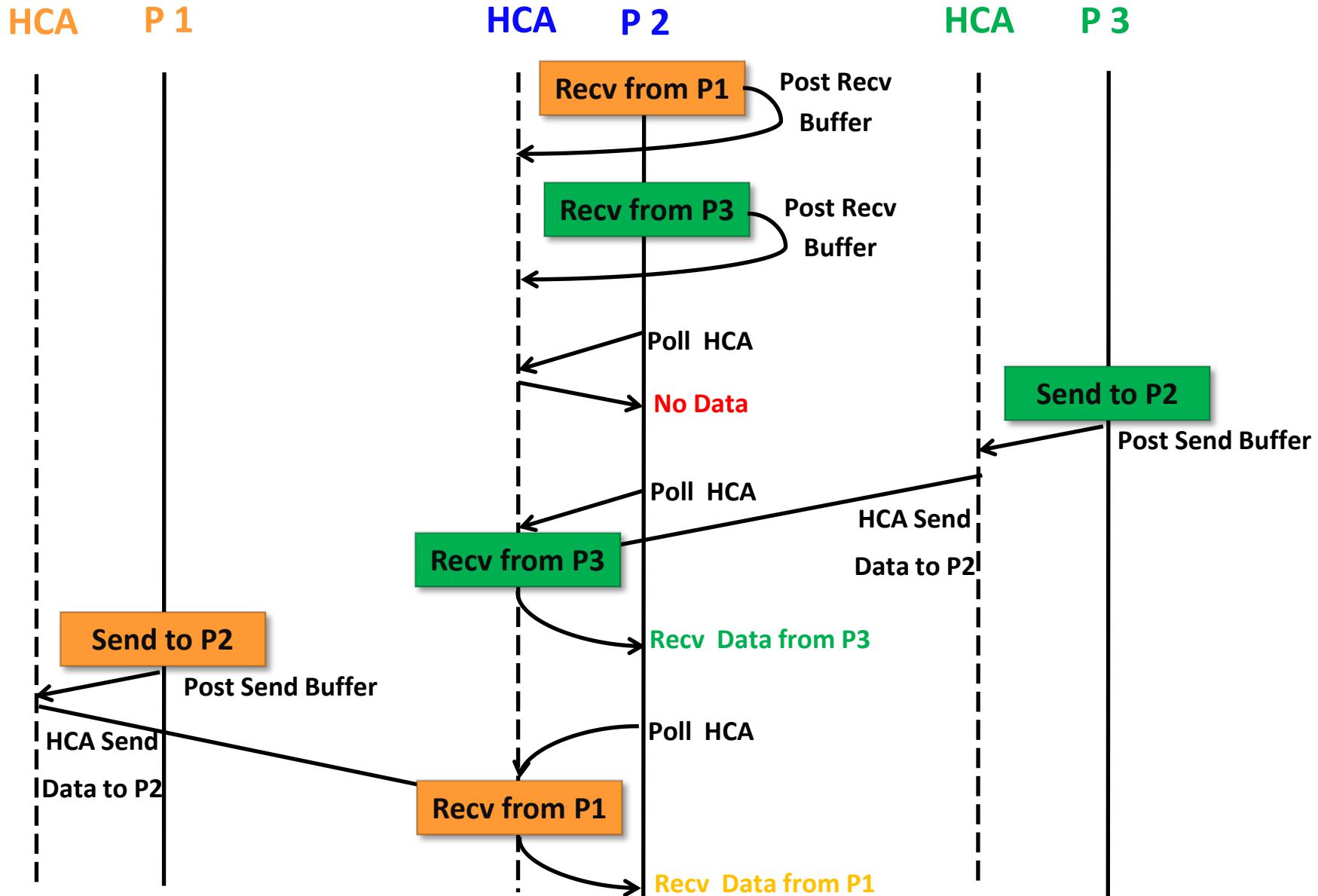
- **InfiniBand**
  - Architecture and Basic Hardware Components
  - **Communication Model and Semantics**
    - **Communication Model**
    - **Memory registration and protection**
    - **Channel and memory semantics**
  - Novel Features
    - Hardware Protocol Offload
      - Link, network and transport layer features
    - Subnet Management and Services
    - Sockets Direct Protocol (SDP) stack
    - RSocket Protocol Stack

# IB Communication Model

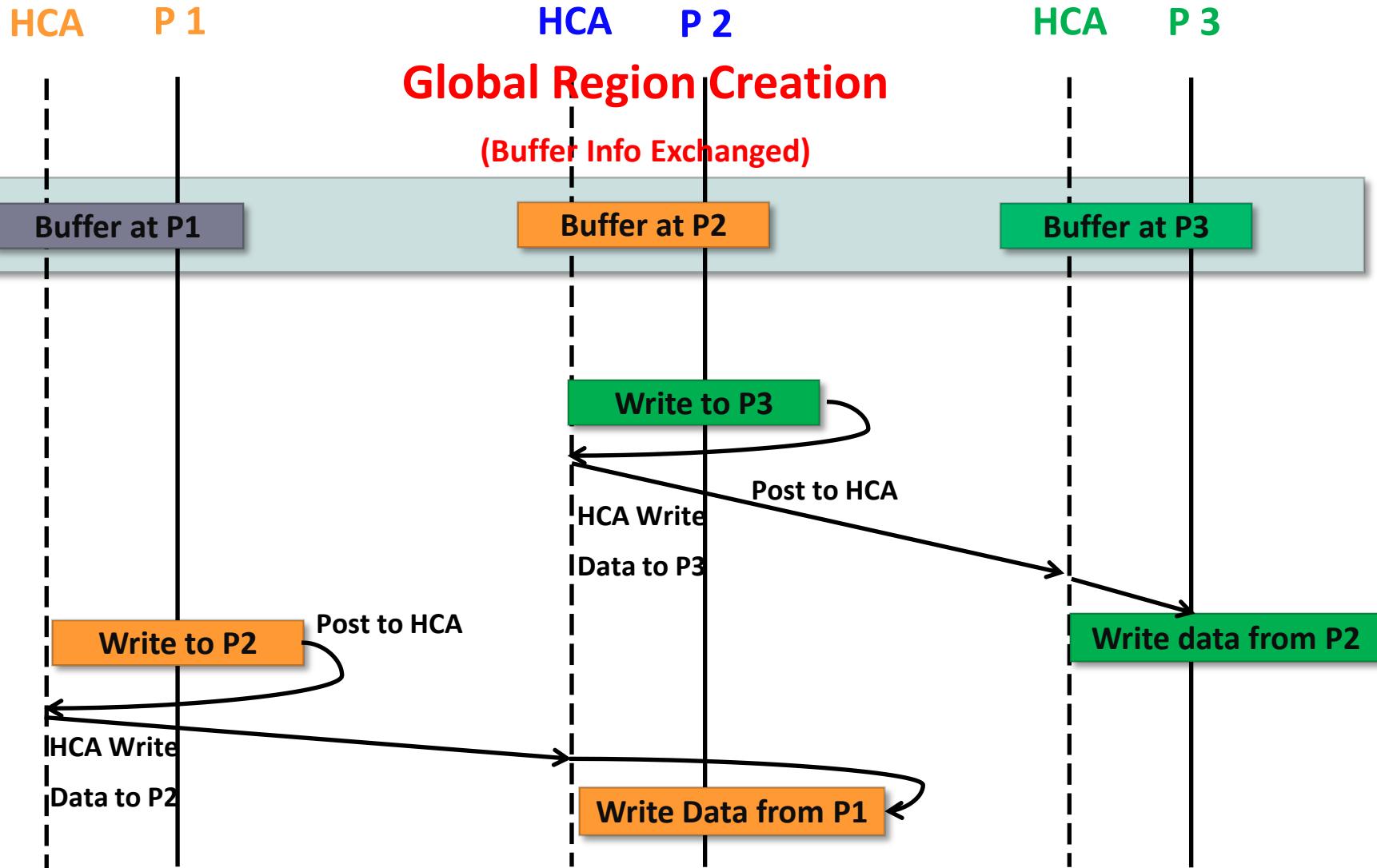
## Basic InfiniBand Communication Semantics



## Two-sided Communication Model

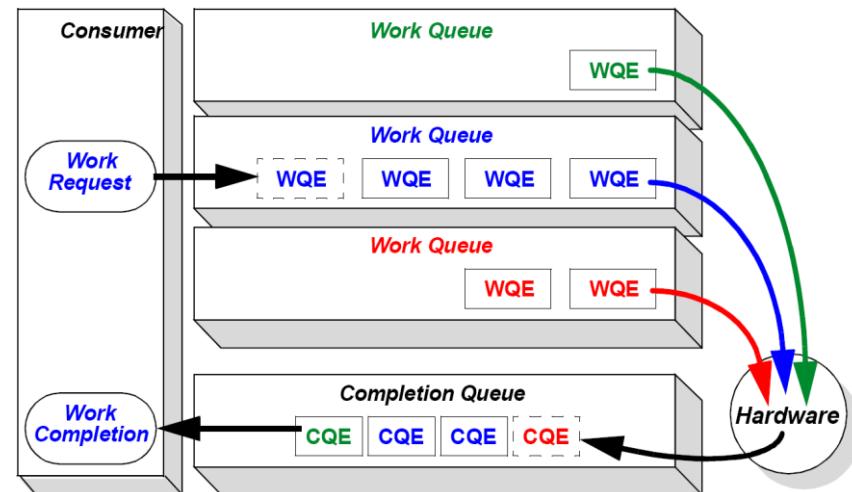
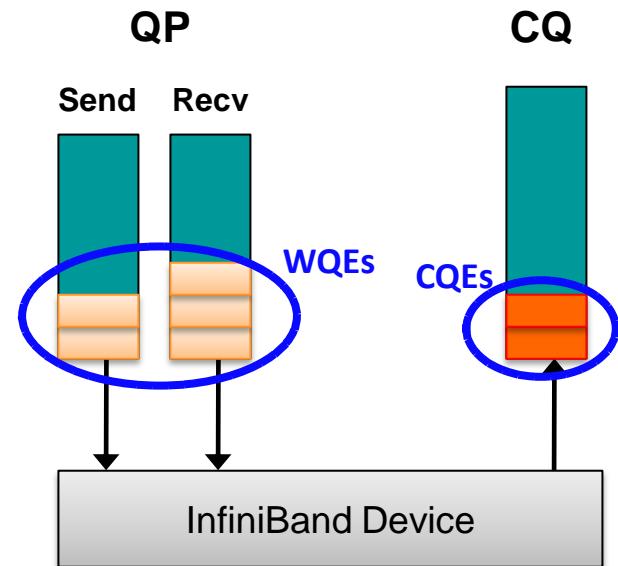


# One-sided Communication Model



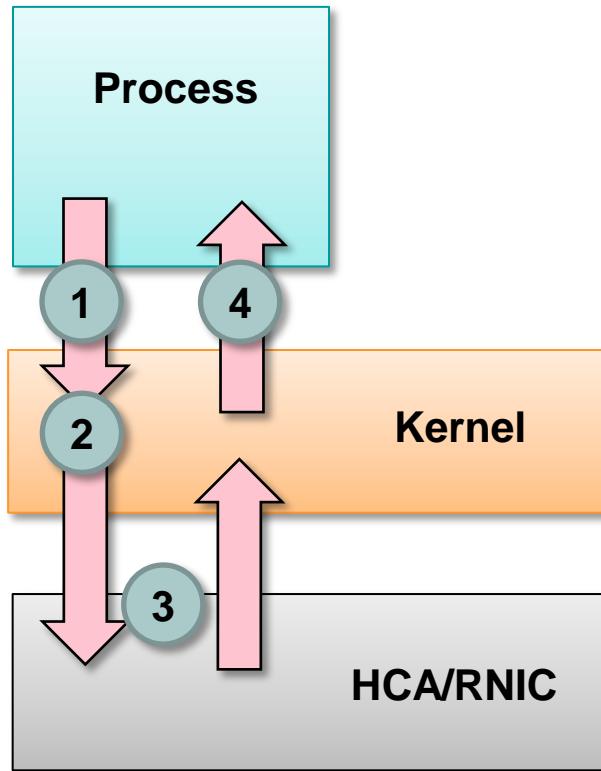
# Queue Pair Model

- Each QP has two queues
  - Send Queue (SQ)
  - Receive Queue (RQ)
  - Work requests are queued to the QP (WQEs: “Wookies”)
- QP to be linked to a Complete Queue (CQ)
  - Gives notification of operation completion from QPs
  - Completed WQEs are placed in the CQ with additional information (CQEs: “Cookies”)



# Memory Registration

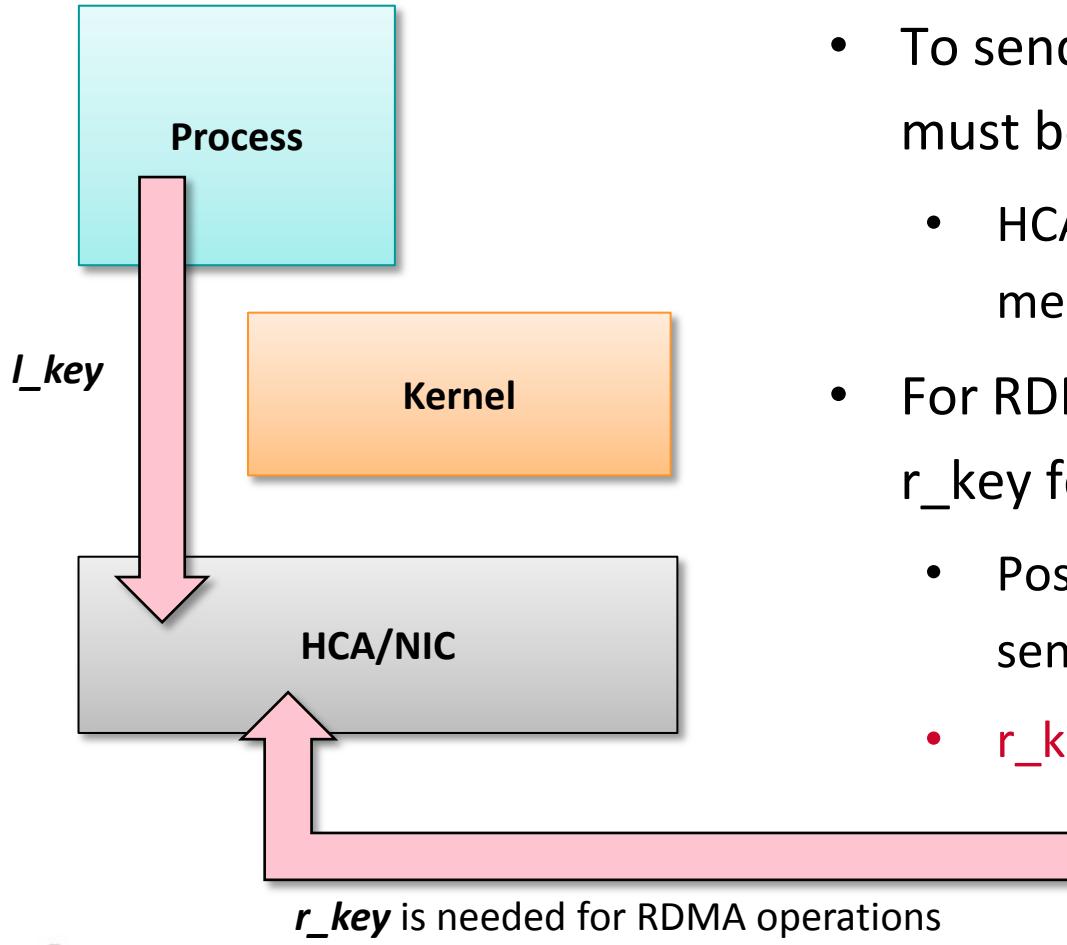
Before we do any communication:  
All memory used for communication must  
be registered



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

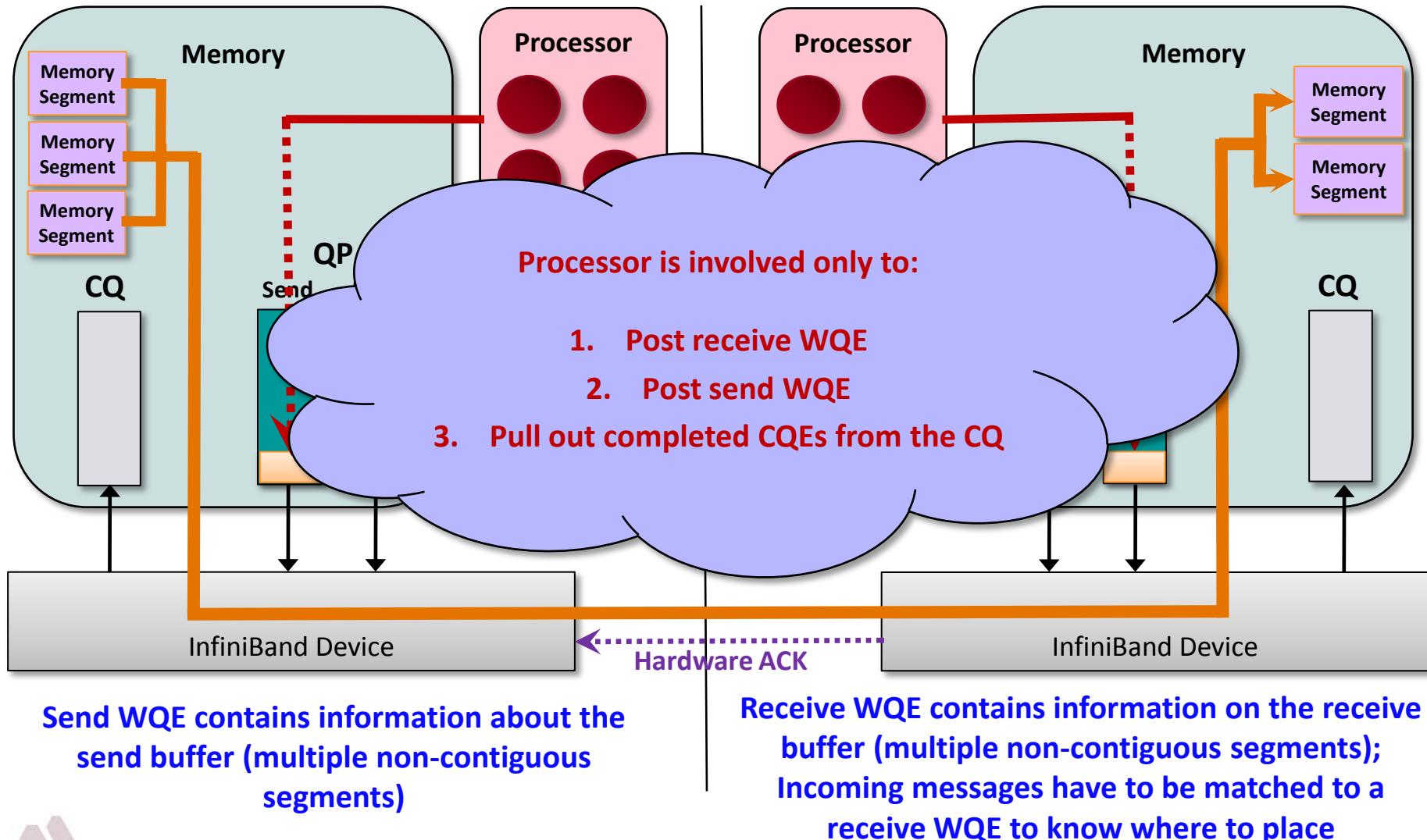
# Memory Protection

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

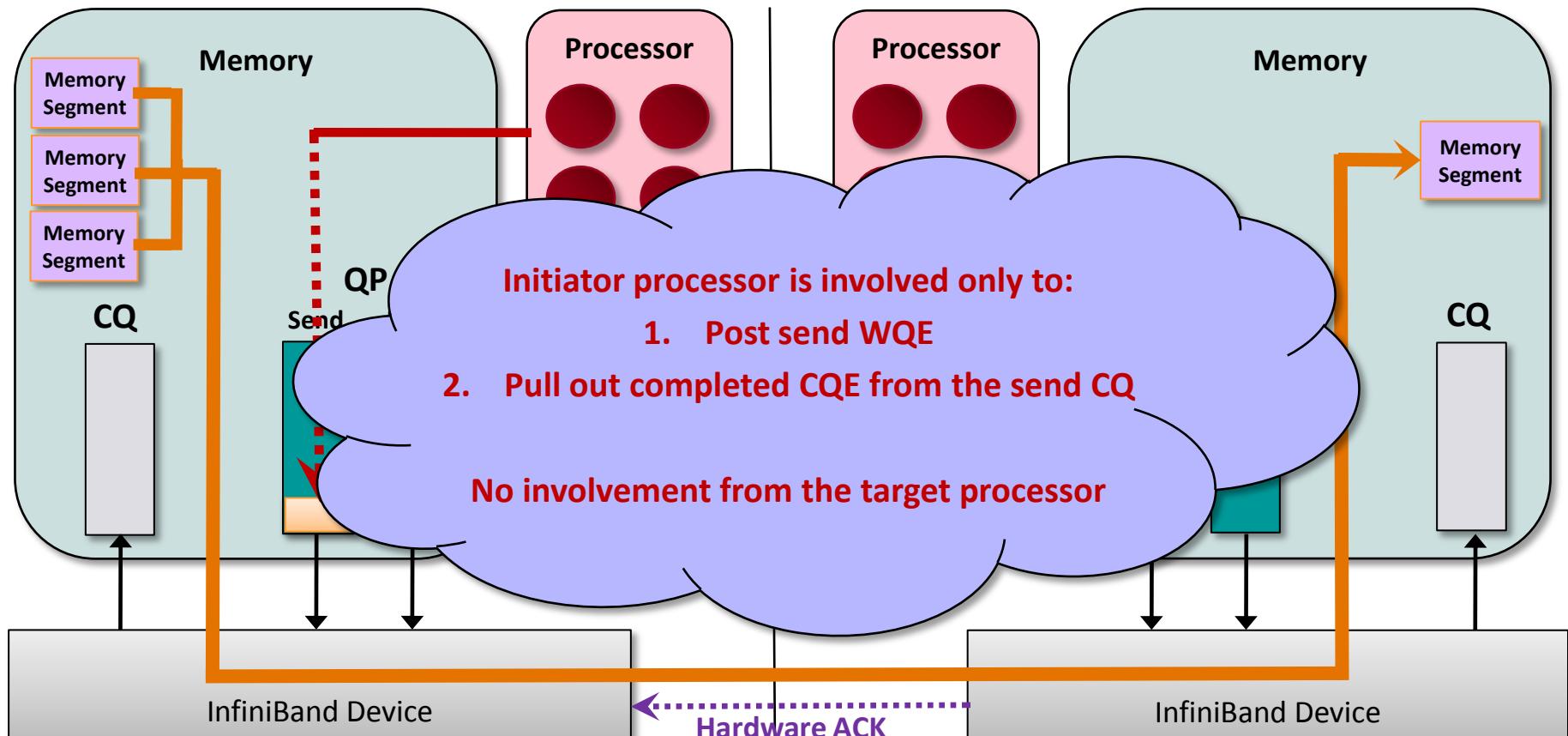


- To send or receive data the *l\_key* must be provided to the HCA
  - HCA verifies access to local memory
- For RDMA, initiator must have the *r\_key* for the remote virtual address
  - Possibly exchanged with a send/recv
  - *r\_key* is not encrypted in IB

# Communication in the Channel Semantics (Send/Receive Model)

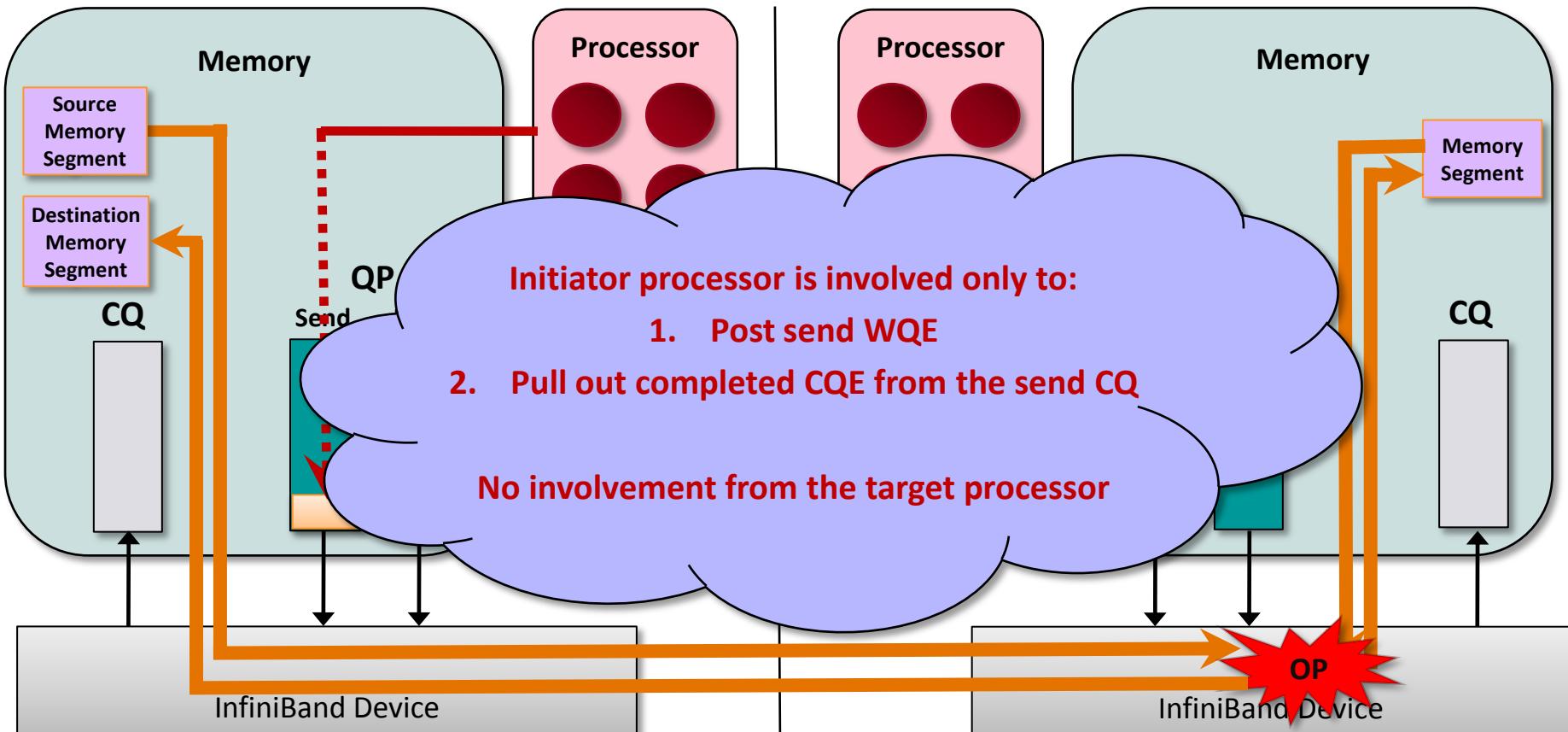


# Communication in the Memory Semantics (RDMA Model)



Send WQE contains information about the send buffer (multiple segments) and the receive buffer (single segment)

# Communication in the Memory Semantics (Atomics)

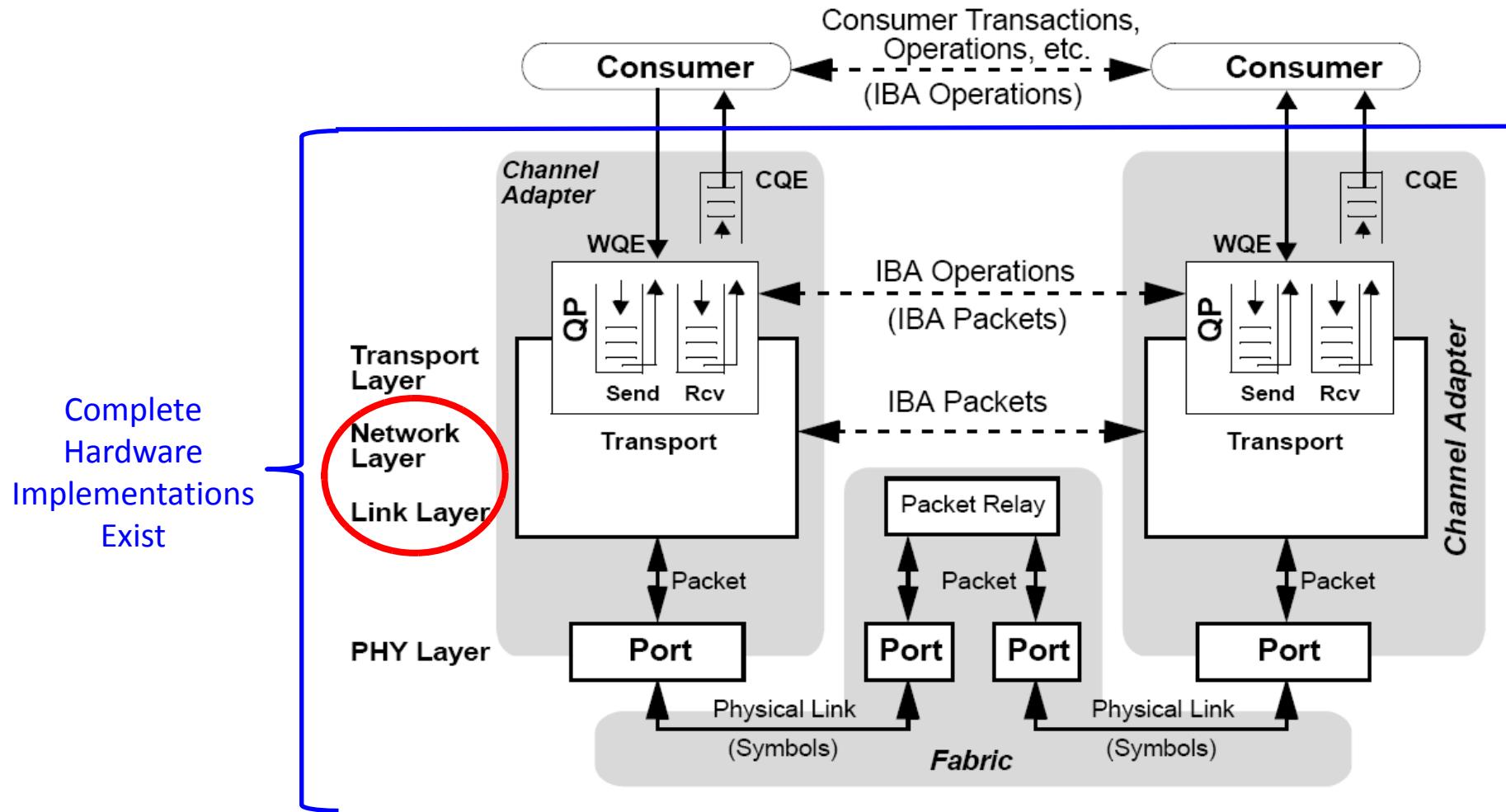


# IB Overview

- **InfiniBand**

- Architecture and Basic Hardware Components
- Communication Model and Semantics
  - Communication Model
  - Memory registration and protection
  - Channel and memory semantics
- **Novel Features**
  - **Hardware Protocol Offload**
    - **Link, network and transport layer features**
  - Subnet Management and Services
  - Sockets Direct Protocol (SDP) stack
  - RSocket Protocol Stack

# Hardware Protocol Offload



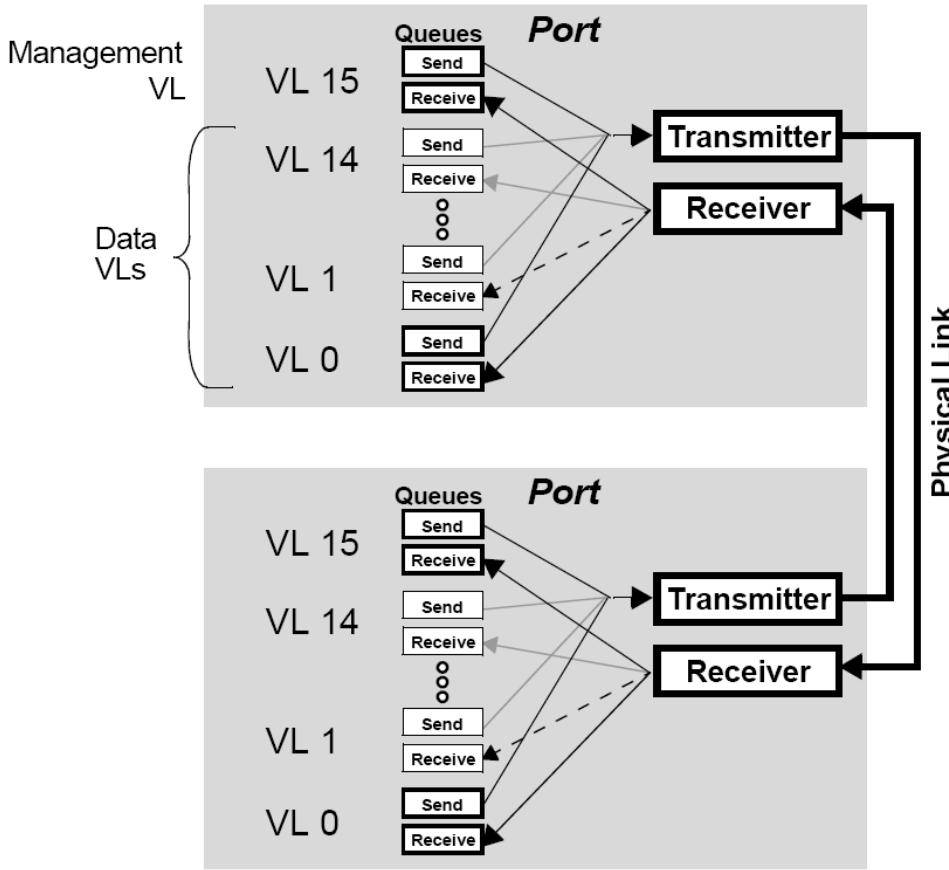
# Link/Network Layer Capabilities

- **Buffering and Flow Control**
- Virtual Lanes, Service Levels and QoS
- Switching and Multicast
- Network Fault Tolerance
- IB WAN Capability

# Buffering and Flow Control

- IB provides three-levels of communication throttling/control mechanisms
  - Link-level flow control (link layer feature)
  - *Message-level flow control (transport layer feature): discussed later*
  - Congestion control (part of the link layer features)
- IB provides an absolute credit-based flow-control
  - Receiver guarantees that enough space is 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)

# Virtual Lanes

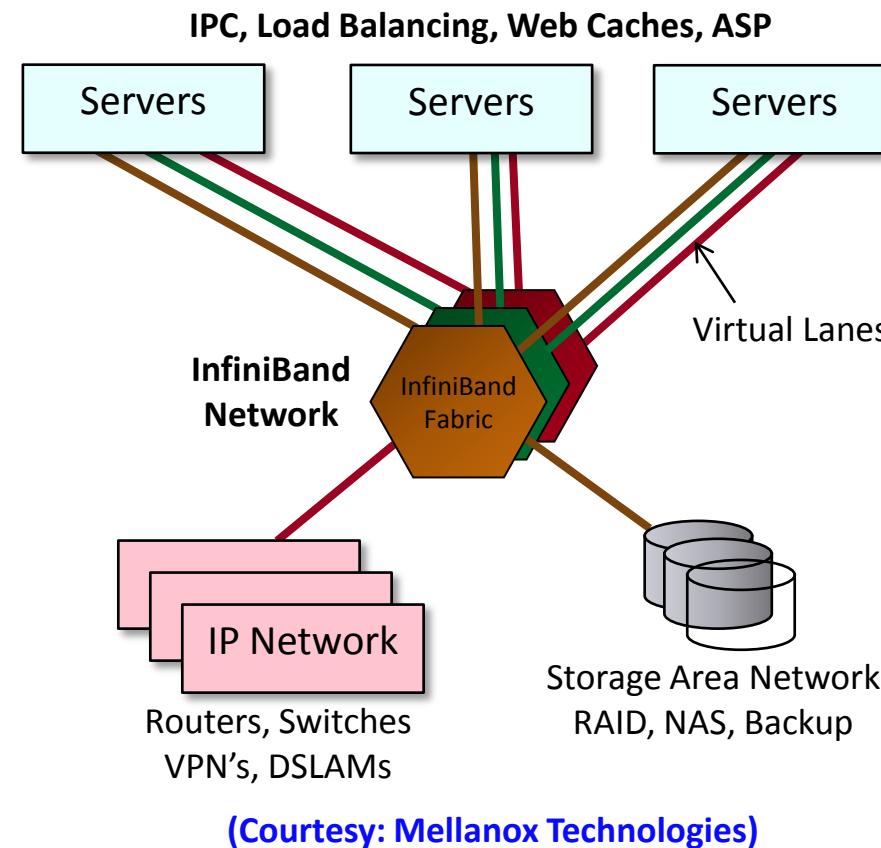


- Multiple virtual links within same physical link
  - Between 2 and 16
- Separate buffers and flow control
  - Avoids Head-of-Line Blocking
- VL15: reserved for management
- Each port supports one or more data VL

# 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



- 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

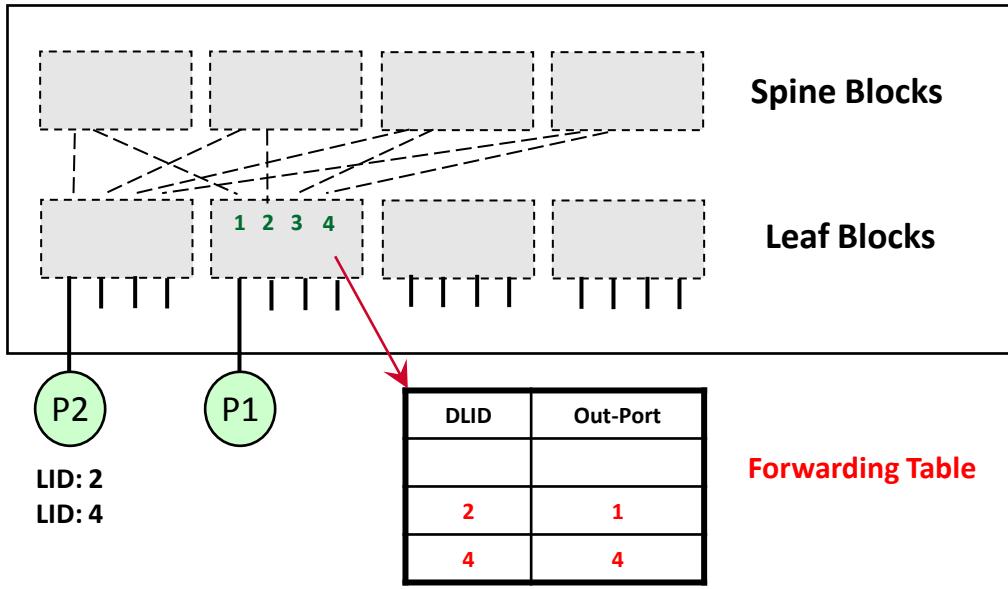
- Each port has one or more associated LIDs (Local Identifiers)
  - Switches look up which port to forward a packet to based on its destination LID (DLID)
  - This information is maintained at the switch
- For multicast packets, the switch needs to maintain multiple output ports to forward the packet 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

# Switch Complex

- Basic unit of switching is a crossbar
  - Current InfiniBand products use either 24-port (DDR) or 36-port (QDR and FDR) crossbars
- Switches available in the market are typically collections of crossbars within a single cabinet
- Do not confuse “non-blocking switches” with “crossbars”
  - Crossbars provide all-to-all connectivity to all connected nodes
    - *For any random node pair selection, all communication is non-blocking*
  - Non-blocking switches provide a fat-tree of many crossbars
    - *For any random node pair selection, there exists a switch configuration such that communication is non-blocking*
    - *If the communication pattern changes, the same switch configuration might no longer provide fully non-blocking communication*

# IB Switching/Routing: An Example

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



Switching: IB supports  
Virtual Cut Through (VCT)

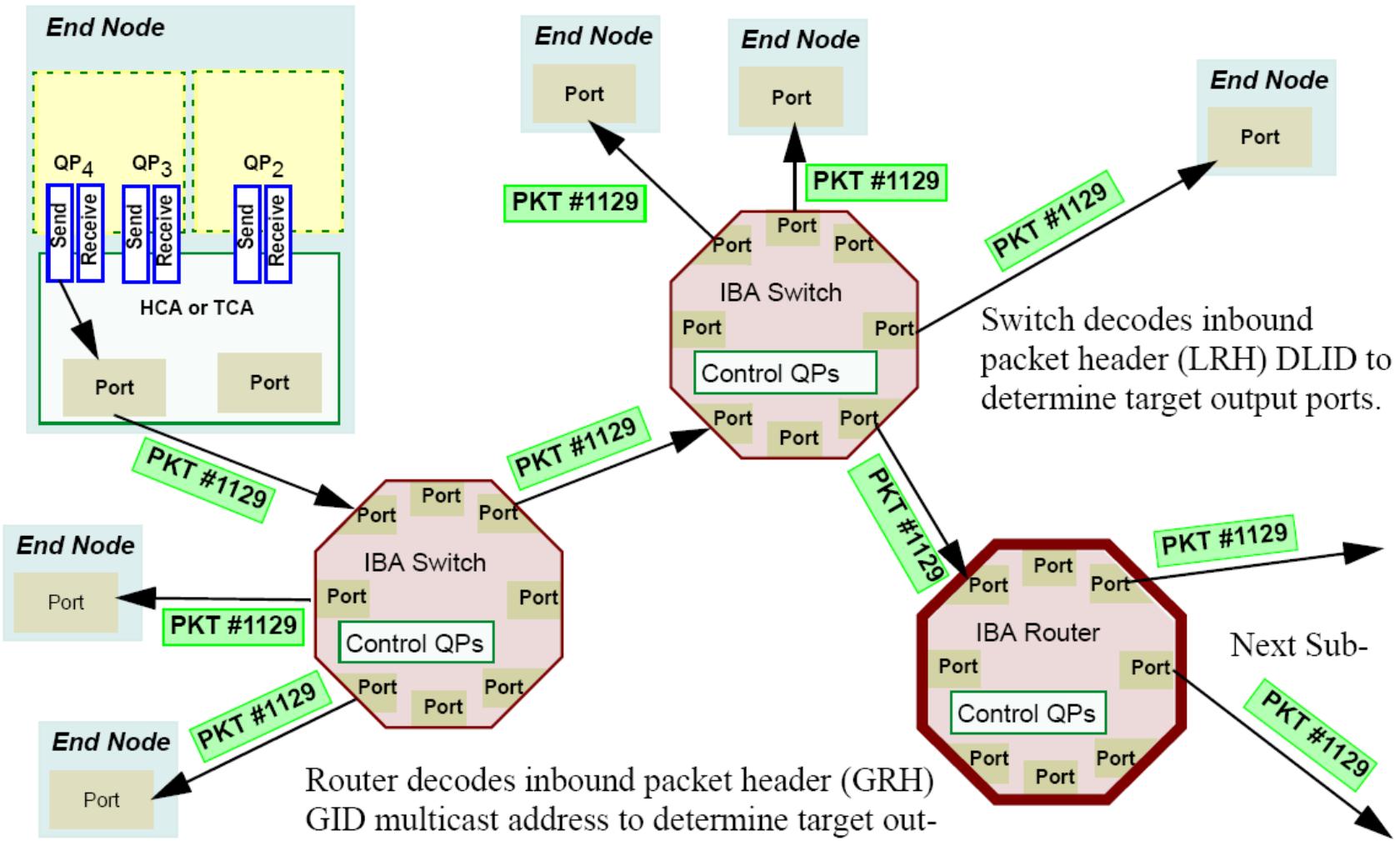
Routing: Unspecified by IB SPEC  
Up\*/Down\*, Shift are popular  
routing engines supported by OFED

- Someone has to setup the forwarding tables and give every port an LID
  - “Subnet Manager” does this work
- Different routing algorithms give different paths
- Fat-Tree is a popular topology for IB Cluster
  - Different over-subscription ratio may be used
- Other topologies
  - 3D Torus (Sandia Red Sky, SDSC Gordon) and SGI Altix (Hypercube)
  - 10D Hypercube (NASA Pleiades)

# More on Multipathing

- Similar to basic switching, except...
  - ... sender can utilize multiple LIDs associated to the same destination port
    - Packets sent to one DLID take a fixed path
    - Different packets can be sent using different DLIDs
    - Each DLID can have a different path (switch can be configured differently for each DLID)
- Can cause out-of-order arrival of packets
  - IB uses a simplistic approach:
    - If packets in one connection arrive out-of-order, they are dropped
  - Easier to use different DLIDs for different connections
    - This is what most high-level libraries using IB do!

# IB Multicast Example



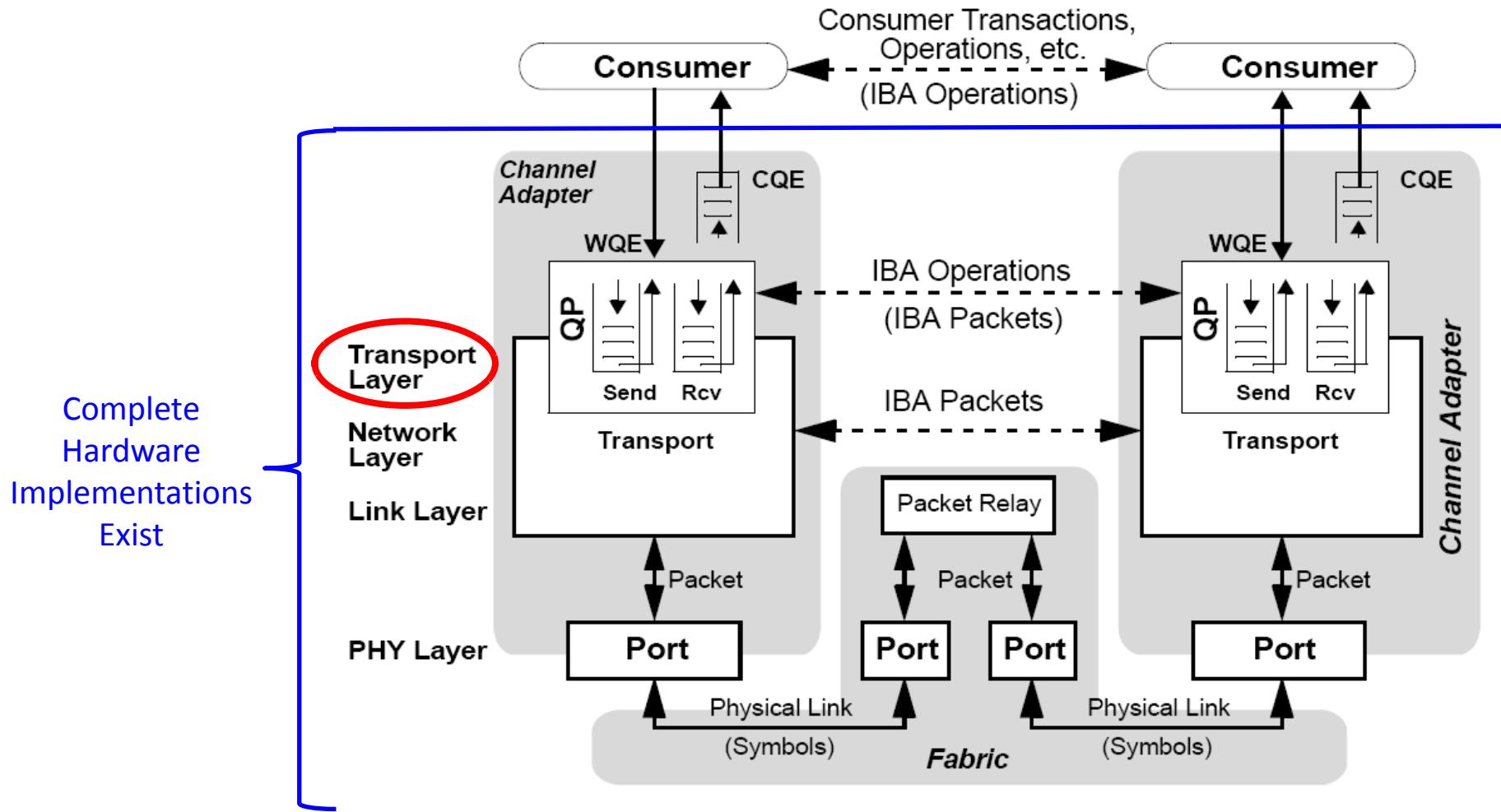
## Network Level Fault Tolerance: Automatic Path Migration

- Automatically utilizes multipathing for network fault-tolerance (optional feature)
- Idea is that the high-level library (or application) using IB will have one primary path, and one fall-back path
  - Enables migrating connections to a different path
    - Connection recovery in the case of failures
- Available for RC, UC, and RD
- Reliability guarantees for service type maintained during migration
- Issue is that there is only one fall-back path (in hardware). If there is more than one failure (or a failure that affects both paths), the application will have to handle this in software

# IB WAN Capability

- Getting increased attention for:
  - Remote Storage, Remote Visualization
  - Cluster Aggregation (Cluster-of-clusters)
- IB-Optical switches by multiple vendors
  - Mellanox Technologies: [www.mellanox.com](http://www.mellanox.com)
  - Obsidian Research Corporation: [www.obsidianresearch.com](http://www.obsidianresearch.com) & Bay Microsystems: [www.baymicrosystems.com](http://www.baymicrosystems.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



# 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

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

# Trade-offs in Different Transport Types

Attribute	Reliable Connection	Reliable Datagram	eXtended Reliable Connection	Unreliable Connection	Unreliable Datagram	Raw Datagram
Scalability (M processes, N nodes)	$M^2N$ QPs per HCA	M QPs per HCA	$MN$ QPs per HCA	$M^2N$ QPs per HCA	M QPs per HCA	1 QP per HCA
Reliability	Corrupt data detected	Yes				
	Data Delivery Guarantee	Data delivered exactly once			No guarantees	
	Data Order Guarantees	Per connection	One source to multiple destinations	Per connection	Unordered, duplicate data detected	No
	Data Loss Detected	Yes				No
	Error Recovery	Errors (retransmissions, alternate path, etc.) handled by transport layer. Client only involved in handling fatal errors (links broken, protection violation, etc.)			Packets with errors and sequence errors are reported to responder	None

# Transport Layer Capabilities

- **Data Segmentation**
- **Transaction Ordering**
- **Message-level Flow Control**
- **Static Rate Control and Auto-negotiation**

# Data Segmentation

- IB transport layer provides a message-level communication granularity, not byte-level (unlike TCP)
- Application can hand over a large message
  - Network adapter segments it to MTU sized packets
  - Single notification when the entire message is transmitted or received (not per packet)
- Reduced host overhead to send/receive messages
  - Depends on the number of messages, not the number of bytes

## Transaction Ordering

- IB follows a strong transaction ordering for RC
- Sender network adapter transmits messages in the order in which WQEs were posted
- Each QP utilizes a single LID
  - All WQEs posted on same QP take the same path
  - All packets are received by the receiver in the same order
  - All receive WQEs are completed in the order in which they were posted

# Message-level Flow-Control

- Also called as End-to-end Flow-control
  - Does not depend on the number of network hops
- Separate from Link-level Flow-Control
  - Link-level flow-control only relies on the number of bytes being transmitted, not the number of messages
  - Message-level flow-control only relies on the number of messages transferred, not the number of bytes
- If 5 receive WQEs are posted, the sender can send 5 messages (can post 5 send WQEs)
  - If the sent messages are larger than what the receive buffers are posted, flow-control cannot handle it

# Static Rate Control and Auto-Negotiation

- IB allows link rates to be statically changed
  - On a 4X link, we can set data to be sent at 1X
  - For heterogeneous links, rate can be set to the lowest link rate
  - Useful for low-priority traffic
- Auto-negotiation also available
  - E.g., if you connect a 4X adapter to a 1X switch, data is automatically sent at 1X rate
- Only fixed settings available
  - Cannot set rate requirement to 3.16 Gbps, for example

# IB Overview

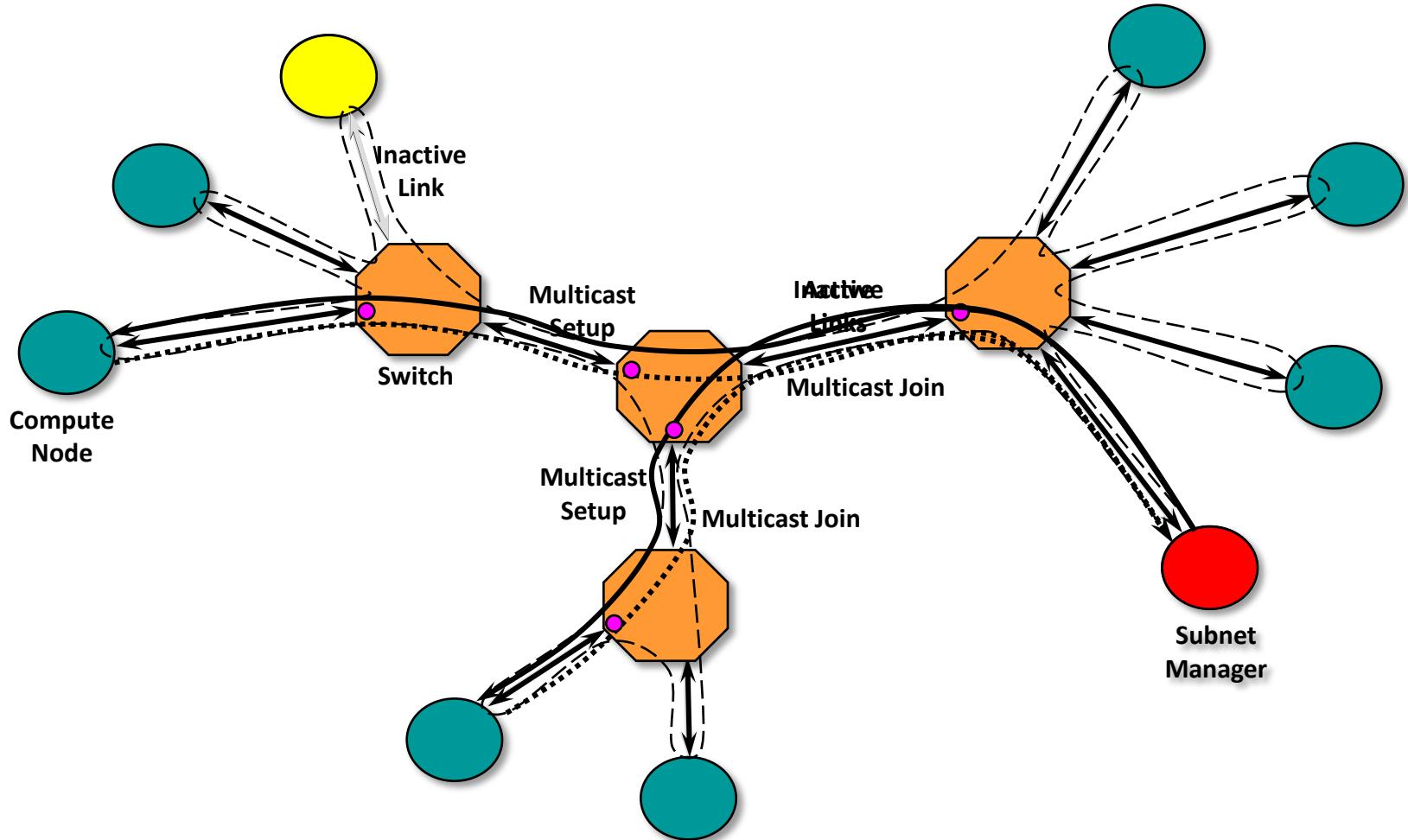
- **InfiniBand**

- Architecture and Basic Hardware Components
- Communication Model and Semantics
  - Communication Model
  - Memory registration and protection
  - Channel and memory semantics
- Novel Features
  - Hardware Protocol Offload
    - Link, network and transport layer features
- **Subnet Management and Services**
- Sockets Direct Protocol (SDP) Stack
- RSocket Protocol Stack

# 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

# Subnet Manager

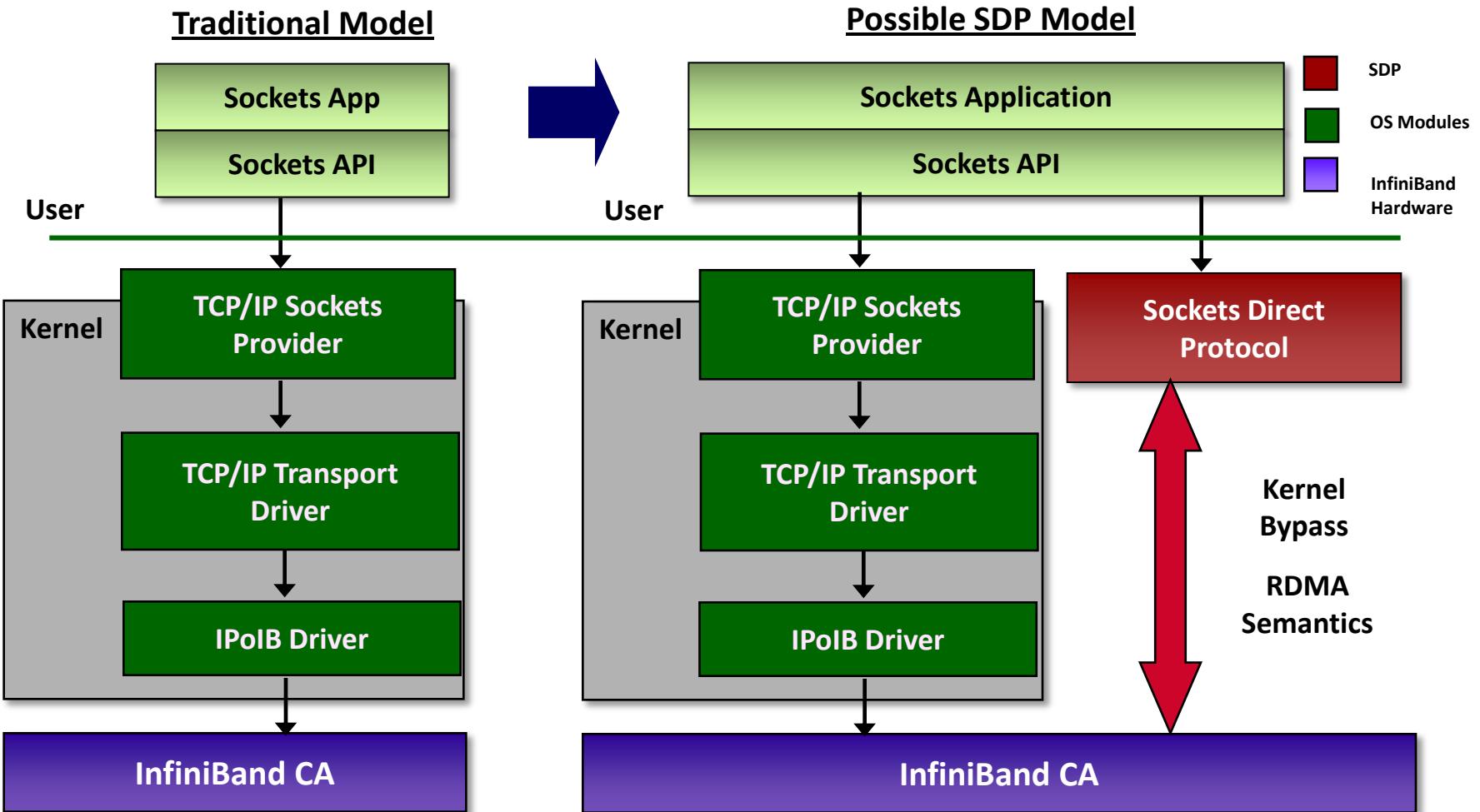


# IB Overview

- **InfiniBand**

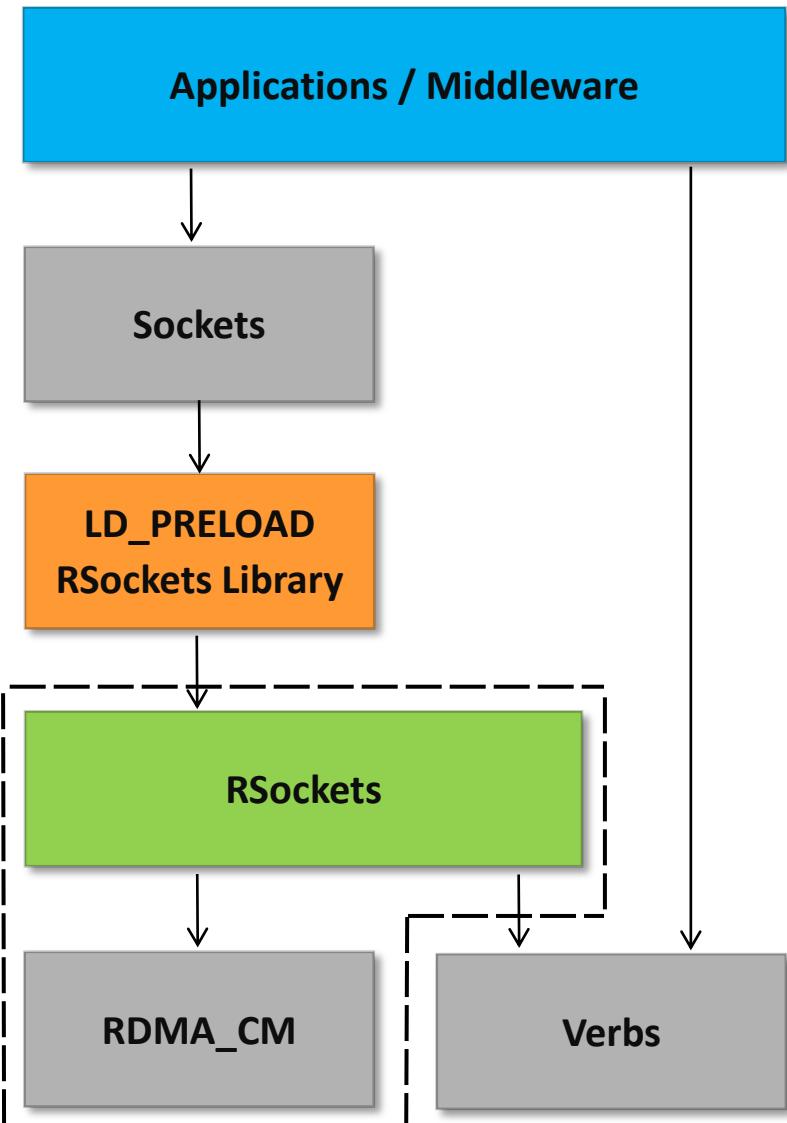
- Architecture and Basic Hardware Components
- Communication Model and Semantics
  - Communication Model
  - Memory registration and protection
  - Channel and memory semantics
- Novel Features
  - Hardware Protocol Offload
    - Link, network and transport layer features
  - Subnet Management and Services
  - **Sockets Direct Protocol (SDP) Stack**
  - **RSockets Protocol Stack**

# IPoIB vs. SDP Architectural Models

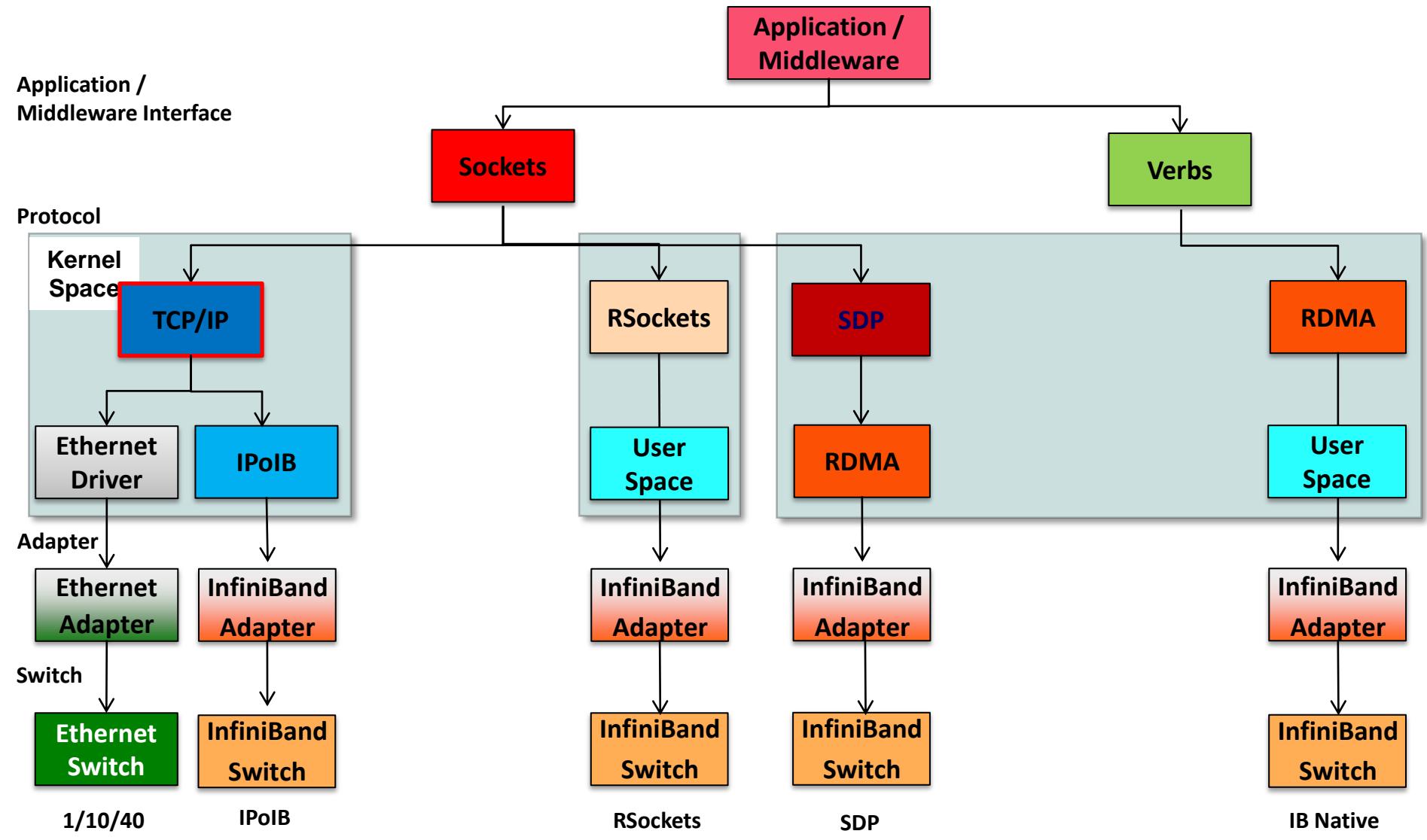


# RSocket Overview

- Implements various socket like functions
  - Functions take same parameters as sockets
- Can switch between regular Sockets and RSocket using LD\_PRELOAD



# TCP/IP, IPoIB, Native IB Verbs, SDP and RSocket



# IB, HSE and their Convergence

- InfiniBand
  - Architecture and Basic Hardware Components
  - Communication Model and Semantics
  - Novel Features
  - Subnet Management and Services
- **High-speed Ethernet Family**
  - Internet Wide Area RDMA Protocol (**iWARP**)
  - **Alternate vendor-specific protocol stacks**
- InfiniBand/Ethernet Convergence Technologies
  - Virtual Protocol Interconnect (VPI)
  - RDMA over Converged Enhanced Ethernet (RoCE)

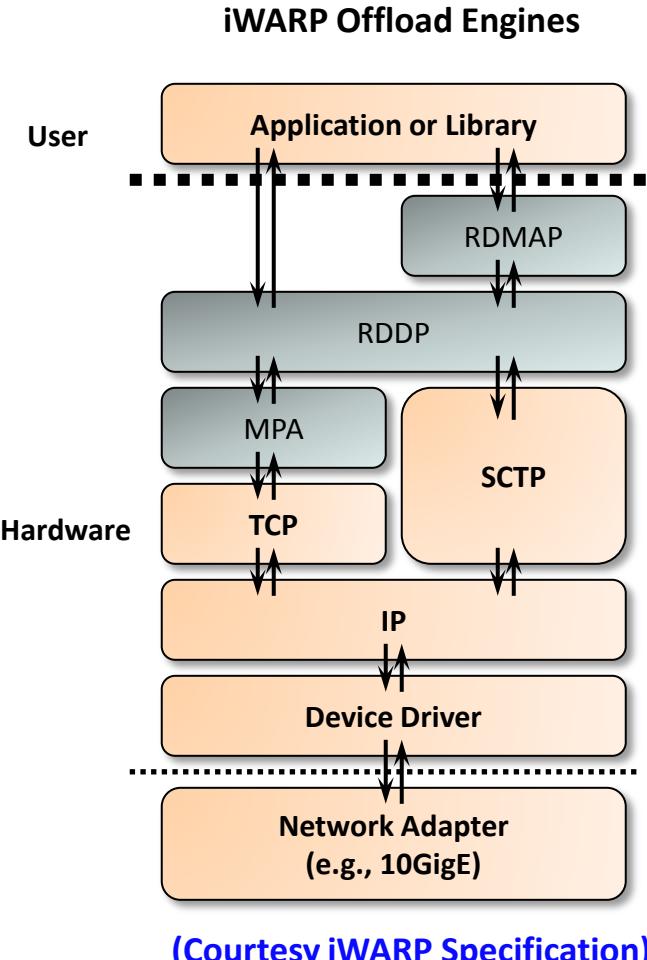
# HSE Overview

- **High-speed Ethernet Family**
  - Internet Wide-Area RDMA Protocol (**iWARP**)
    - **Architecture and Components**
    - Features
      - Out-of-order data placement
      - Dynamic and Fine-grained Data Rate control
    - Existing Implementations of HSE/iWARP
  - Alternate Vendor-specific Stacks
    - MX over Ethernet (for Myricom 10GE adapters)
    - Datagram Bypass Layer (for Myricom 10GE adapters)
    - Solarflare OpenOnload (for Solarflare 10GE adapters)

# IB and 10GE RDMA Models: Commonalities and Differences

Features	IB	iWARP/HSE
Hardware Acceleration	Supported	Supported
RDMA	Supported	Supported
Atomic Operations	Supported	Not supported
Multicast	Supported	Supported
Congestion Control	Supported	Supported
Data Placement	Ordered	Out-of-order
Data Rate-control	Static and Coarse-grained	Dynamic and Fine-grained
QoS	Prioritization	Prioritization and Fixed Bandwidth QoS
Multipathing	Using DLIDs	Using VLANs

# iWARP Architecture and Components



- ***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)

# HSE Overview

- **High-speed Ethernet Family**
  - Internet Wide-Area RDMA Protocol (**iWARP**)
    - Architecture and Components
    - **Features**
      - **Out-of-order data placement**
      - **Dynamic and Fine-grained Data Rate control**
    - Existing Implementations of HSE/iWARP
  - Alternate Vendor-specific Stacks
    - MX over Ethernet (for Myricom 10GE adapters)
    - Datagram Bypass Layer (for Myricom 10GE adapters)
    - Solarflare OpenOnload (for Solarflare 10GE adapters)

# Decoupled Data Placement and Data Delivery

- Place data as it arrives, whether in or out-of-order
- If data is out-of-order, place it at the appropriate offset
- Issues from the application's perspective:
  - Second half of the message has been placed 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 the previous messages have been placed
- Issues from protocol stack's perspective
  - The receiver network stack has to understand each frame of data
    - If the frame is unchanged during transmission, this is easy!
  - The MPA protocol layer adds appropriate information at regular intervals to allow the receiver to identify fragmented frames

# 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 Gbps network refers to a bandwidth allocation of 5 Gbps
  - 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

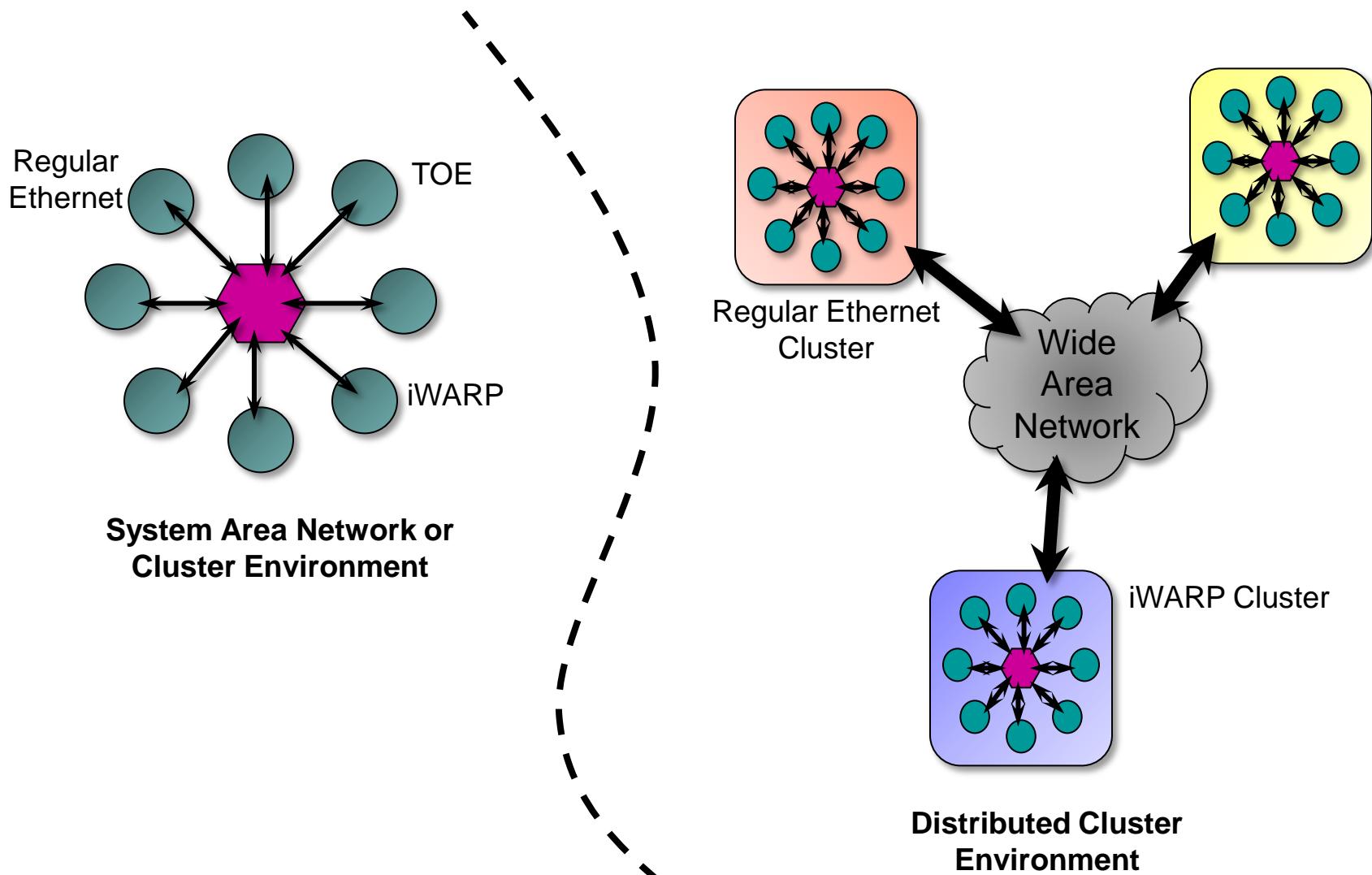
# Prioritization and Fixed Bandwidth QoS

- 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”

# HSE Overview

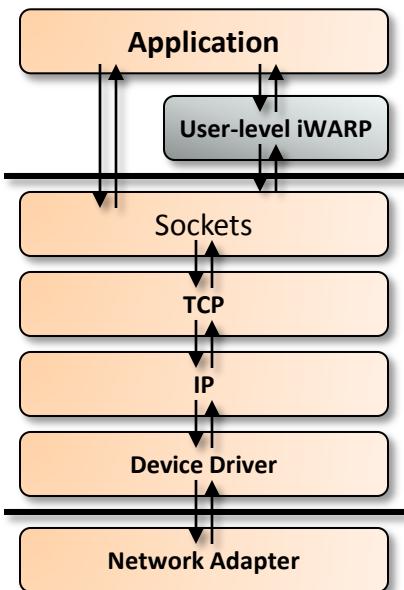
- **High-speed Ethernet Family**
  - Internet Wide-Area RDMA Protocol (**iWARP**)
    - Architecture and Components
    - Features
      - Out-of-order data placement
      - Dynamic and Fine-grained Data Rate control
    - **Existing Implementations of HSE/iWARP**
  - Alternate Vendor-specific Stacks
    - MX over Ethernet (for Myricom 10GE adapters)
    - Datagram Bypass Layer (for Myricom 10GE adapters)
    - Solarflare OpenOnload (for Solarflare 10GE adapters)

# Current Usage of Ethernet



# Different iWARP Implementations

OSU, OSC, IBM



Application

Sockets

Kernel-level iWARP

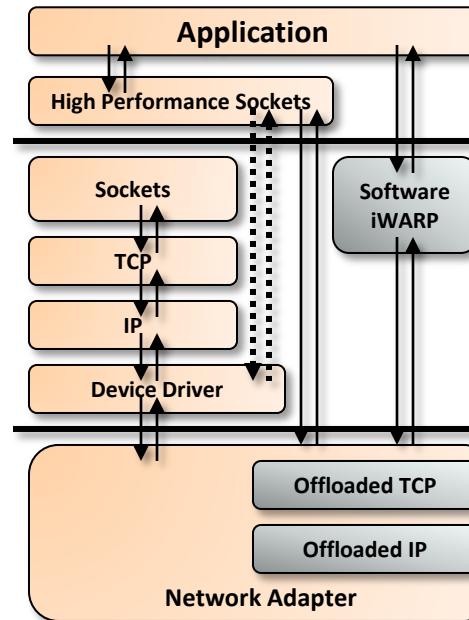
TCP (Modified with MPA)

IP

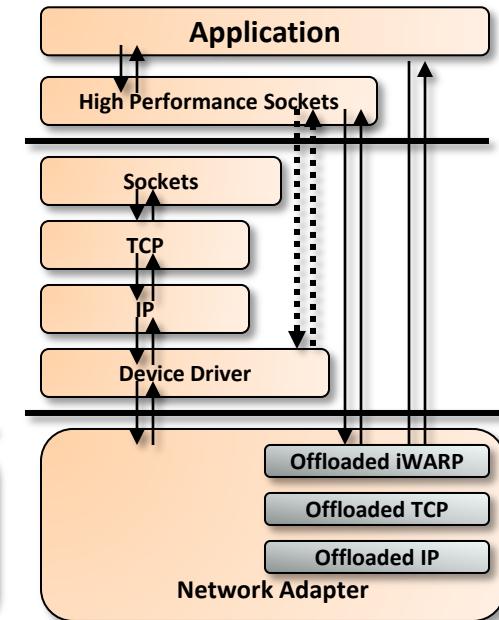
Device Driver

Network Adapter

OSU, ANL



Chelsio, NetEffect (Intel)

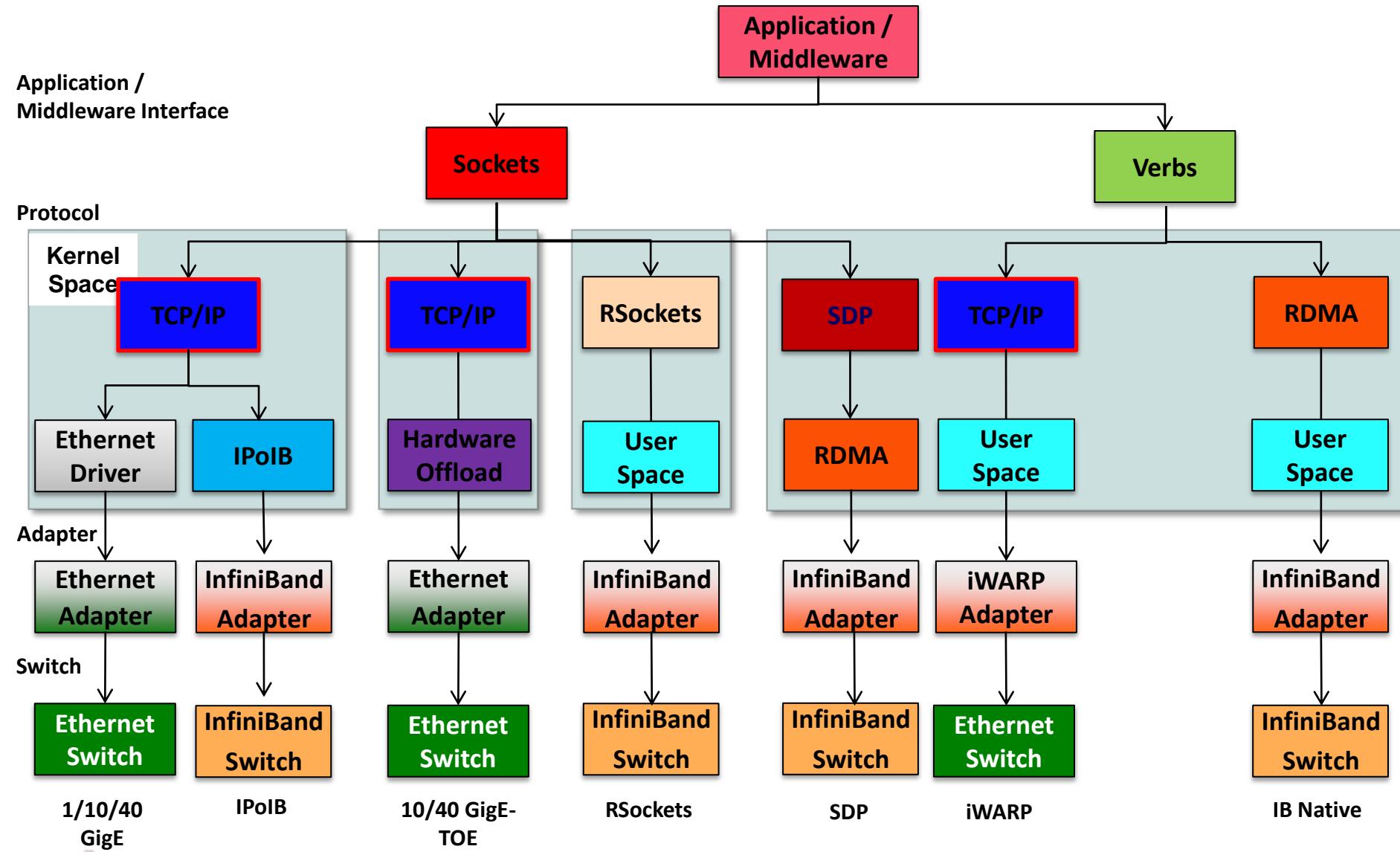


Regular Ethernet Adapters

TCP Offload Engines

iWARP compliant  
Adapters

# iWARP and TOE



# HSE Overview

- **High-speed Ethernet Family**
  - Internet Wide-Area RDMA Protocol (iWARP)
    - Architecture and Components
    - Features
      - Out-of-order data placement
      - Dynamic and Fine-grained Data Rate control
    - Existing Implementations of HSE/iWARP
  - **Alternate Vendor-specific Stack**
    - **MX over Ethernet (for Myricom 10GE adapters)**
    - **Datagram Bypass Layer (for Myricom 10GE adapters)**
    - **Solarflare OpenOnload (for Solarflare 10GE adapters)**
    - **Emulex FastStack DBL (for OneConnect OCe12000-D 10GE adapters)**

## Myrinet Express (MX)

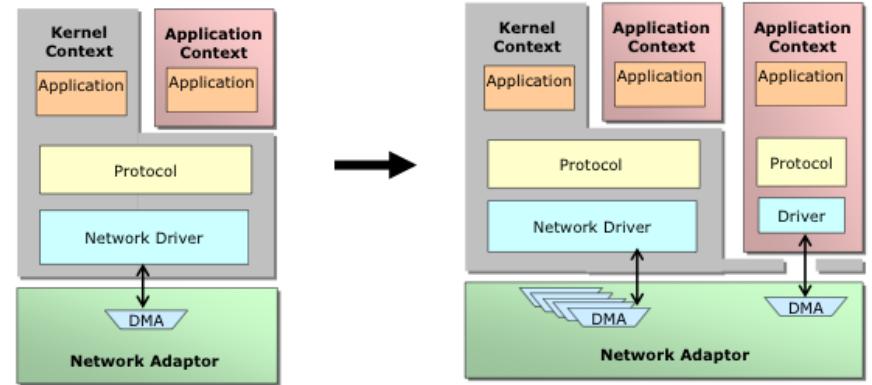
- Proprietary communication layer developed by Myricom for their Myrinet adapters
  - Third generation communication layer (after FM and GM)
  - Supports Myrinet-2000 and the newer Myri-10G adapters
- Low-level “MPI-like” messaging layer
  - Almost one-to-one match with MPI semantics (including connection-less model, implicit memory registration and tag matching)
  - Later versions added some more advanced communication methods such as RDMA to support other programming models such as ARMCI (low-level runtime for the Global Arrays PGAS library)
- Open-MX
  - New open-source implementation of the MX interface for non-Myrinet adapters from INRIA, France

## Datagram Bypass Layer (DBL)

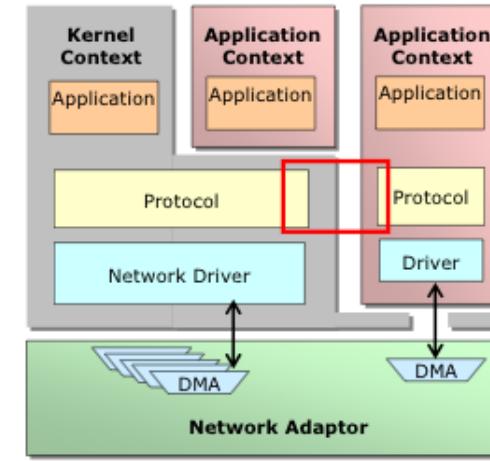
- Another proprietary communication layer developed by Myricom
  - Compatible with regular UDP sockets (embraces and extends)
  - Idea is to bypass the kernel stack and give UDP applications direct access to the network adapter
    - High performance and low-jitter
- Primary motivation: Financial market applications (e.g., stock market)
  - Applications prefer unreliable communication
  - Timeliness is more important than reliability
- *This stack is covered by NDA; more details can be requested from Myricom*

# Solarflare Communications: OpenOnload Stack

- HPC Networking Stack provides many performance benefits, but has limitations for certain types of scenarios, especially where applications tend to fork(), exec() and need asynchronous advancement (per application)



- Solarflare approach:
  - Network hardware provides user-safe interface to route packets directly to apps based on flow information in headers
  - Protocol processing can happen in **both kernel and user space**
  - Protocol state shared **between app and kernel** using shared memory



*Solarflare approach to networking stack*

*Courtesy Solarflare communications ([www.openonload.org/openonload-google-talk.pdf](http://www.openonload.org/openonload-google-talk.pdf))*

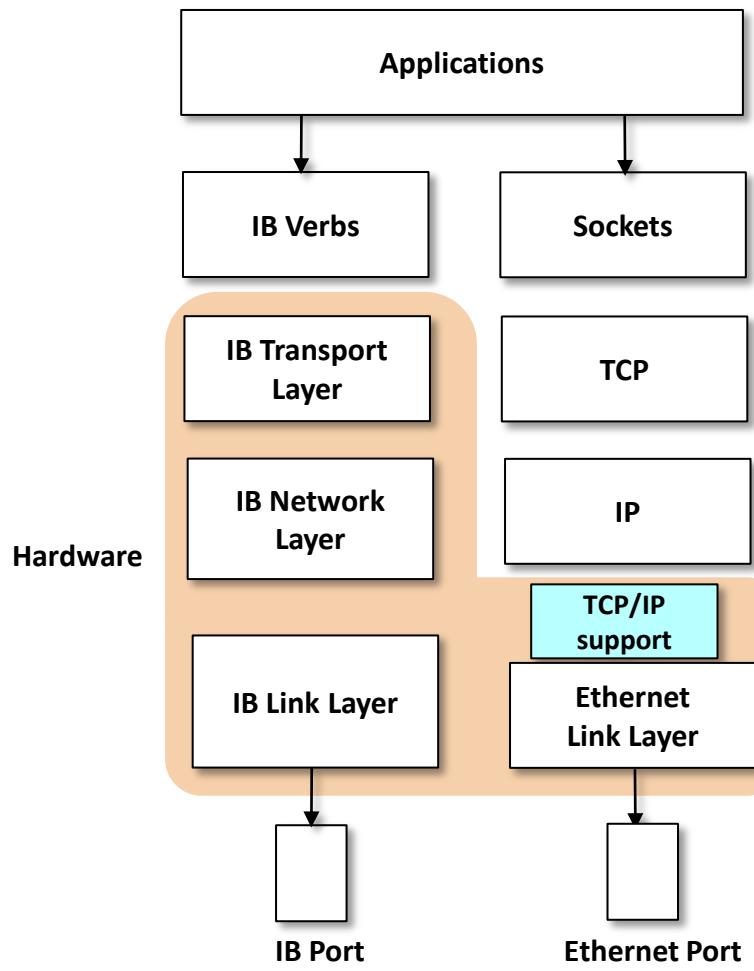
# FastStack DBL

- Proprietary communication layer developed by Emulex
  - Compatible with regular UDP and TCP sockets
  - Idea is to bypass the kernel stack
    - High performance, low-jitter and low latency
  - Available In multiple modes
    - Transparent Acceleration (TA)
      - Accelerate existing sockets applications for UDP/TCP
    - DBL API
      - UDP-only, socket-like semantics but requires application changes
- Primary motivation: Financial market applications (e.g., stock market)
  - Applications prefer unreliable communication
  - Timeliness is more important than reliability
- *This stack is covered by NDA; more details can be requested from Emulex*

# IB, HSE and their Convergence

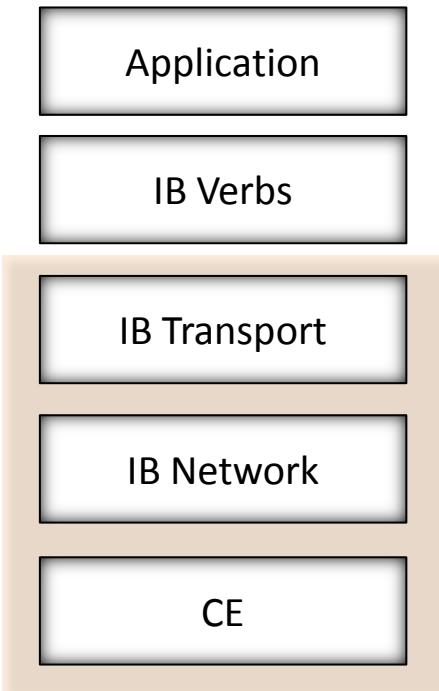
- InfiniBand
  - Architecture and Basic Hardware Components
  - Communication Model and Semantics
  - Novel Features
  - Subnet Management and Services
- High-speed Ethernet Family
  - Internet Wide Area RDMA Protocol (iWARP)
  - Alternate vendor-specific protocol stacks
- **InfiniBand/Ethernet Convergence Technologies**
  - **Virtual Protocol Interconnect (VPI)**
  - **RDMA over Converged Enhanced Ethernet (RoCE)**

# Virtual Protocol Interconnect (VPI)



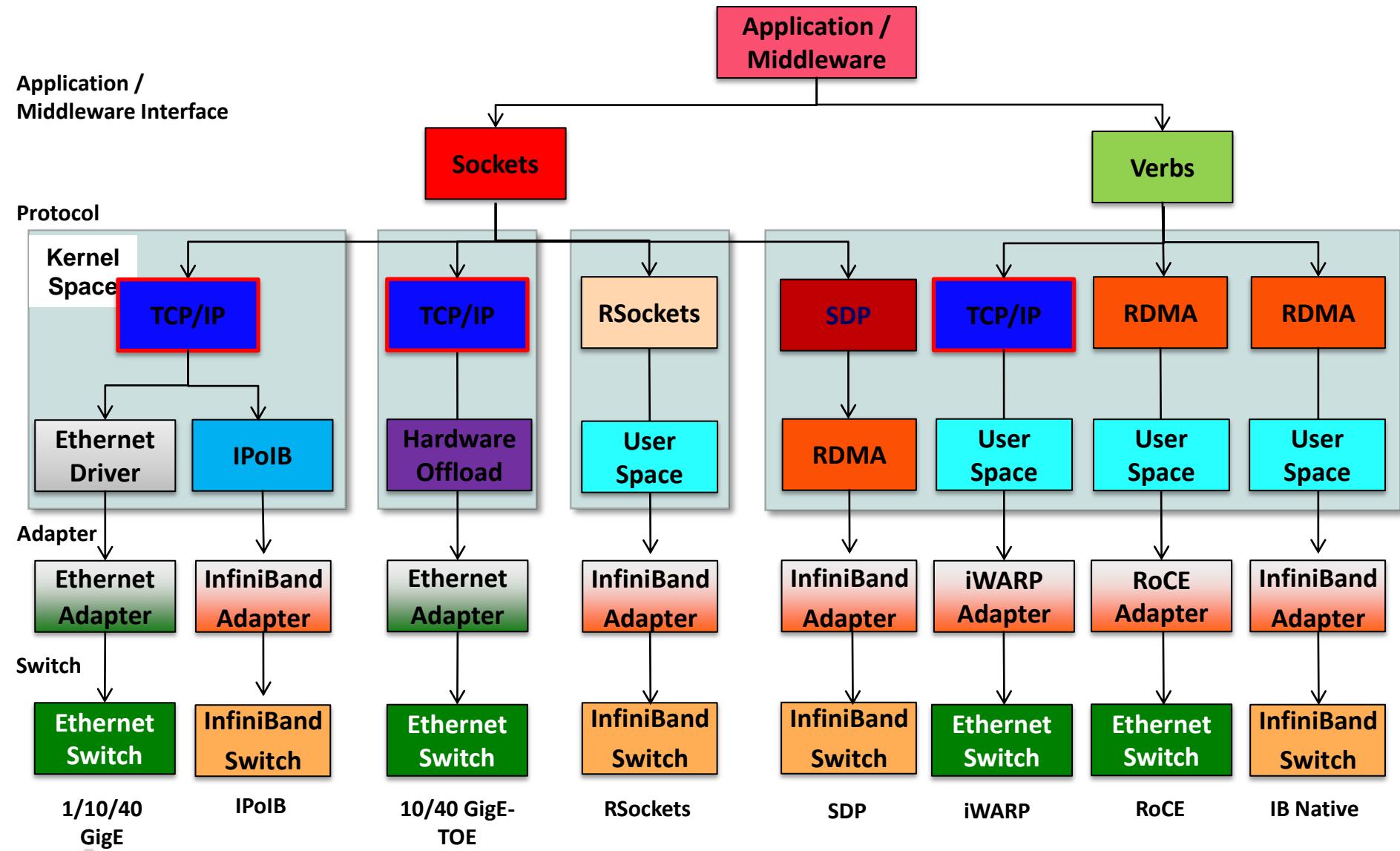
- Single network firmware to support both IB and Ethernet
- Autosensing of layer-2 protocol
  - Can be configured to automatically work with either IB or Ethernet networks
- Multi-port adapters can use one port on IB and another on Ethernet
- Multiple use modes:
  - Datacenters with IB inside the cluster and Ethernet outside
  - Clusters with IB network and Ethernet management

# RDMA over Converged Enhanced Ethernet (RoCE)



- Takes advantage of IB and Ethernet
  - Software written with IB-Verbs
  - Link layer is Converged (Enhanced) Ethernet (CE)
- Pros:
  - Works natively in Ethernet environments (entire Ethernet management ecosystem is available)
  - Has all the benefits of IB verbs
  - CE is very similar to the link layer of native IB, so there are no missing features
- Cons:
  - Network bandwidth might be limited to Ethernet switches: 10/40GE switches available; 56 Gbps IB is available

# All interconnects and protocols including RoCE



# IB and HSE: Feature Comparison

Features	IB	iWARP/HSE	RoCE
Hardware Acceleration	Yes	Yes	Yes
RDMA	Yes	Yes	Yes
Congestion Control	Yes	Optional	Yes
Multipathing	Yes	Yes	Yes
Atomic Operations	Yes	No	Yes
Multicast	Optional	No	Optional
Data Placement	Ordered	Out-of-order	Ordered
Prioritization	Optional	Optional	Yes
Fixed BW QoS (ETS)	No	Optional	Yes
Ethernet Compatibility	No	Yes	Yes
TCP/IP Compatibility	Yes (using IPoIB)	Yes	Yes (using IPoIB)

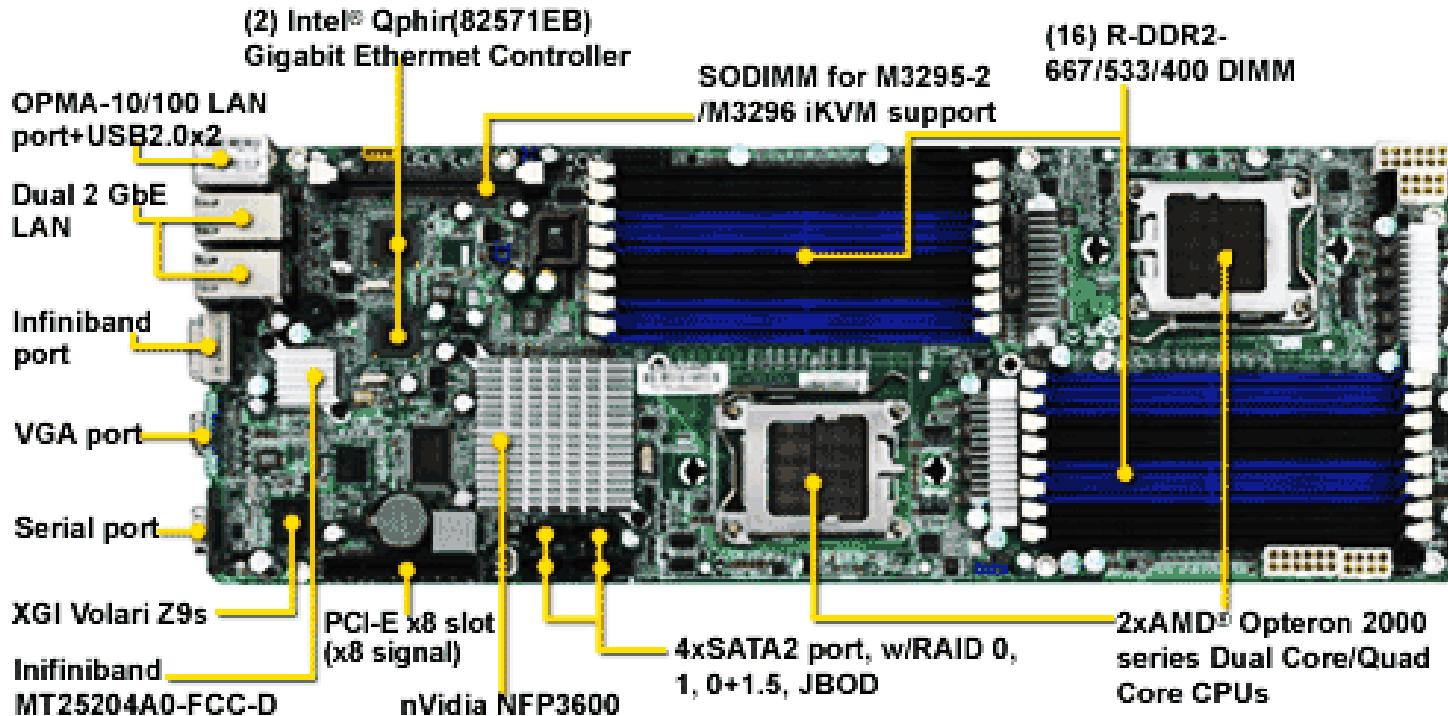
# Presentation Overview

- Introduction
- Why InfiniBand and High-speed Ethernet?
- Overview of IB, HSE, their Convergence and Features
- **IB and HSE HW/SW Products and Installations**
- Sample Case Studies and Performance Numbers
- Conclusions and Final Q&A

# IB Hardware Products

- Many IB vendors: Mellanox+Voltaire and Qlogic (acquired by Intel)
  - Aligned with many server vendors: Intel, IBM, Oracle, 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, PCI 3.0 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), QDR (32 Gbps), FDR (56 Gbps), Dual-FDR (100Gbps)
- ConnectX-2, ConnectX-3 and ConnectIB adapters from Mellanox supports offload for collectives (Barrier, Broadcast, etc.)

# Tyan Thunder S2935 Board



(Courtesy Tyan)

Similar boards from Supermicro with LOM features are also available

# IB Hardware Products (contd.)

- Switches:
  - 4X SDR and DDR (8-288 ports); 12X SDR (small sizes)
  - 3456-port “Magnum” switch from SUN → used at TACC
    - 72-port “nano magnum”
  - 36-port Mellanox InfiniScale IV QDR switch silicon in 2008
    - Up to 648-port QDR switch by Mellanox and SUN
    - Some internal ports are 96 Gbps (12X QDR)
  - IB switch silicon from Qlogic introduced at SC ’08
    - Up to 846-port QDR switch by Qlogic
  - FDR (54.6 Gbps) switch silicon (Bridge-X) and associated switches (18-648 ports) are available
  - Switch-X-2 silicon from Mellanox with VPI and SDN (Software Defined Networking) support announced in Oct ‘12
- Switch Routers with Gateways
  - IB-to-FC; IB-to-IP

# 10G, 40G and 100G Ethernet Products

- 10GE adapters: Intel, Intilop, Myricom, Emulex, Mellanox (ConnectX)
- 10GE/iWARP adapters: Chelsio, NetEffect (now owned by Intel)
- 40GE adapters: Mellanox ConnectX3-EN 40G, Chelsio (4 x 10GigE)
- 10GE switches
  - Fulcrum Microsystems (acquired by Intel recently)
    - Low latency switch based on 24-port silicon
    - FM4000 switch with IP routing, and TCP/UDP support
  - Arista, Brocade, Cisco, Extreme, Force10, Fujitsu, Juniper, Gnodal and Myricom
- 40GE and 100GE switches
  - Gnodal, Arista, Brocade and Mellanox 40GE (SX series)
  - Broadcom has switch architectures for 10/40/100GE
  - Nortel Networks
    - 10GE downlinks with 40GE and 100GE uplinks

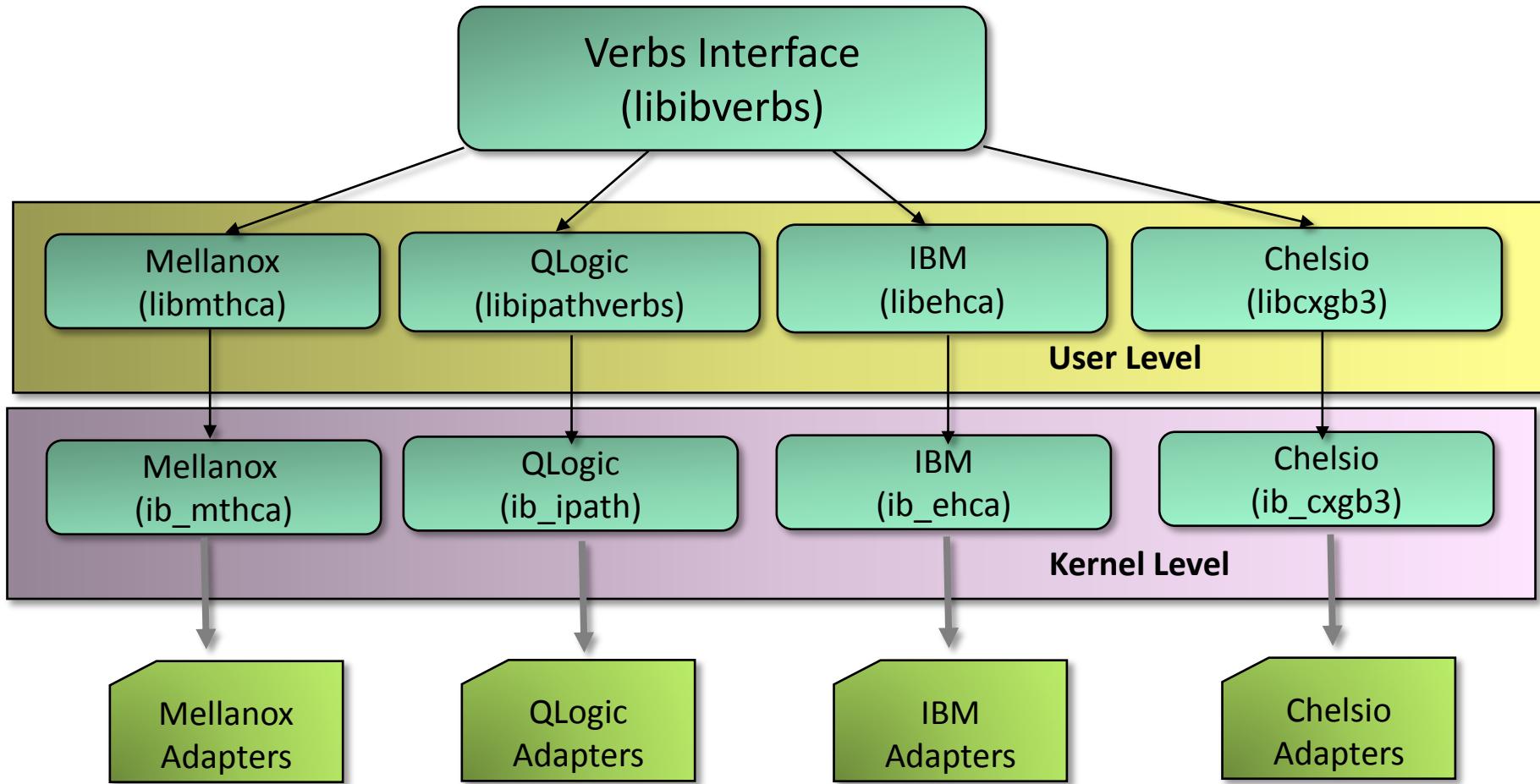
## Products Providing IB and HSE Convergence

- Mellanox ConnectX Adapter
- Supports IB and HSE convergence
- Ports can be configured to support IB or HSE
- Support for VPI and RoCE
  - 8 Gbps (SDR), 16Gbps (DDR), 32Gbps (QDR) and 54.6 Gbps (FDR) rates available for IB
  - 10GE and 40GE rates available for RoCE

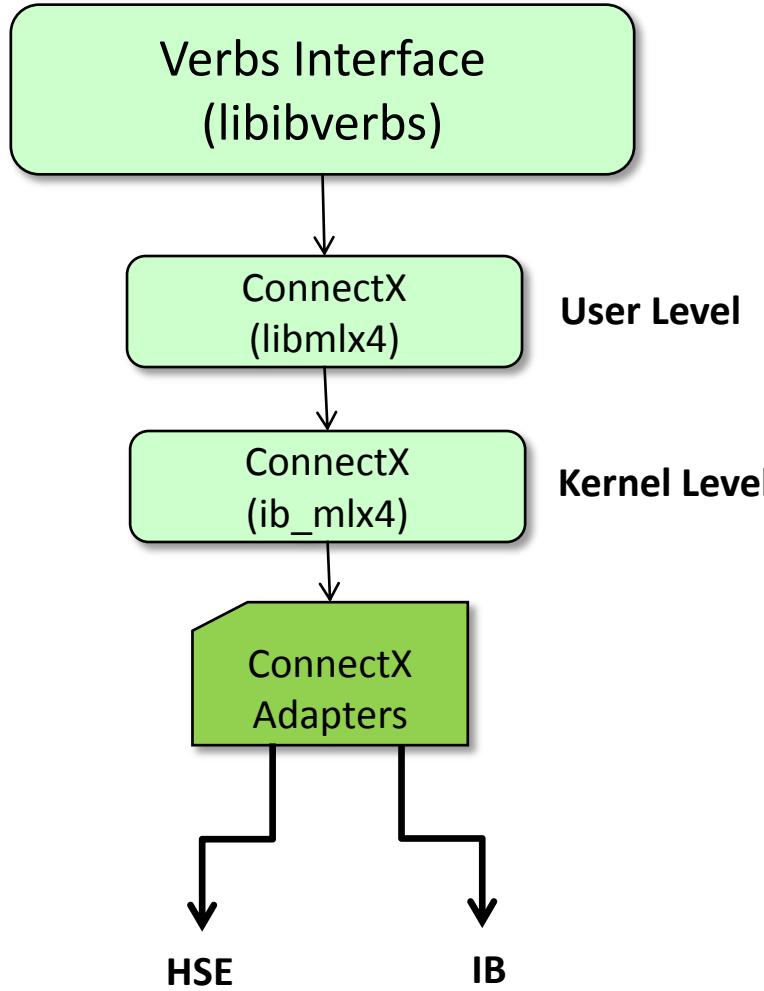
# Software Convergence with OpenFabrics

- Open source organization (formerly OpenIB)
  - [www.openfabrics.org](http://www.openfabrics.org)
- Incorporates both IB and iWARP in a unified manner
  - Support for Linux and Windows
- Users can download the entire stack and run
  - Latest release is OFED 3.5
    - New naming convention to get aligned with Linux Kernel Development

# OpenFabrics Stack with Unified Verbs Interface

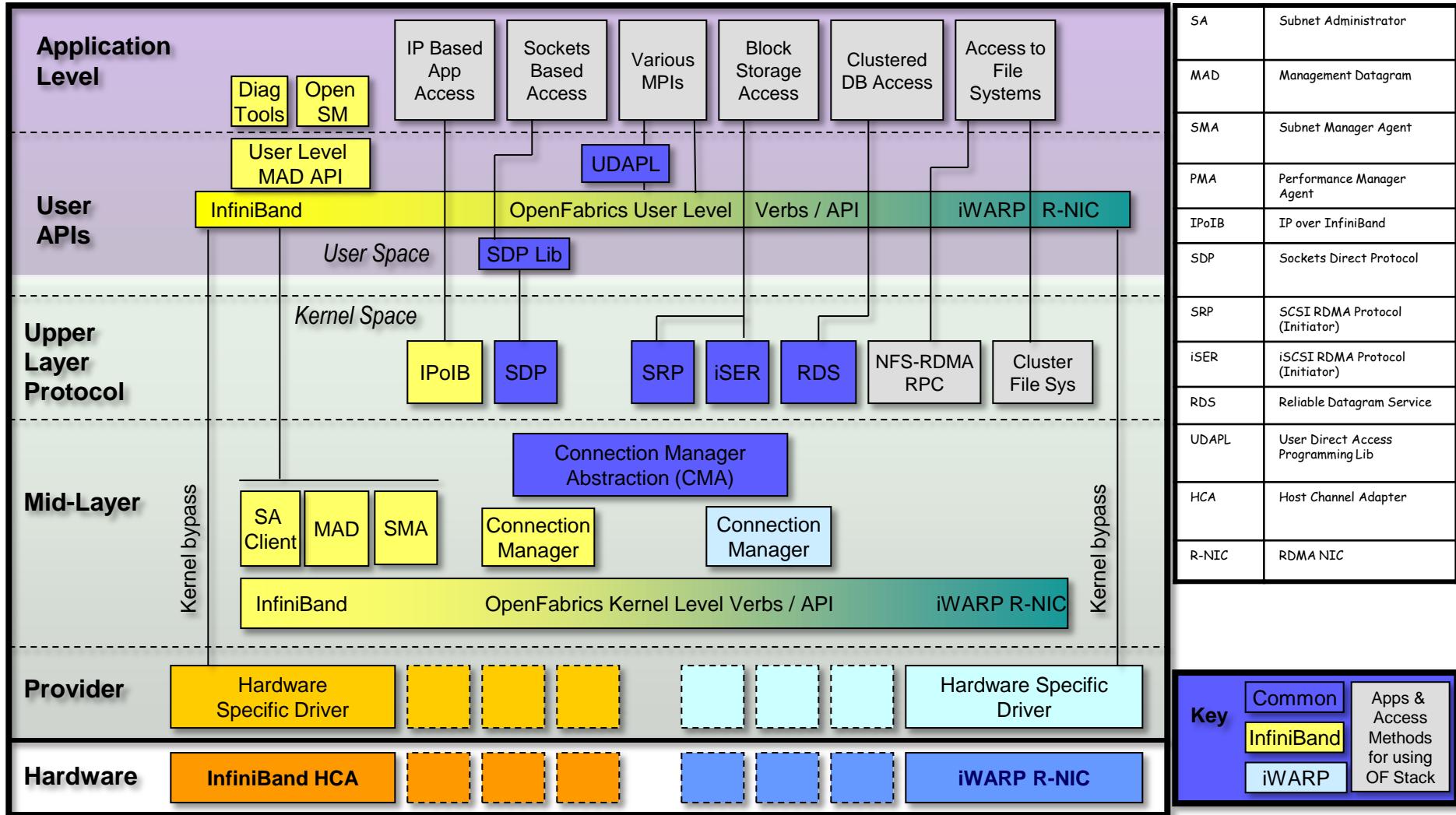


# OpenFabrics on Convergent IB/HSE



- For IBoE and RoCE, the upper-level stacks remain completely unchanged
- Within the hardware:
  - Transport and network layers remain completely unchanged
  - Both IB and Ethernet (or CEE) link layers are supported on the network adapter
- Note: The OpenFabrics stack is not valid for the Ethernet path in VPI
  - That still uses sockets and TCP/IP

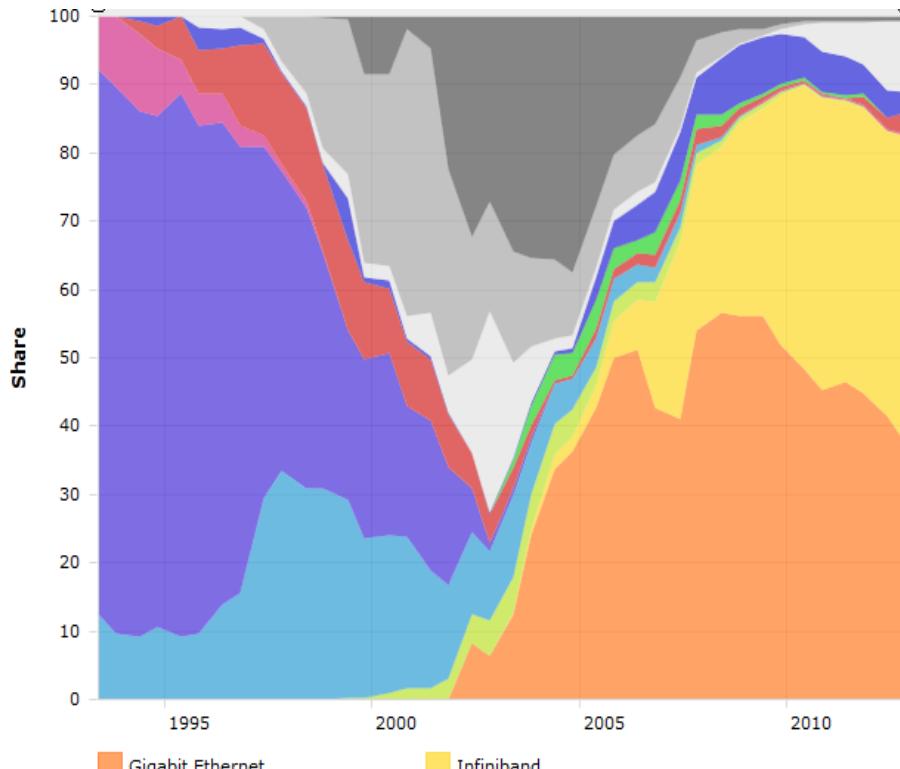
# OpenFabrics Software Stack



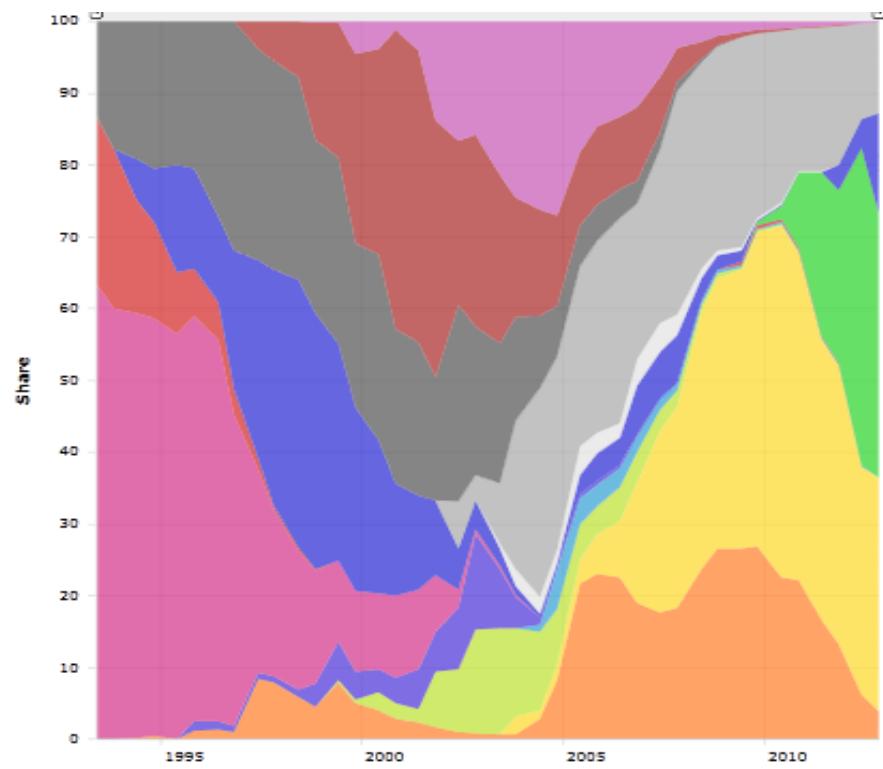
# Trends of Networking Technologies in TOP500 Systems

Percentage share of InfiniBand is steadily increasing

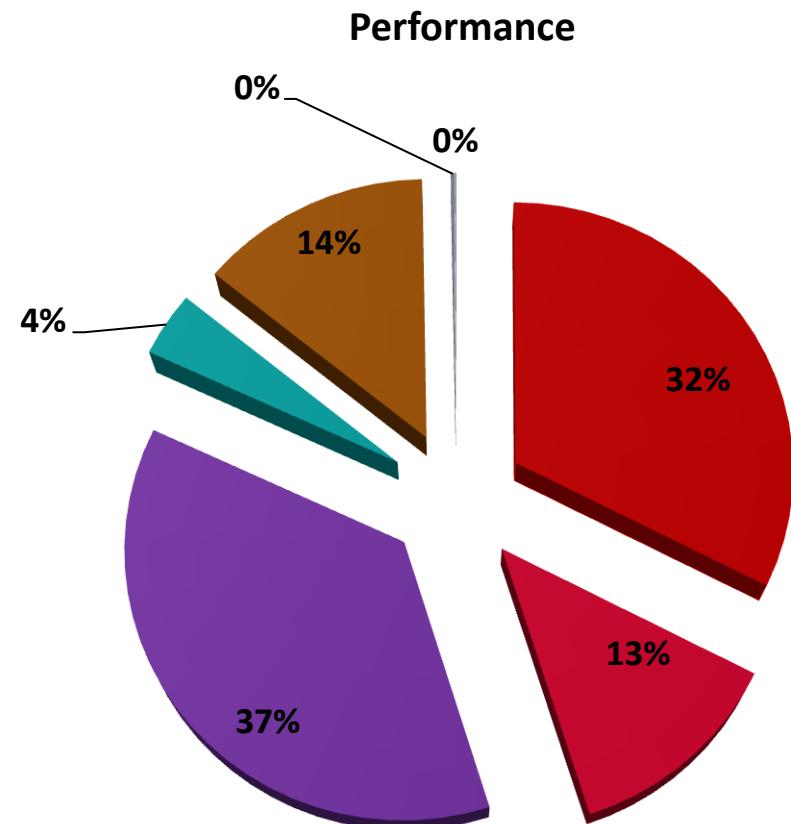
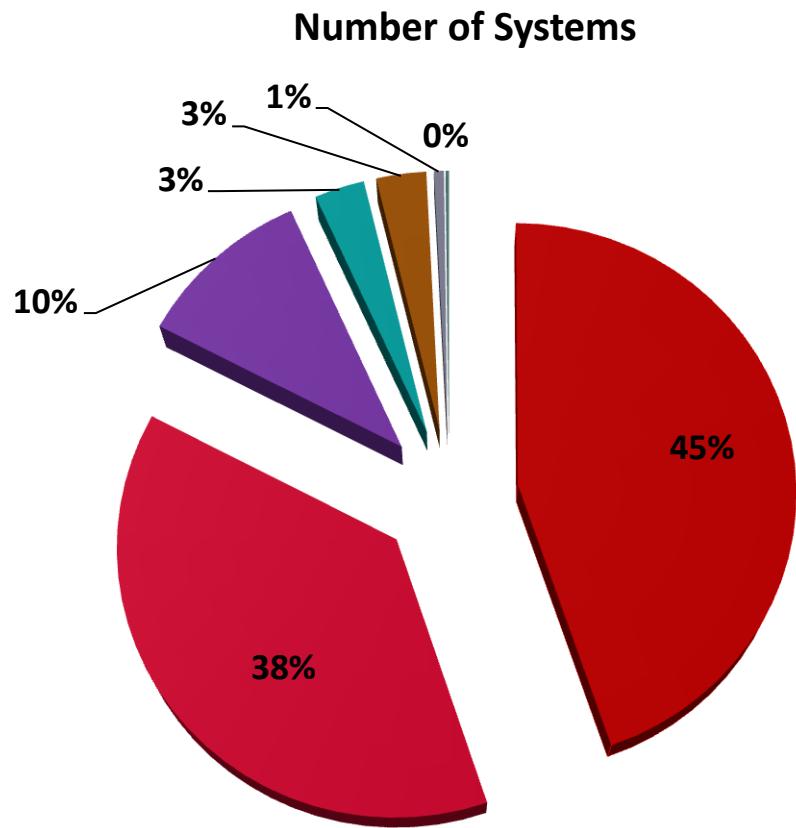
Interconnect Family – Systems Share



Interconnect Family – Performance Share



# InfiniBand in the Top500 (November 2012)



- Infiniband
- Custom Interconnect
- Cray Interconnect
- Fat Tree

- Gigabit Ethernet
- Proprietary Network
- Myrinet

- Infiniband
- Custom Interconnect
- Cray Interconnect
- Fat Tree

- Gigabit Ethernet
- Proprietary Network
- Myrinet

# Large-scale InfiniBand Installations

- 224 IB Clusters (44.8%) in the November 2012 Top500 list  
(<http://www.top500.org>)
- Installations in the Top 40 (16 systems):

147,456 cores (Super MUC) in Germany (6 <sup>th</sup> )	122,400 cores (Roadrunner) at LANL (22 <sup>nd</sup> )
204,900 cores (Stampede) at TACC (7 <sup>th</sup> )	53,504 (PRIMERGY) at Australia/NCI (24 <sup>th</sup> )
77,184 cores (Curie thin nodes) at France/CEA (11 <sup>th</sup> )	78,660 cores (Lomonosov) in Russia (26 <sup>th</sup> )
120,640 cores (Nebulae) at China/NSCS (12 <sup>th</sup> )	137,200 cores (Sunway Blue Light) in China (28 <sup>th</sup> )
72,288 cores (Yellowstone) at NCAR (13 <sup>th</sup> )	46,208 cores (Zin) at LLNL (29 <sup>th</sup> )
125,980 cores (Pleiades) at NASA/Ames (14 <sup>th</sup> )	33,664 (MareNostrum) at Spain/BSC (36 <sup>th</sup> )
70,560 cores (Helios) at Japan/IFERC (15 <sup>th</sup> )	32,256 (SGI Altix X) at Japan/CRIEPI (39 <sup>th</sup> )
73,278 cores (Tsubame 2.0) at Japan/GSIC (17 <sup>th</sup> )	<b>More are getting installed !</b>
138,368 cores (Tera-100) at France/CEA (20 <sup>th</sup> )	

# HSE Scientific Computing Installations

- HSE compute systems with ranking in the Nov'12 Top500 list
  - 32,256-core installation in United States (#48)
  - 25,568-core installation in United States (#60)
  - 17,024-core installation at the Amazon EC2 Cluster (#102)
  - 15,369-core installation in United States (#112)
  - 14,272-core installation in United States (#122)
  - 16,064-core installation in United States (#147, #148)
  - 9,488-core installation in United States (#173)
  - 9,216-core installation in United States (#191)
  - 12,032-core installation in United States (#194, #195, #196, #197)
  - 8,572-core installation in India (#199, #200)
  - 16,224-core installation in United States (#226)
  - 7,680-core installation in United States (#247, #248)
  - 19,504-core installation in United States (#254 to #260)
  - 8,000-core installation in United States (#263, #264)
  - 11,712-core installation in United States (#332)
  - 8,640-core installation at Columbia University, United States (#339)
  - 16,512-core installation in United States (#341)
- Integrated Systems
  - BG/P uses 10GE for I/O (ranks 47, 54, 126, 225 and 328 in the Top 500)

## Other HSE Installations

- HSE has most of its popularity in enterprise computing and other non-scientific markets including Wide-area networking
- Example Enterprise Computing Domains
  - Enterprise Datacenters (HP, Intel)
  - Animation firms (e.g., Universal Studios (“The Hulk”), 20<sup>th</sup> Century Fox (“Avatar”), and many new movies using 10GE)
  - Amazon’s HPC cloud offering uses 10GE internally
  - Heavily used in financial markets (users are typically undisclosed)
- Many Network-attached Storage devices come integrated with 10GE network adapters
- ESnet is installing 100GE infrastructure for US DOE

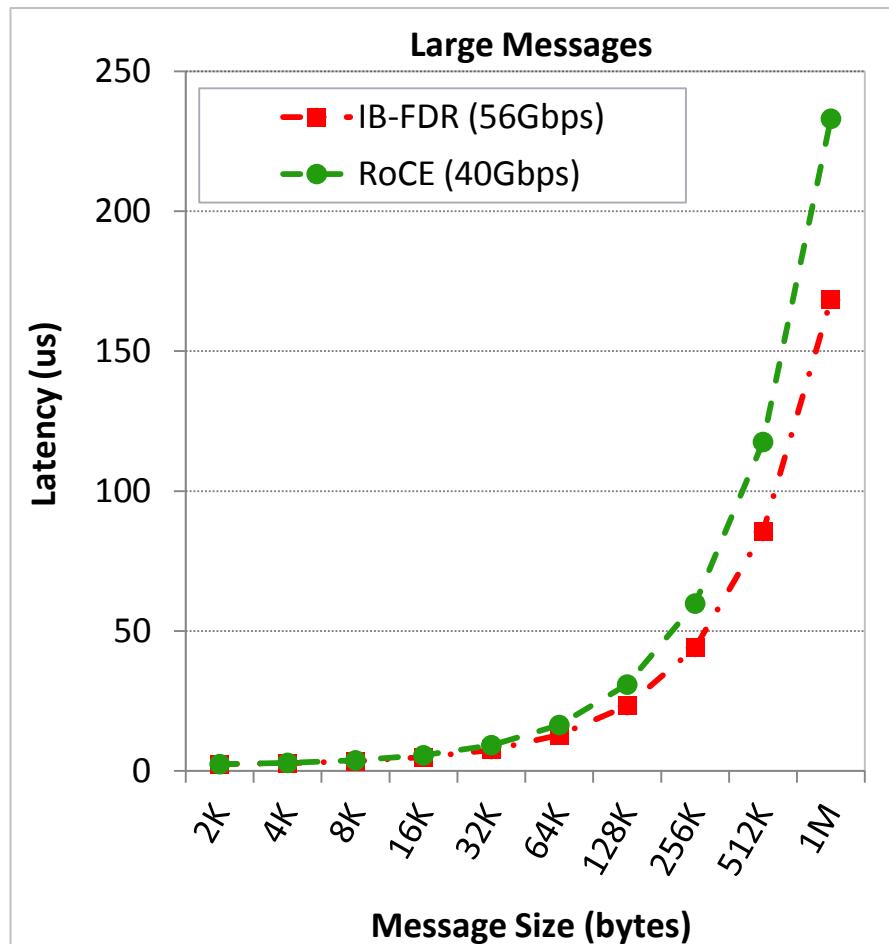
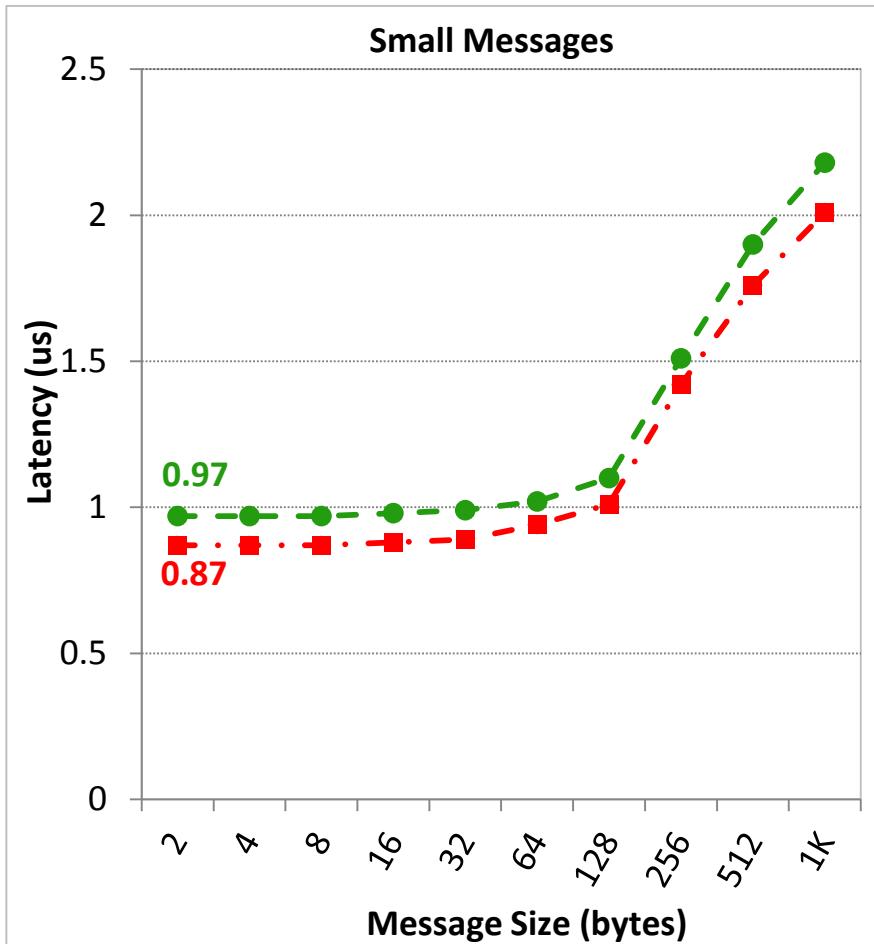
# Presentation Overview

- Introduction
- Why InfiniBand and High-speed Ethernet?
- Overview of IB, HSE, their Convergence and Features
- IB and HSE HW/SW Products and Installations
- **Sample Case Studies and Performance Numbers**
- Conclusions and Final Q&A

# Case Studies

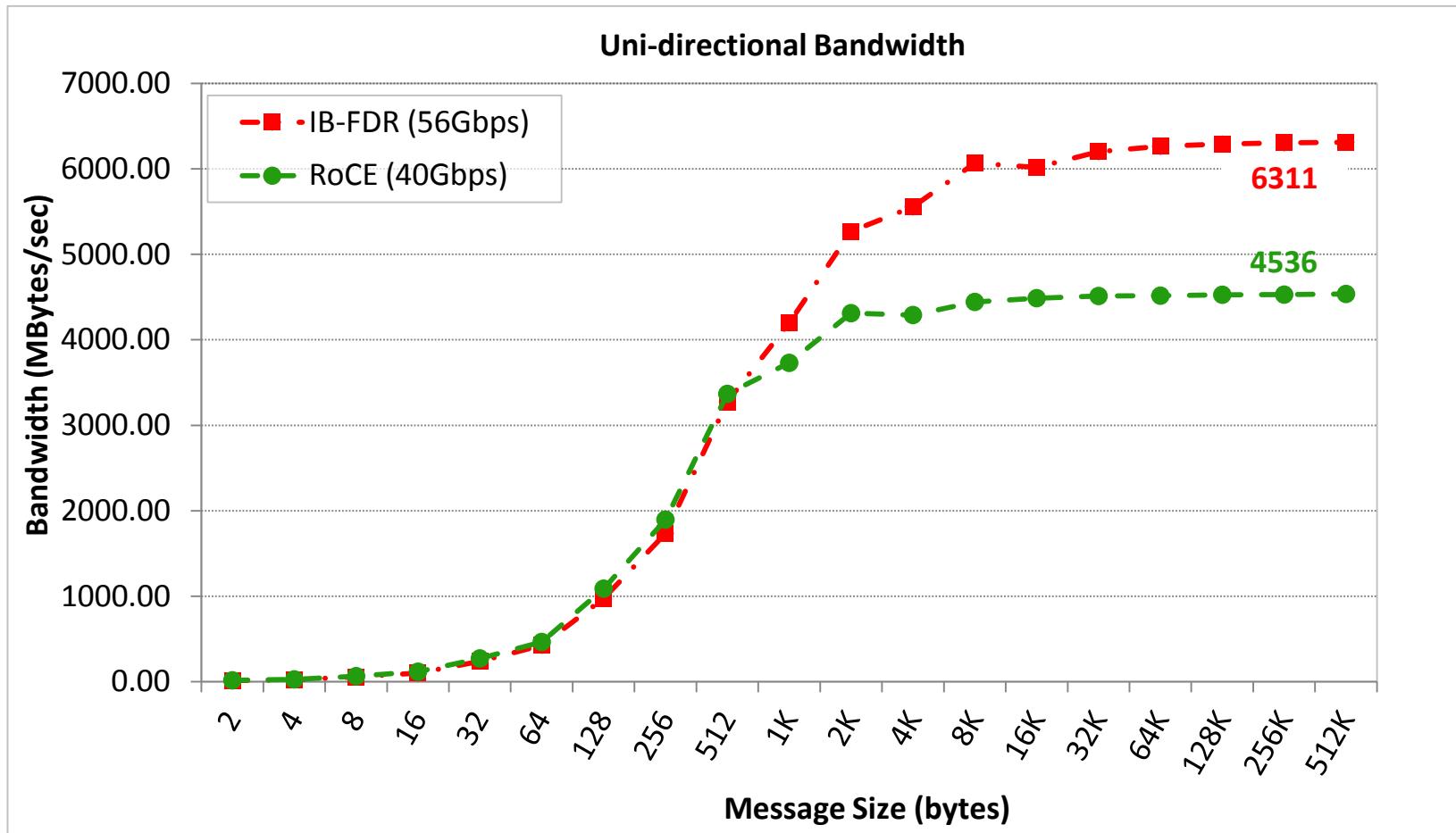
- **Low-level Performance**
- Message Passing Interface (MPI)

# Low-level Latency Measurements



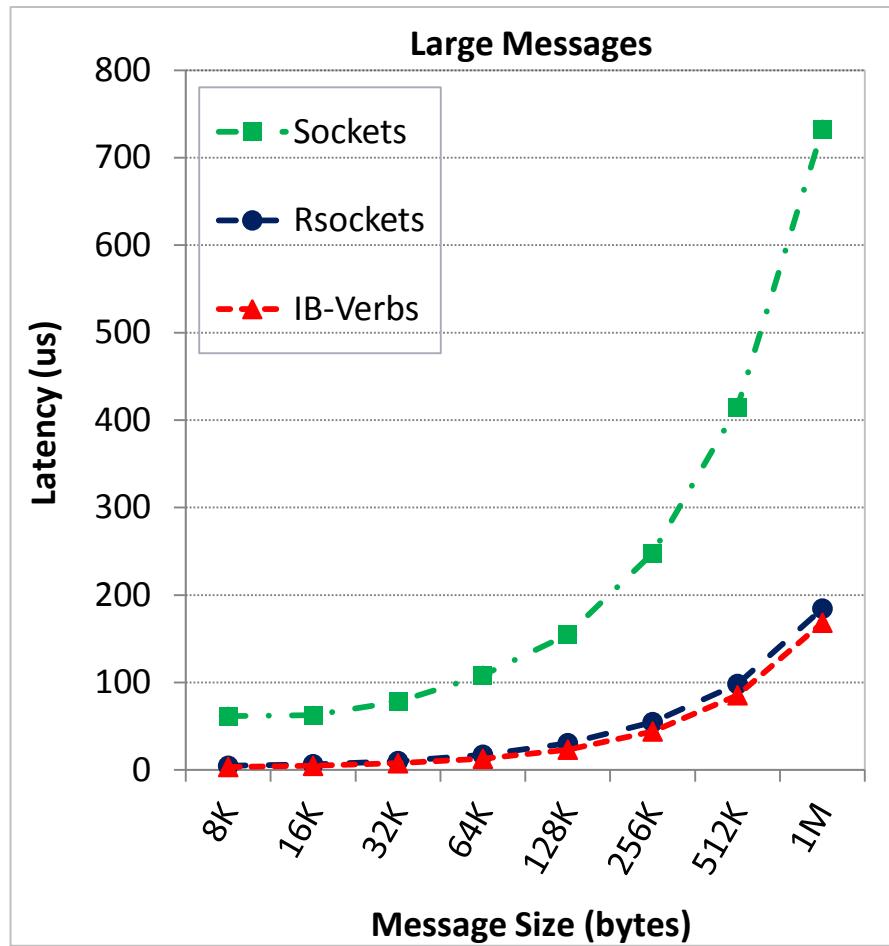
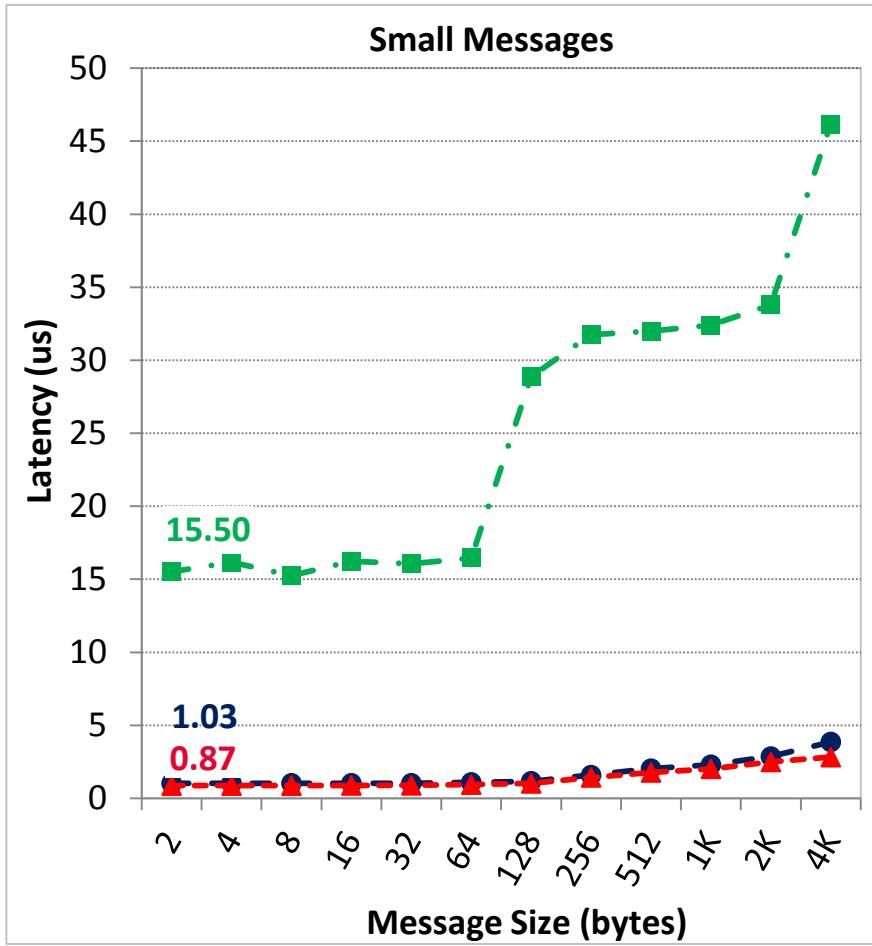
ConnectX-3 FDR (54 Gbps): 2.6 GHz Octa-core (SandyBridge) Intel with IB (FDR) switches  
ConnectX-3 EN (40 GigE): 2.6 GHz Octa-core (SandyBridge) Intel with 40GE switches

# Low-level Uni-directional Bandwidth Measurements



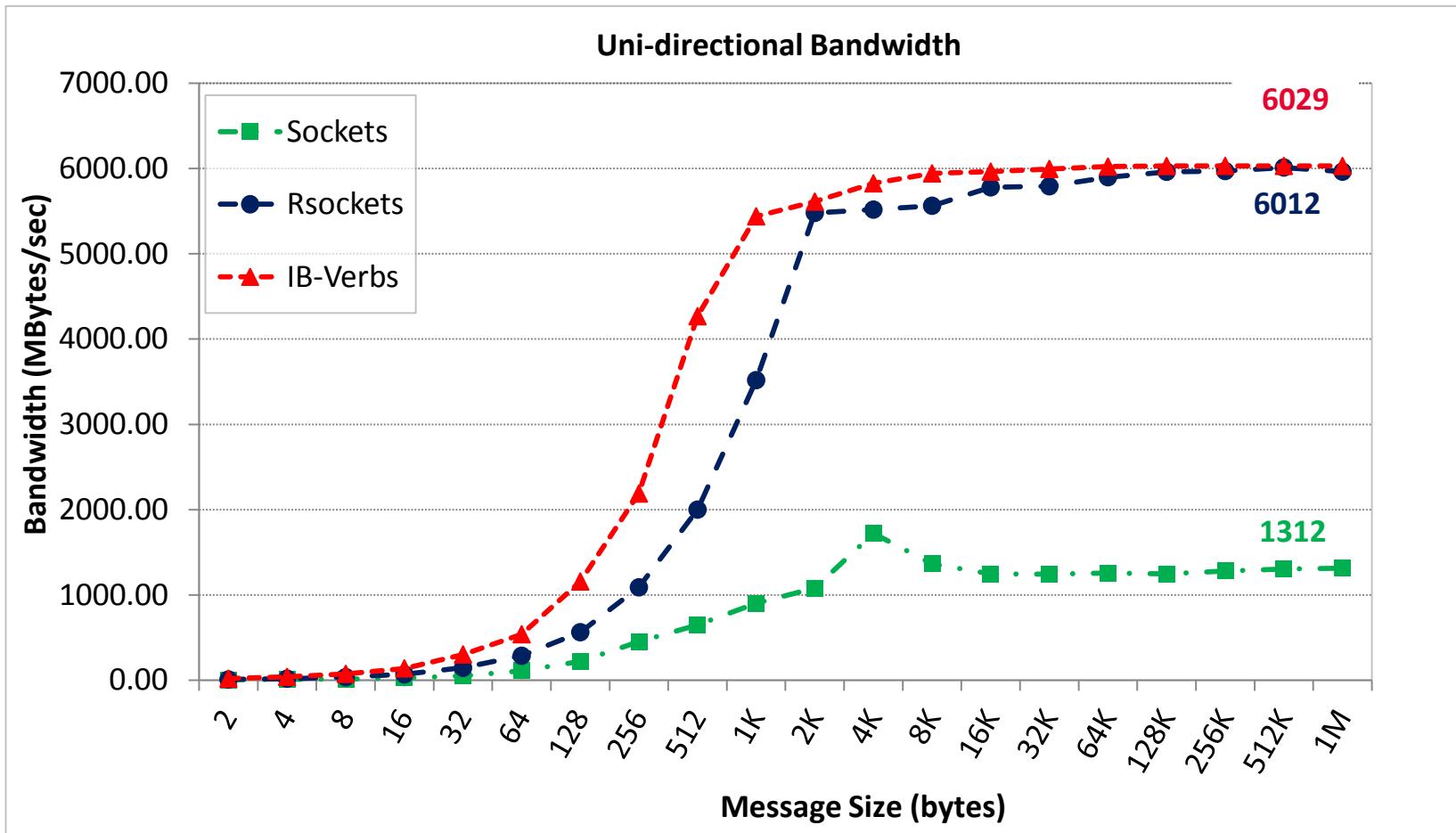
ConnectX-3 FDR (54 Gbps): 2.6 GHz Octa-core (SandyBridge) Intel with IB (FDR) switches  
ConnectX-3 EN (40 GigE): 2.6 GHz Octa-core (SandyBridge) Intel with 40GE switches

# Low-level Latency Measurements



ConnectX-3 FDR (54 Gbps): 2.6 GHz Octa-core (SandyBridge) Intel with IB (FDR) switches

# Low-level Uni-directional Bandwidth Measurements



ConnectX-3 FDR (54 Gbps): 2.6 GHz Octa-core (SandyBridge) Intel with IB (FDR) switches

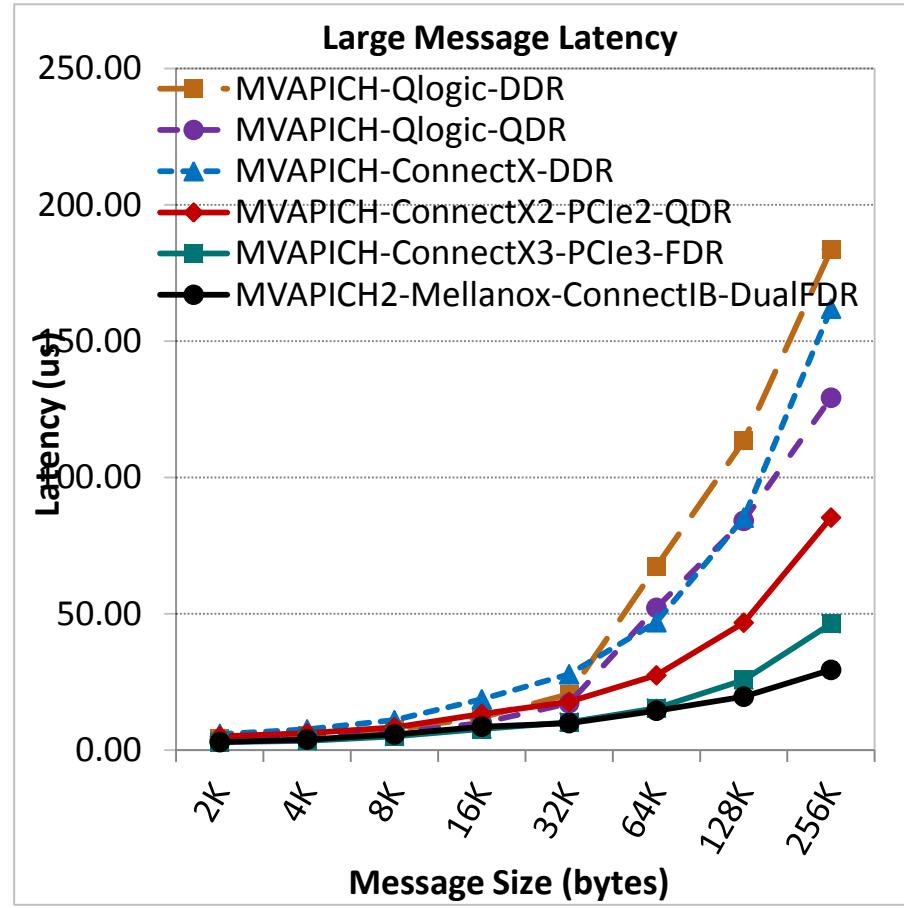
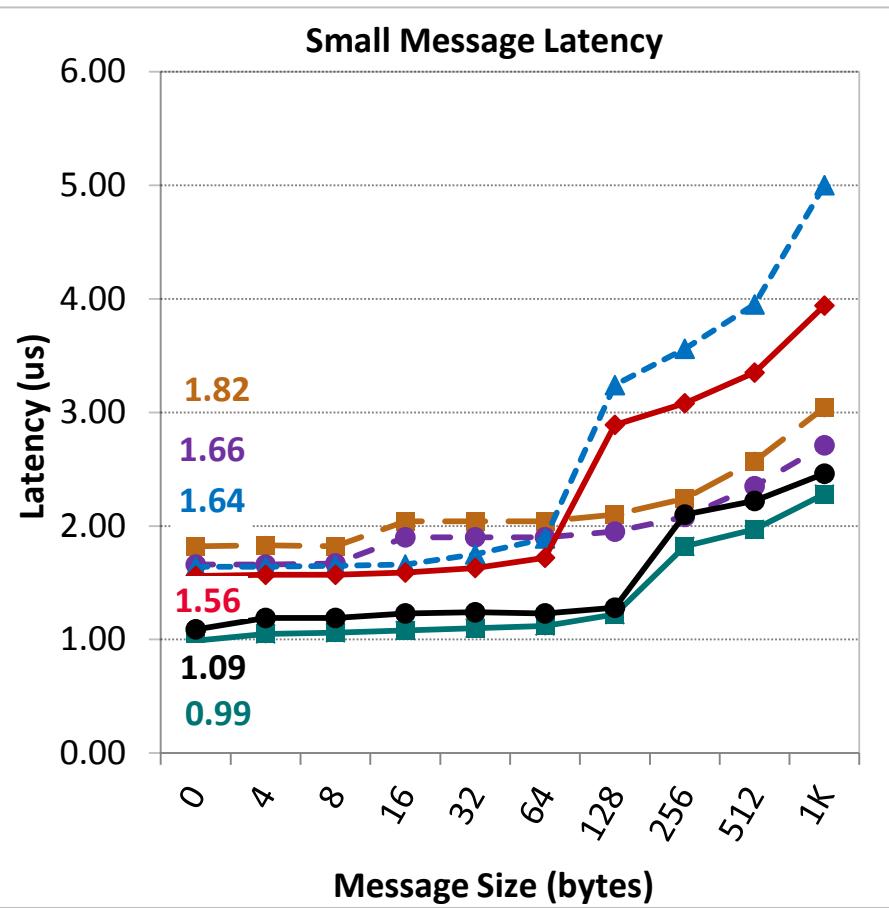
# Case Studies

- Low-level Performance
- **Message Passing Interface (MPI)**

# MVAPICH2/MVAPICH2-X Software

- High Performance open-source MPI Library for InfiniBand, 10Gig/iWARP and RDMA over Converged Enhanced Ethernet (RoCE)
  - MVAPICH (MPI-1) ,MVAPICH2 (MPI-3.0), Available since 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2012
  - Used by more than 2,000 organizations (HPC Centers, Industry and Universities) in 70 countries
  - More than 173,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 7<sup>th</sup> ranked 204,900-core cluster (Stampede) at TACC
    - 14<sup>th</sup> ranked 125,980-core cluster (Pleiades) at NASA
    - 17<sup>th</sup> ranked 73,278-core cluster (Tsubame 2.0) at Tokyo Institute of Technology
    - and many others
  - Available with software stacks of many IB, HSE and server vendors including Linux Distros (RedHat and SuSE)
  - <http://mvapich.cse.ohio-state.edu>
- Partner in the U.S. NSF-TACC Stampede System

# One-way Latency: MPI over IB

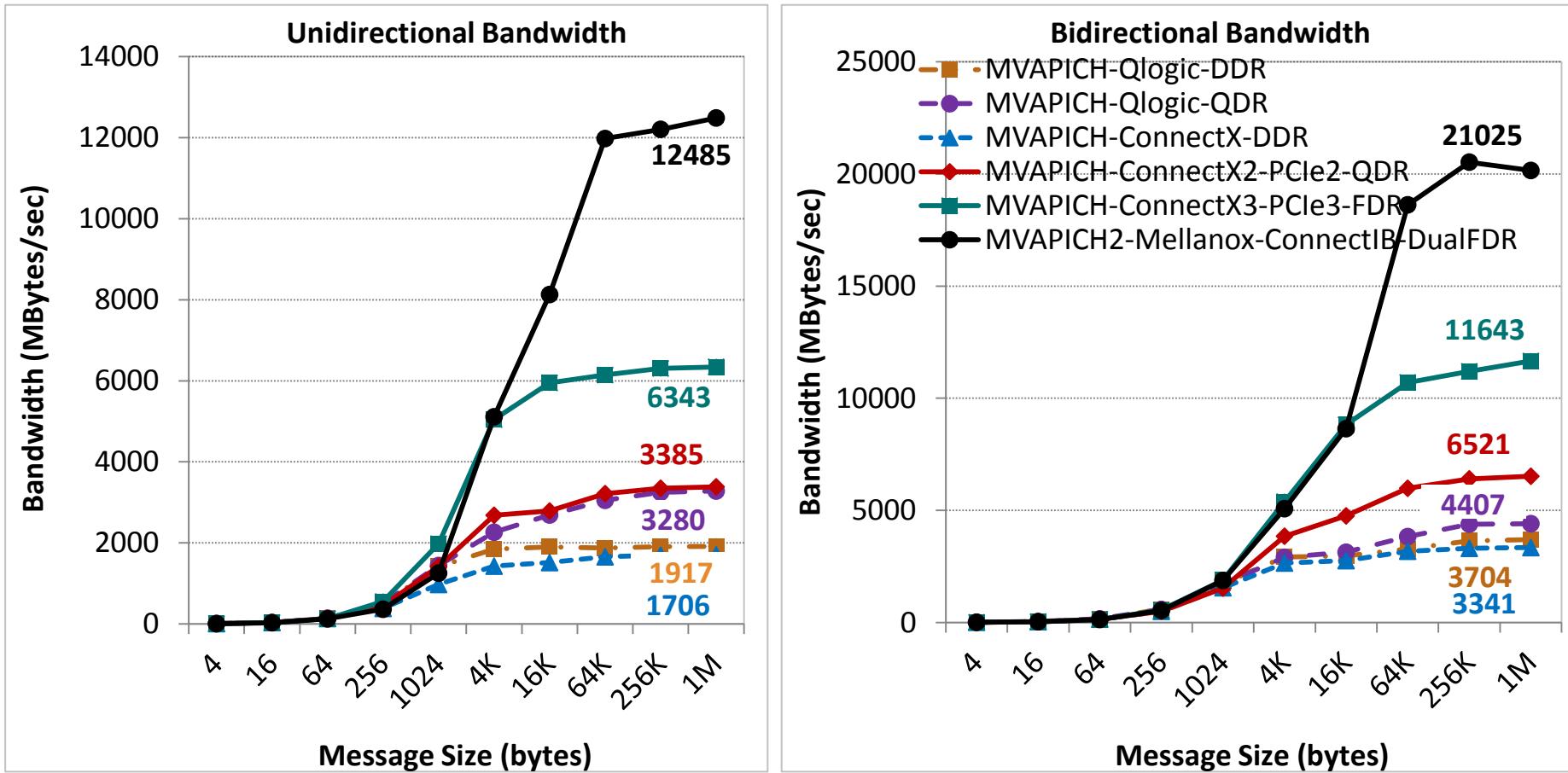


DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch

FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch

ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch

# Bandwidth: MPI over IB

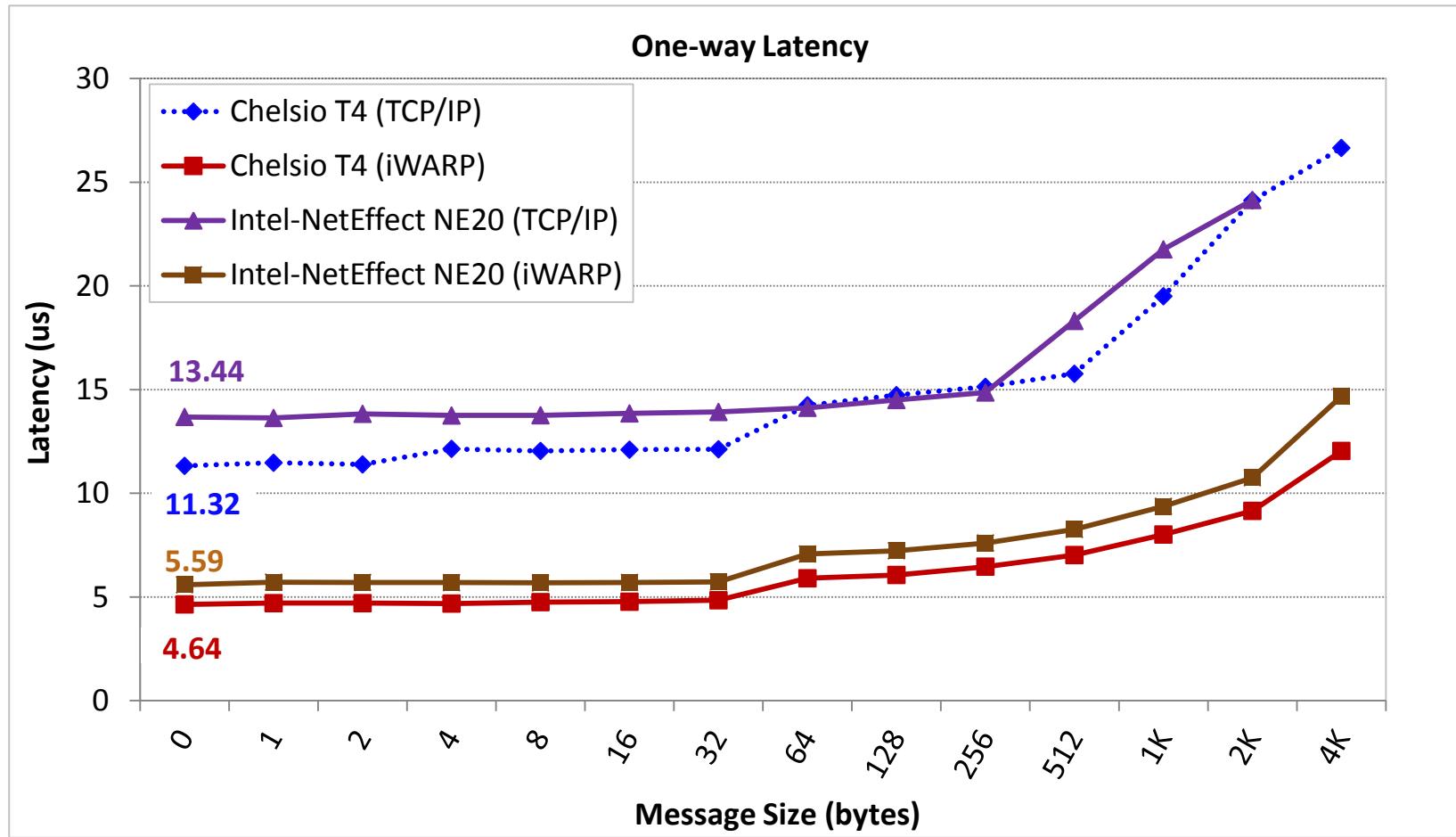


DDR, QDR - 2.4 GHz Quad-core (Westmere) Intel PCI Gen2 with IB switch

FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch

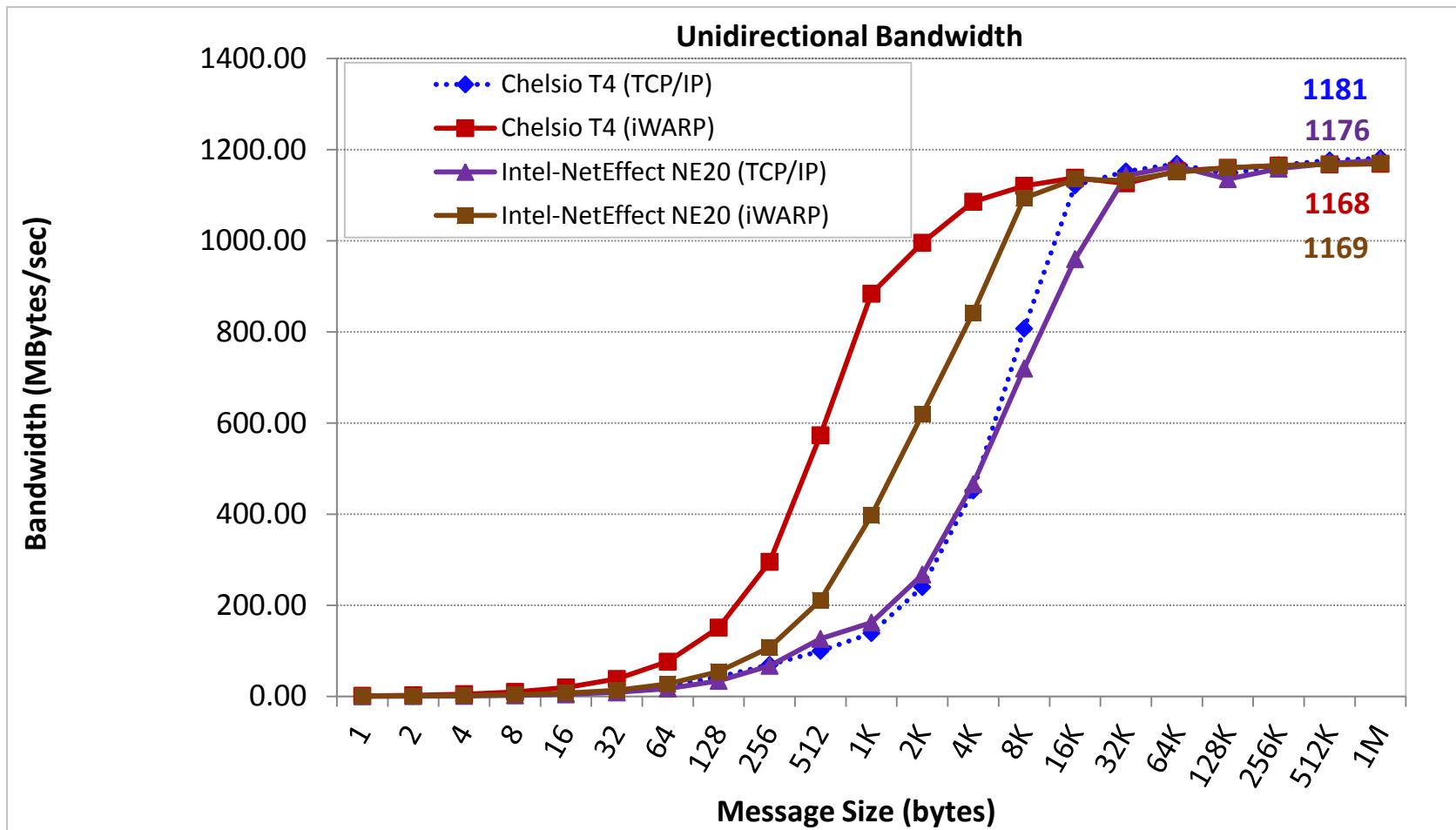
ConnectIB-Dual FDR - 2.6 GHz Octa-core (SandyBridge) Intel PCI Gen3 with IB switch

# One-way Latency: MPI over iWARP



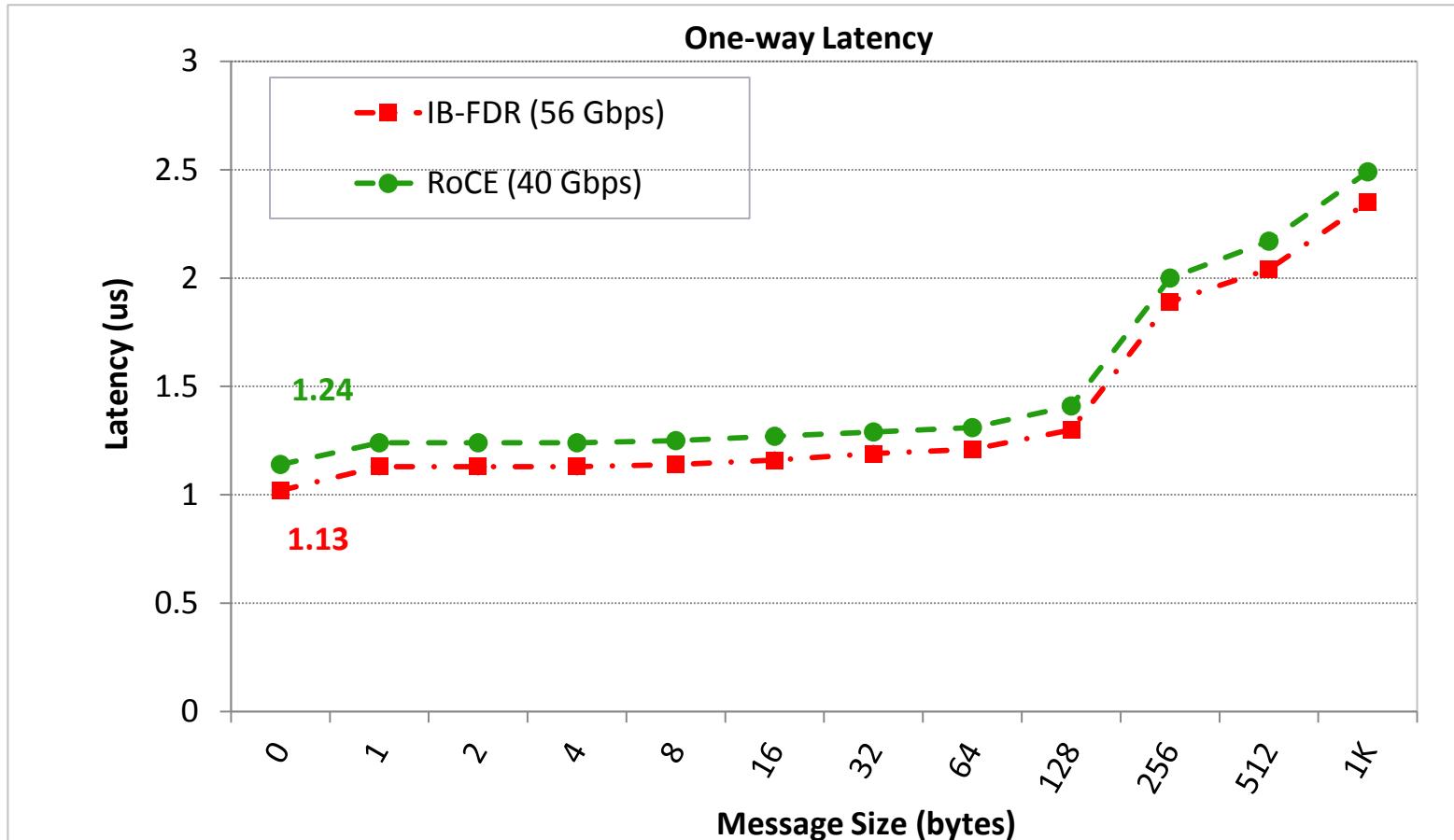
2.6 GHz Dual Eight-core (SandyBridge) Intel  
Chelsio T4 cards connected through Fujitsu xg2600 10GigE switch  
Intel NetEffect cards connected through Fulcrum 10GigE switch

# Bandwidth: MPI over iWARP



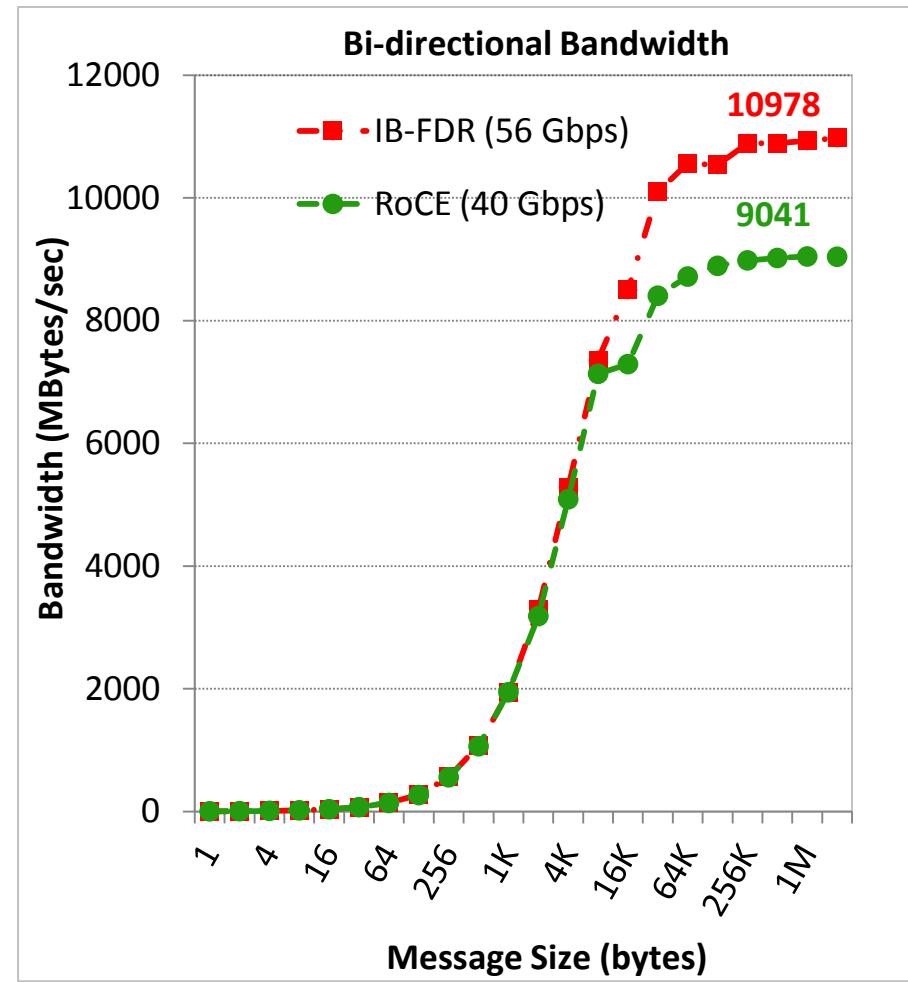
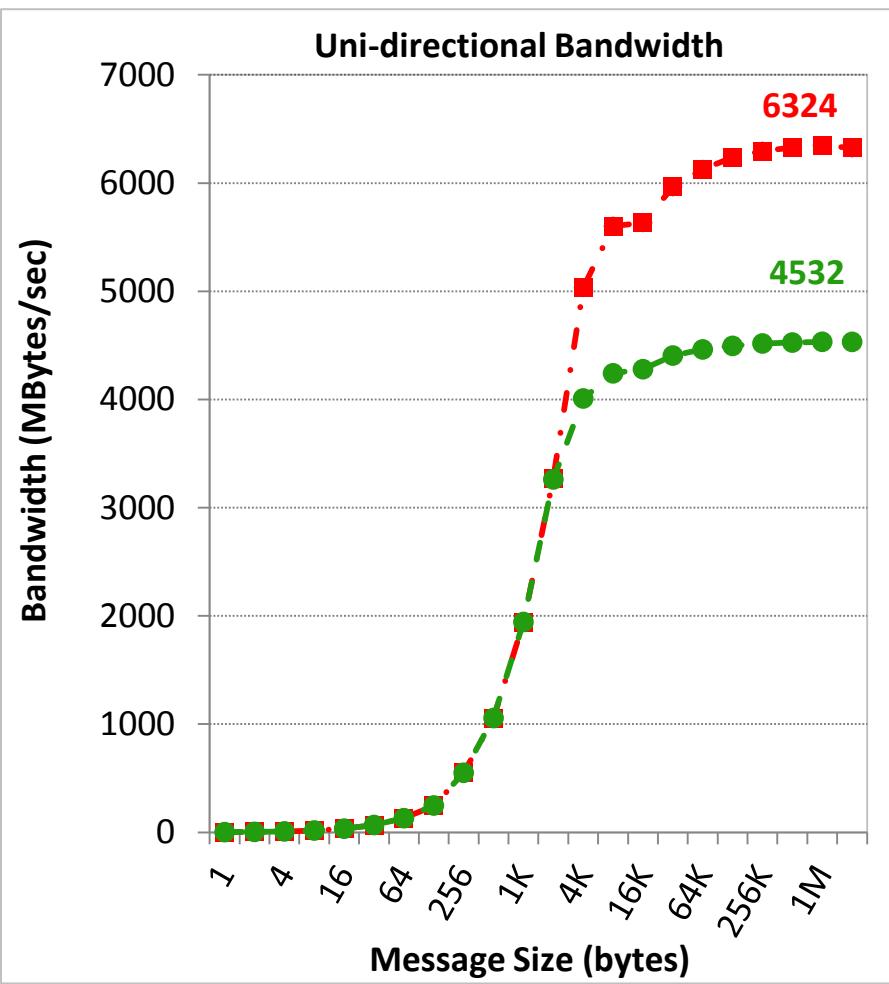
2.6 GHz Dual Eight-core (SandyBridge) Intel  
Chelsio T4 cards connected through Fujitsu xg2600 10GigE switch  
Intel NetEffect cards connected through Fulcrum 10GigE switch

# Convergent Technologies: MPI Latency



ConnectX-3 FDR (54 Gbps): 2.6 GHz Octa-core (SandyBridge) Intel with IB (FDR) switches  
ConnectX-3 EN (40 GigE): 2.6 GHz Octa-core (SandyBridge) Intel with 40GE switches

# Convergent Technologies: MPI Uni- and Bi-directional Bandwidth



ConnectX-3 FDR (54 Gbps): 2.6 GHz Octa-core (SandyBridge) Intel with IB (FDR) switches  
ConnectX-3 EN (40 GigE): 2.6 GHz Octa-core (SandyBridge) Intel with 40GE switches

# Presentation Overview

- Introduction
- Why InfiniBand and High-speed Ethernet?
- Overview of IB, HSE, their Convergence and Features
- IB and HSE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
- **Conclusions and Final Q&A**

# Concluding Remarks

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

# Funding Acknowledgments

*Funding Support by*



*Equipment Support by*



# Personnel Acknowledgments

## Current Students

- N. Islam (Ph.D.)
- J. Jose (Ph.D.)
- K. Kandalla (Ph.D.)
- M. Li (Ph.D.)
- M. Luo (Ph.D.)
- S. Potluri (Ph.D.)
- R. Rajachandrakar (Ph.D.)
- M. Rahman (Ph.D.)
- R. Shir (Ph.D.)
- H. Subramoni (Ph.D.)
- A. Venkatesh (Ph.D.)

## Past Students

- P. Balaji (Ph.D.)
- D. Buntinas (Ph.D.)
- S. Bhagvat (M.S.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)
- W. Huang (Ph.D.)
- W. Jiang (M.S.)
- S. Kini (M.S.)
- M. Koop (Ph.D.)
- R. Kumar (M.S.)
- S. Krishnamoorthy (M.S.)
- P. Lai (M.S.)

## Current Post-Docs

- X. Lu
- K. Hamidouche

## Current Programmers

- M. Arnold
- D. Bureddy
- J. Perkins

## Past Post-Docs

- H. Wang
- X. Besseron
- H.-W. Jin
- E. Mancini
- S. Marcarelli
- J. Vienne

## Past Research Scientist

- S. Sur

# Web Pointers

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

<http://www.cse.ohio-state.edu/~subramon>

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

MVAPICH Web Page

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



[panda@cse.ohio-state.edu](mailto:panda@cse.ohio-state.edu)

[subramon@cse.ohio-state.edu](mailto:subramon@cse.ohio-state.edu)

# MVAPICH User Group (MUG) Meeting

## August 26-27, 2013, Columbus, Ohio, U.S.A

- The MUG meeting will provide an open forum for all attendees (users, researchers, system administrators, engineers, and students) to share their knowledge about MVAPICH2/MVAPICH2-X on large-scale systems and diverse applications.
- The event includes:
  - Talks from experts in the field
  - Presentations from the MVAPICH team on tuning and optimization strategies
  - Troubleshooting guidelines
  - Contributed presentations
  - Open mic session
  - Interactive one-on-one session with the MVAPICH developers

### Call for Presentation

- The MVAPICH team is requesting the submission of presentations from MVAPICH2 and MVAPICH2-X users to be included in the event.

**Presentation Submission Deadline: July 1, 2013**

**Notification of Acceptance: July 8, 2013**

**Advanced Registration Deadline: July 15, 2013**

The preliminary program has been posted at [mug.mvapich.cse.ohio-state.edu/program/](http://mug.mvapich.cse.ohio-state.edu/program/)