Designing Cloud and Grid Computing Systems with InfiniBand and High-Speed Ethernet

A Tutorial at CCGrid '11

by

Dhabaleswar K. (DK) Panda

The Ohio State University

E-mail: panda@cse.ohio-state.edu

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

Sayantan Sur

The Ohio State University

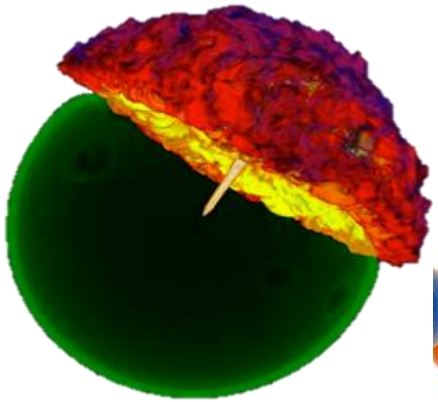
E-mail: surs@cse.ohio-state.edu

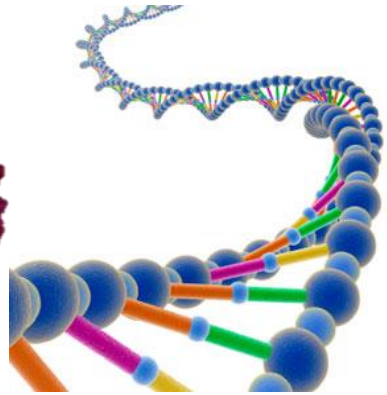
http://www.cse.ohio-state.edu/~surs

Presentation Overview

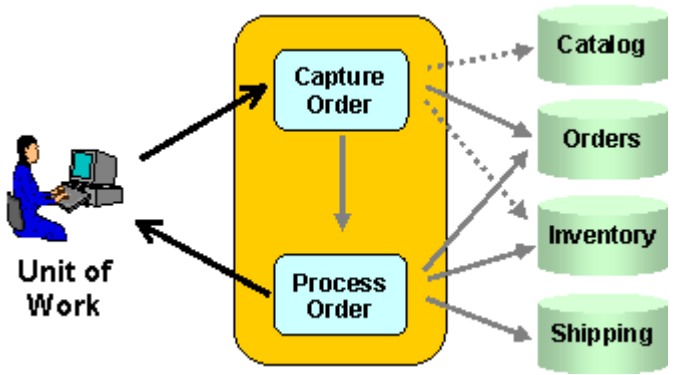
- **Introduction**
- Why InfiniBand and High-speed Ethernet?
- Overview of IB, HSE, their Convergence and Features
- IB and HSE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
- Conclusions and Final Q&A

Current and Next Generation Applications and Computing Systems

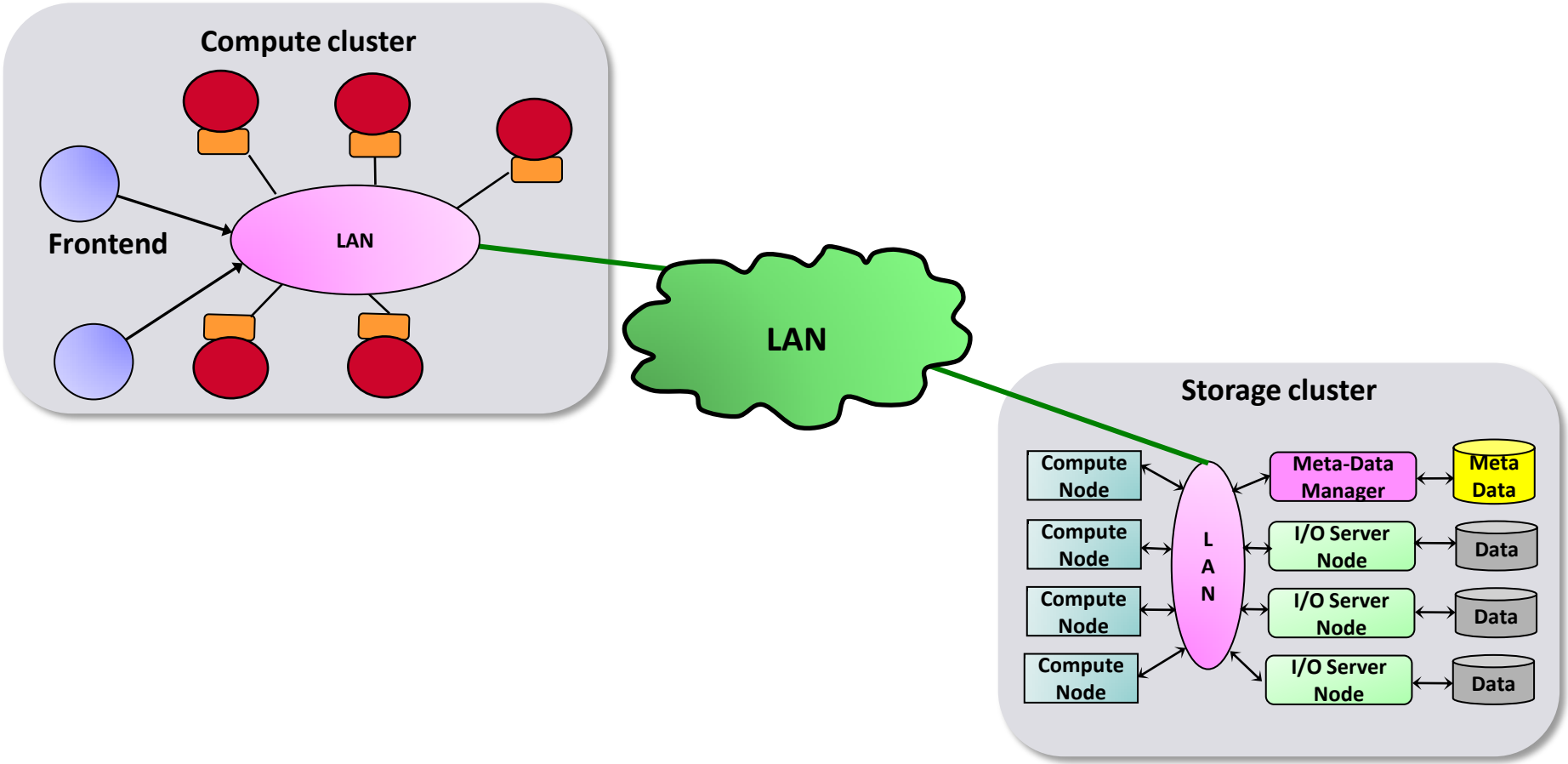




- **Diverse Range of Applications**
  - Processing and dataset characteristics vary
- **Growth of High Performance Computing**
  - Growth in processor performance
    - Chip density doubles every 18 months
  - Growth in commodity networking
    - Increase in speed/features + reducing cost
- **Different Kinds of Systems**
  - Clusters, Grid, Cloud, Datacenters, …..



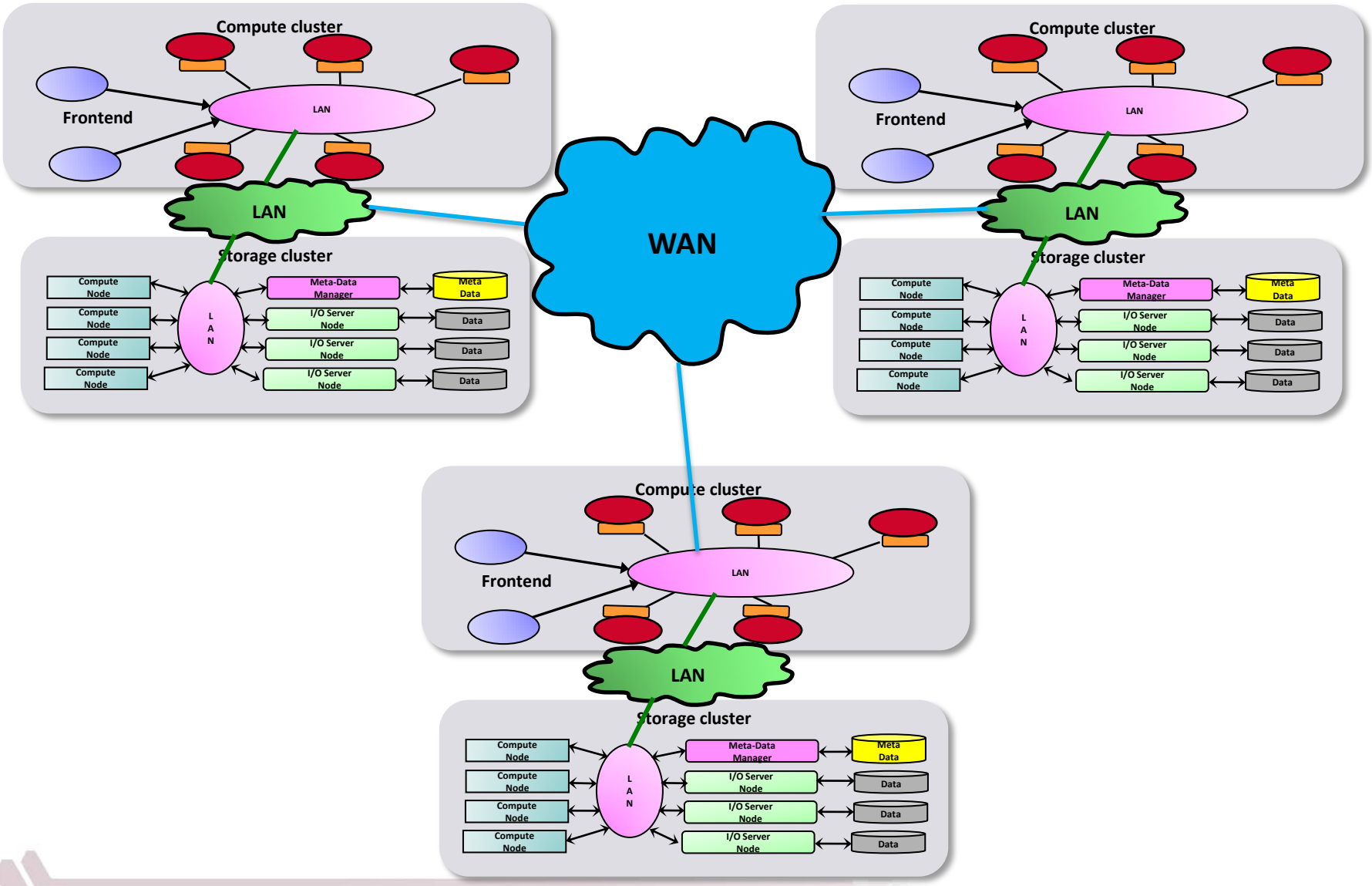
Cluster Computing Environment



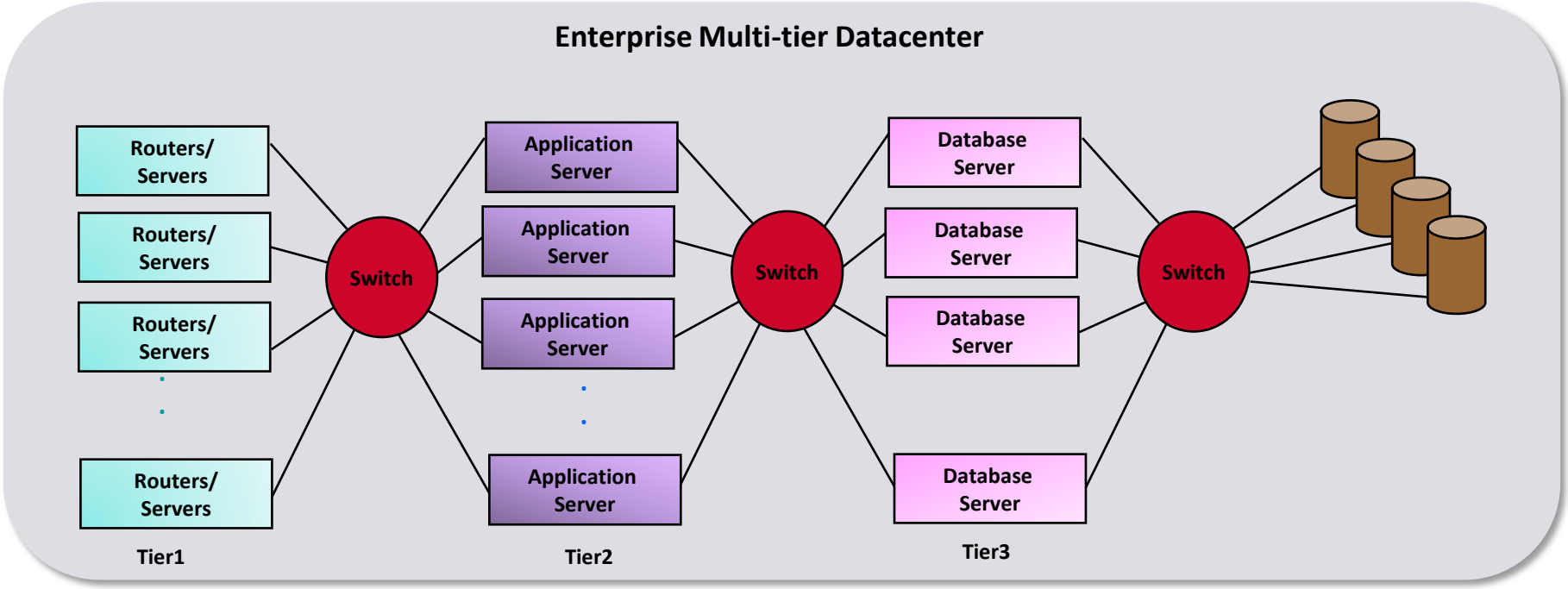
Trends for Computing Clusters in the Top 500 List (http://www.top500.org)

Nov. 1996: 0/500 (0%)	Nov. 2001: 43/500 (8.6%)	Nov. 2006: 361/500 (72.2%)
Jun. 1997: 1/500 (0.2%)	Jun. 2002: 80/500 (16%)	Jun. 2007: 373/500 (74.6%)
Nov. 1997: 1/500 (0.2%)	Nov. 2002: 93/500 (18.6%)	Nov. 2007: 406/500 (81.2%)
Jun. 1998: 1/500 (0.2%)	Jun. 2003: 149/500 (29.8%)	Jun. 2008: 400/500 (80.0%)
Nov. 1998: 2/500 (0.4%)	Nov. 2003: 208/500 (41.6%)	Nov. 2008: 410/500 (82.0%)
Jun. 1999: 6/500 (1.2%)	Jun. 2004: 291/500 (58.2%)	Jun. 2009: 410/500 (82.0%)
Nov. 1999: 7/500 (1.4%)	Nov. 2004: 294/500 (58.8%)	Nov. 2009: 417/500 (83.4%)
Jun. 2000: 11/500 (2.2%)	Jun. 2005: 304/500 (60.8%)	Jun. 2010: 424/500 (84.8%)
Nov. 2000: 28/500 (5.6%)	Nov. 2005: 360/500 (72.0%)	Nov. 2010: 415/500 (83%)
Jun. 2001: 33/500 (6.6%)	Jun. 2006: 364/500 (72.8%)	Jun. 2011: To be announced

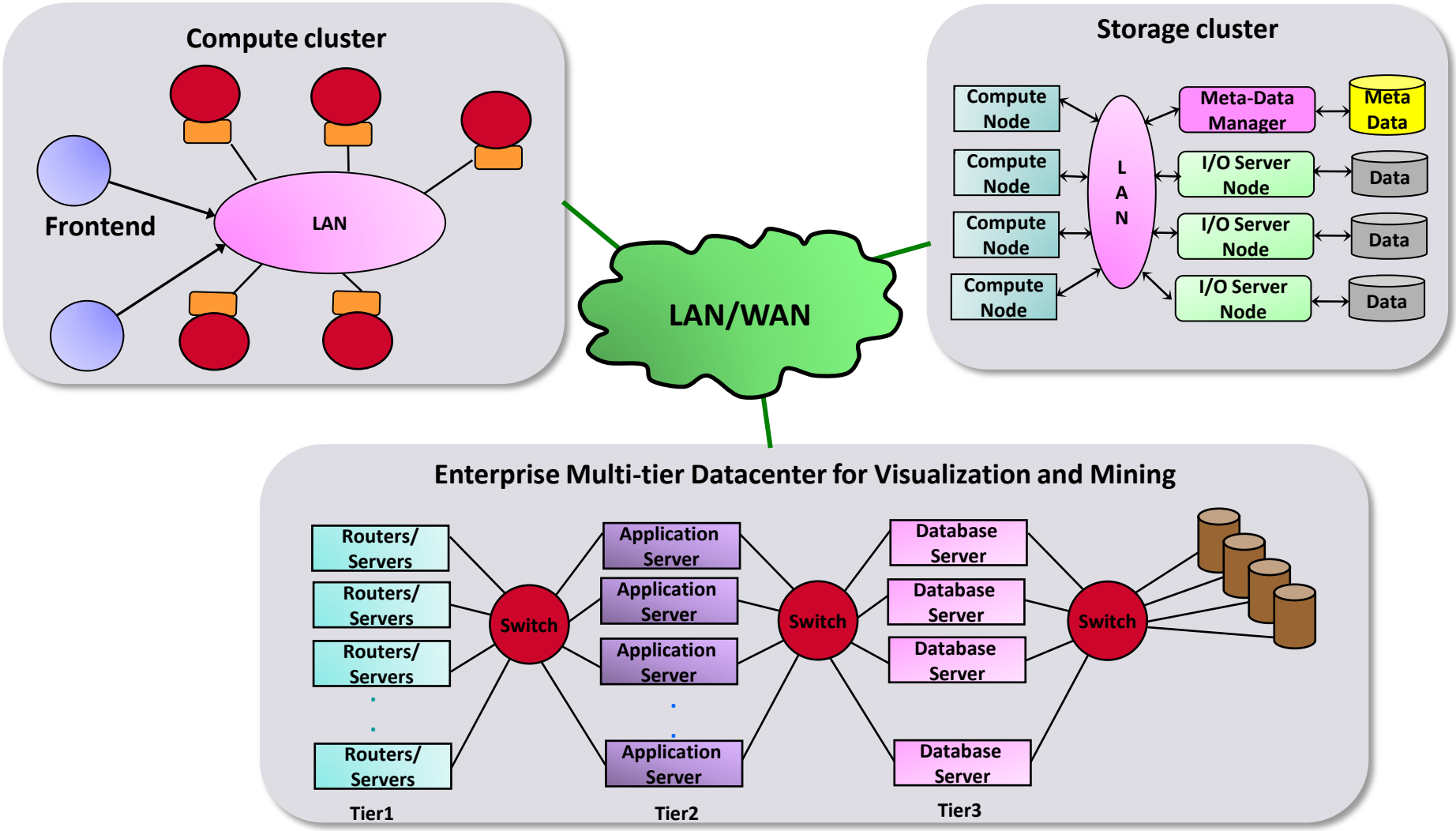
Grid Computing Environment



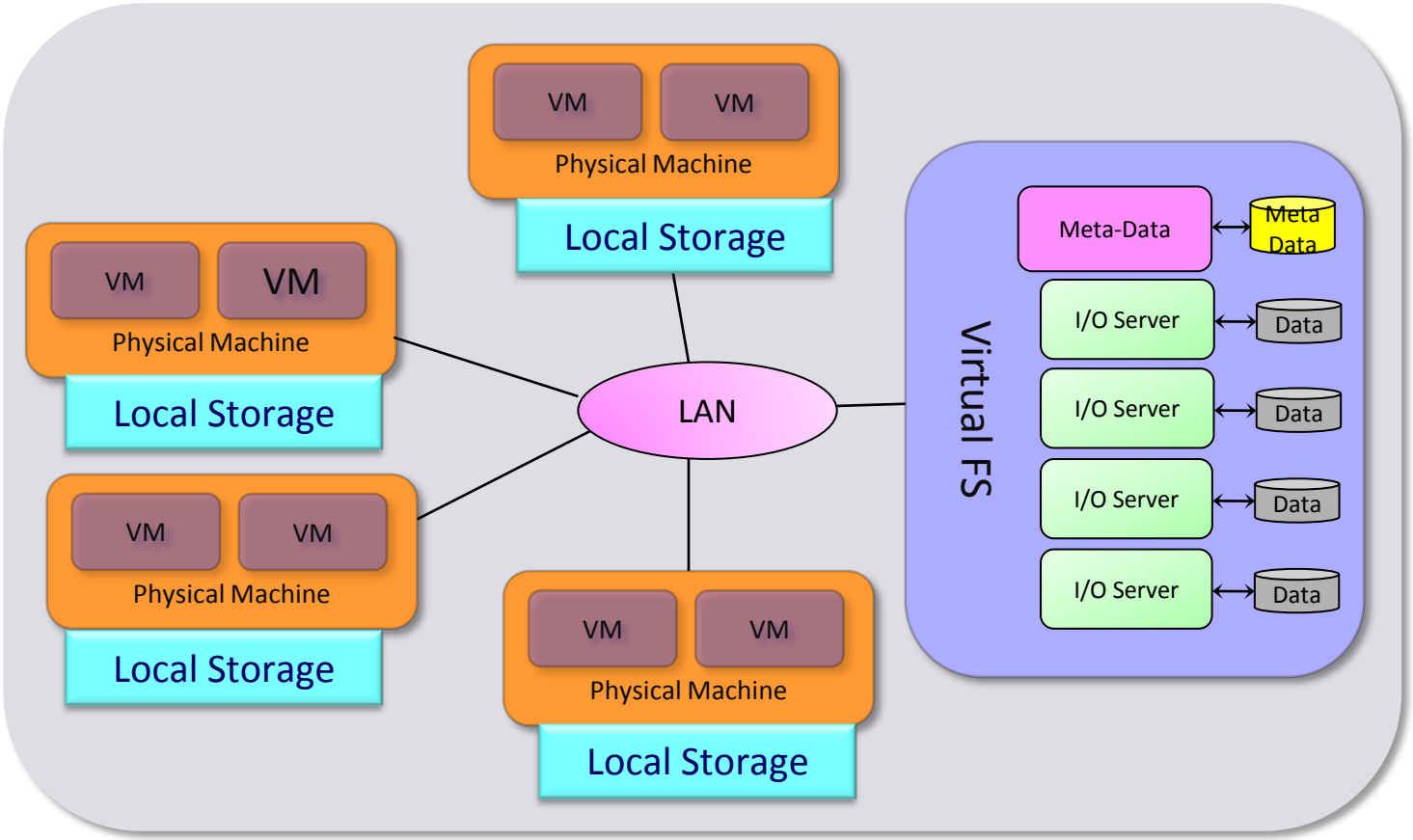
Multi-Tier Datacenters and Enterprise Computing



Integrated High-End Computing Environments

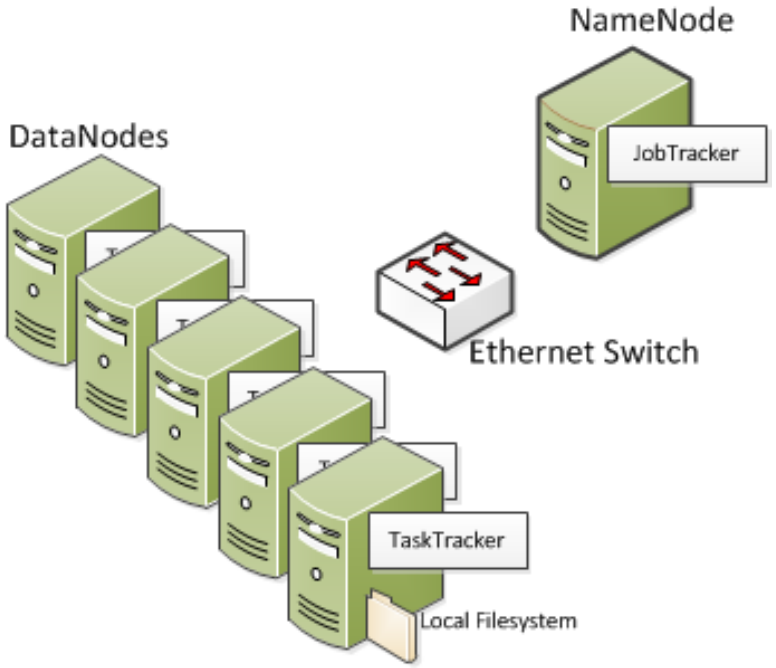


Cloud Computing Environments

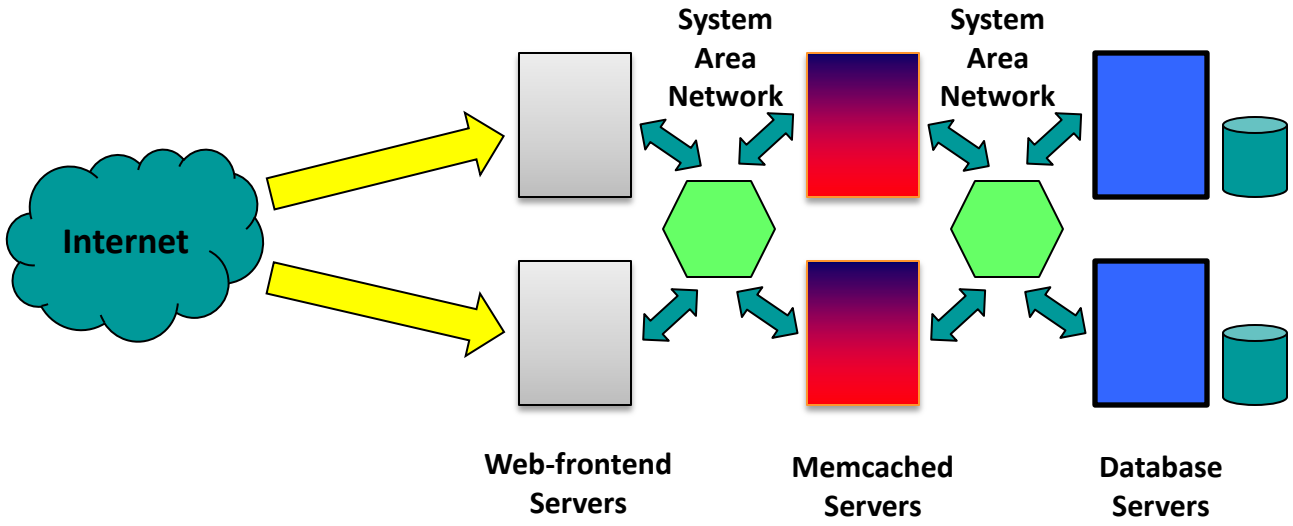


Hadoop Architecture

- 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
- Model scales but high amount of communication during intermediate phases



Memcached Architecture

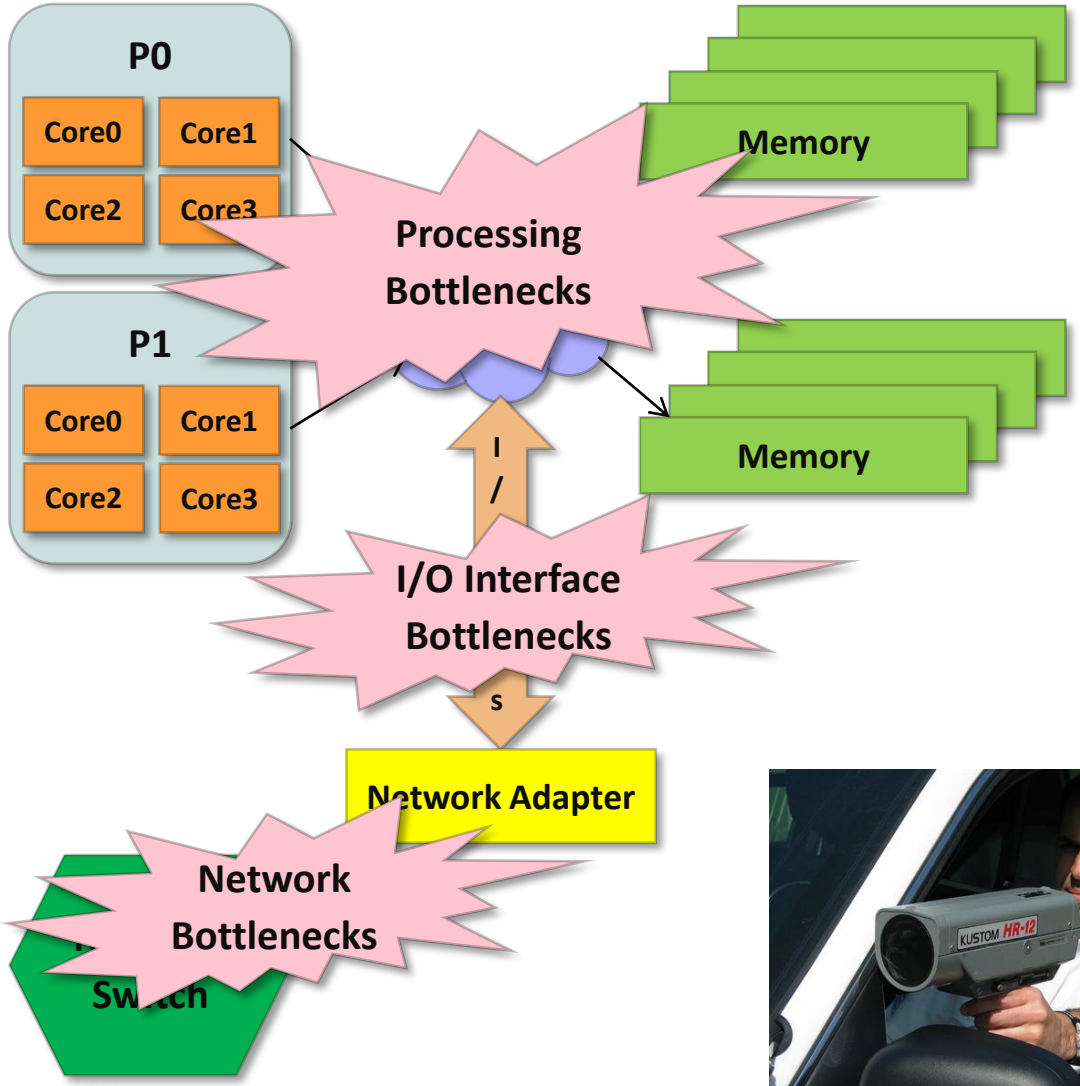


- Distributed Caching Layer
  - Allows to aggregate spare memory from multiple nodes
  - General purpose
- Typically used to cache database queries, results of API calls
- Scalable model, but typical usage very network intensive

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

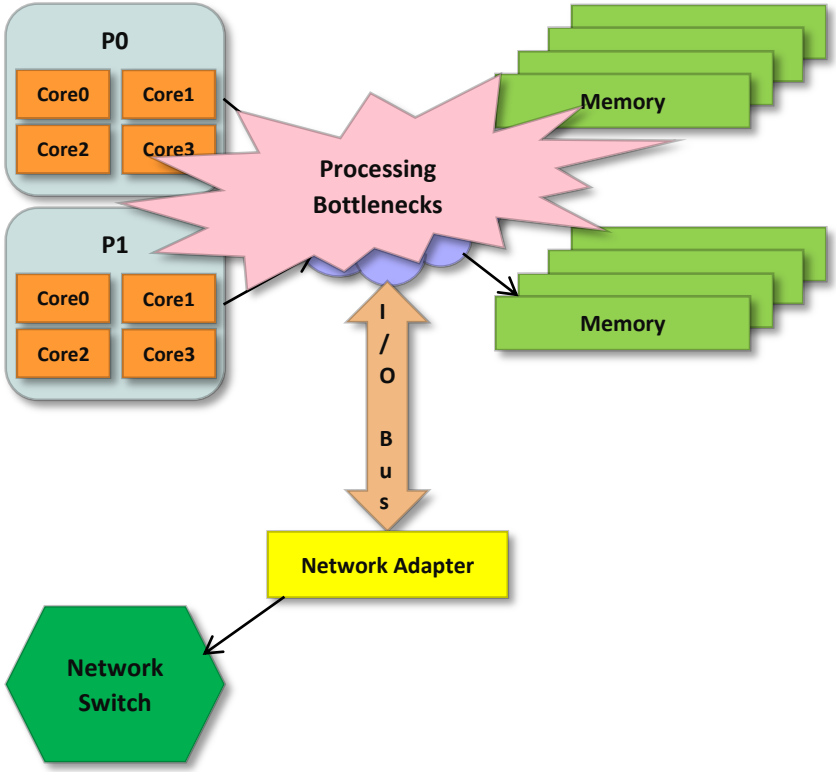
Major Components in Computing Systems



- 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*

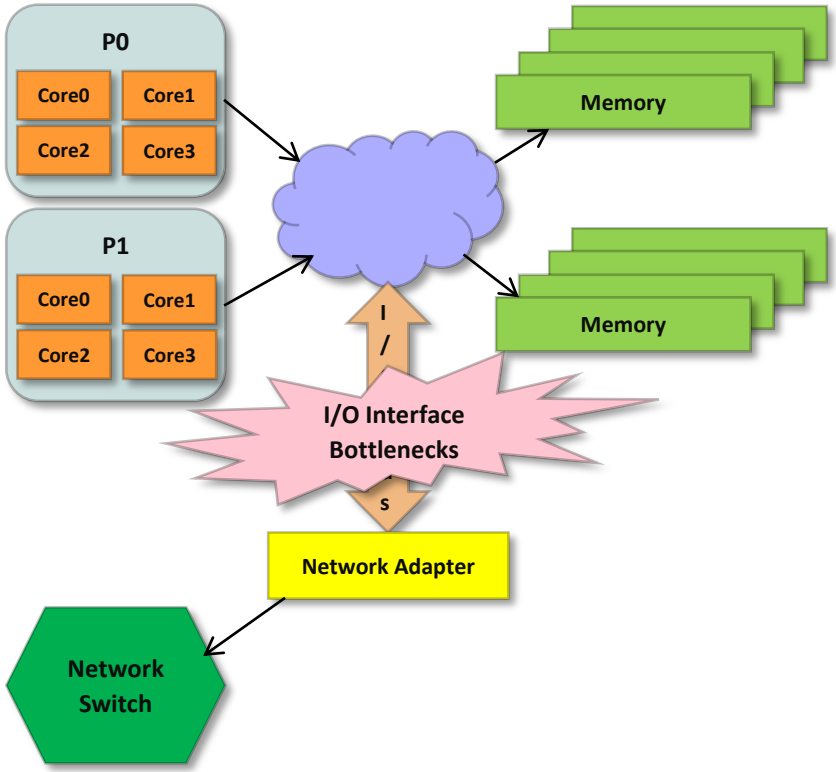
Processing Bottlenecks in Traditional Protocols

- 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

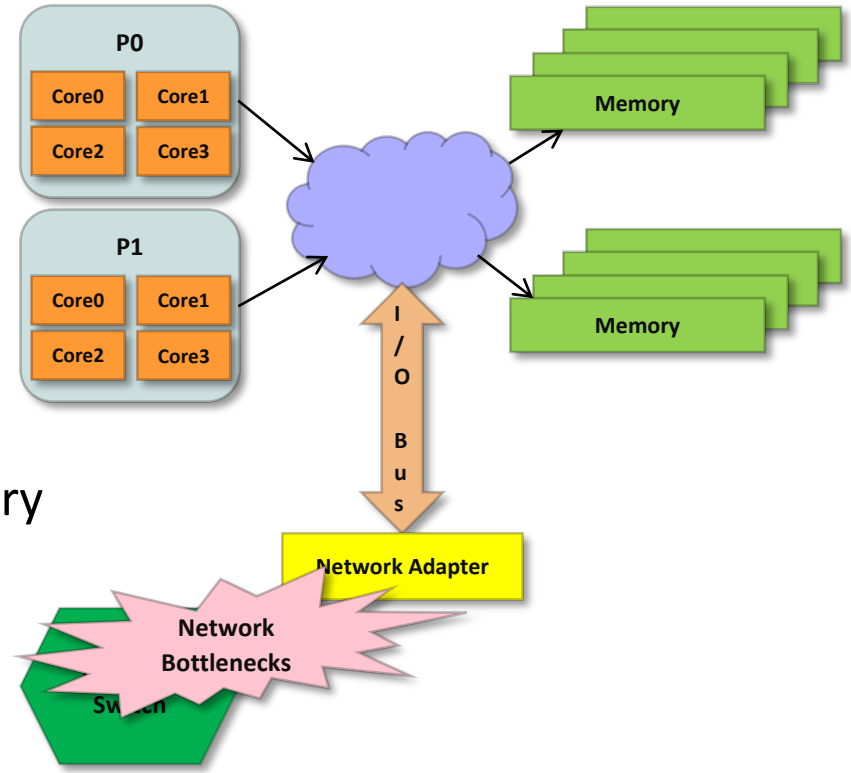
- Traditionally relied on bus-based technologies (last mile bottleneck)
  - E.g., PCI, PCI-X
  - One bit per wire
  - Performance increase through:
    - Increasing clock speed
    - Increasing bus width
  - Not scalable:
    - Cross talk between bits
    - Skew between wires
    - Signal integrity makes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

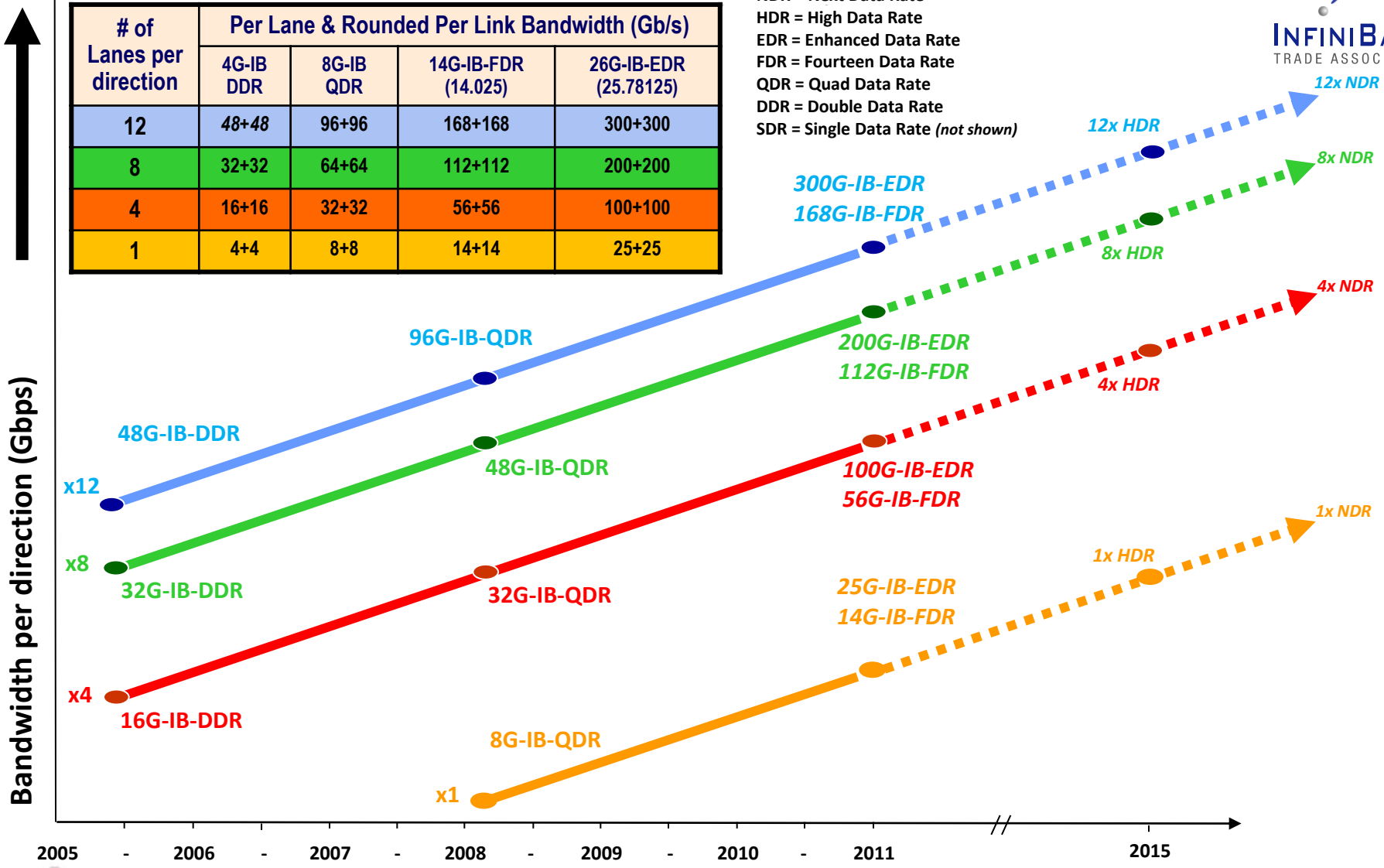
Ethernet (1979 -)	10 Mbit/sec
Fast Ethernet (1993 -)	100 Mbit/sec
Gigabit Ethernet (1995 -)	1000 Mbit /sec
ATM (1995 -)	155/622/1024 Mbit/sec
Myrinet (1993 -)	1 Gbit/sec
Fibre Channel (1994 -)	1 Gbit/sec
InfiniBand (2001 -)	2 Gbit/sec (1X SDR)
10-Gigabit Ethernet (2001 -)	10 Gbit/sec
InfiniBand (2003 -)	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 -)	56 Gbit/sec (4X FDR)
InfiniBand (2012 -)	100 Gbit/sec (4X EDR)

20 times in the last 9 years

InfiniBand Link Speed Standardization Roadmap

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

NDR = Next Data Rate
HDR = High Data Rate
EDR = Enhanced Data Rate
FDR = Fourteen Data Rate
QDR = Quad Data Rate
DDR = Double Data Rate
SDR = Single Data Rate *(not shown)*



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
  - Heavily used in the "server-on-steroids" model
    - High performance servers connected to regular clients

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
  - Support 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)

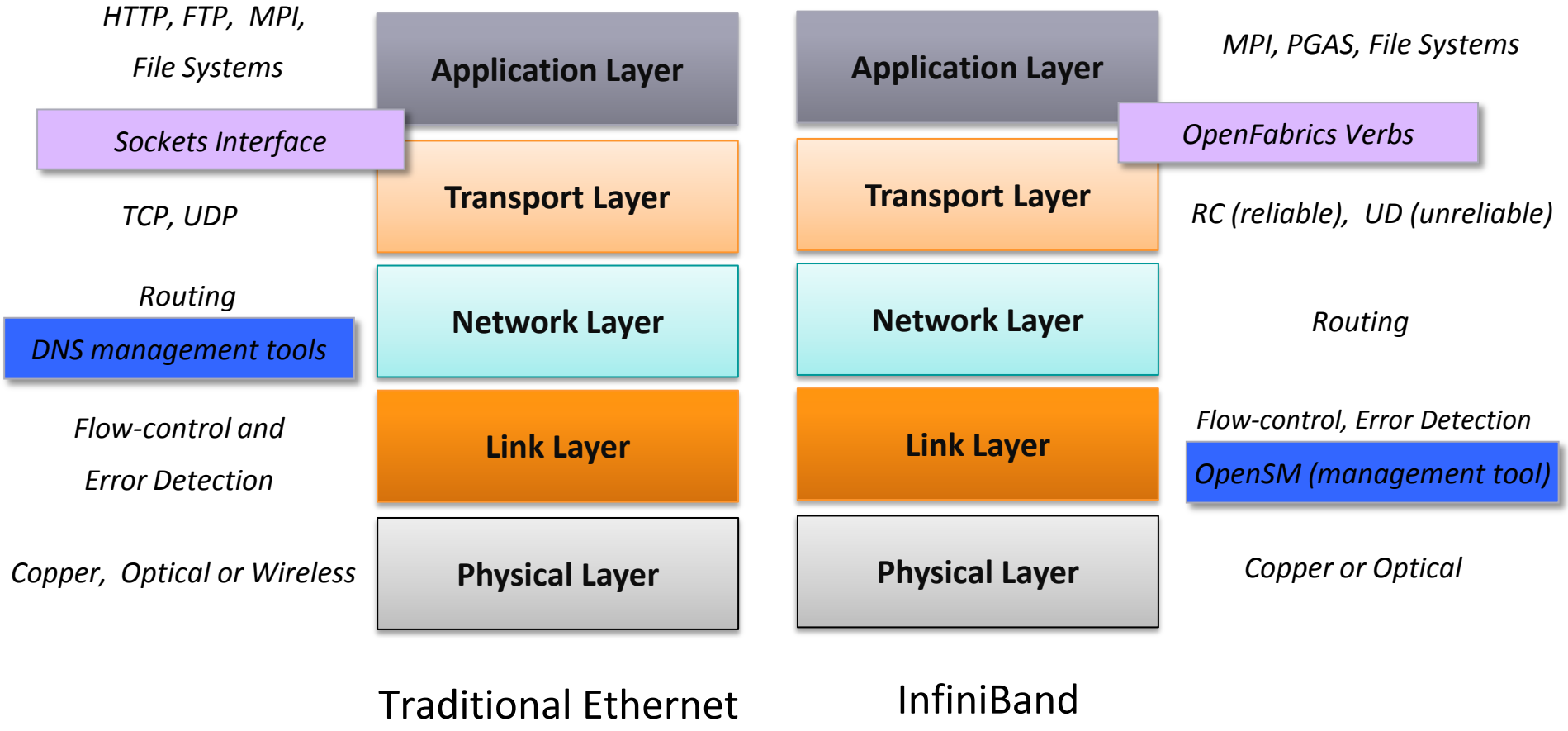
Presentation Overview

- 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 Ethernet (RoE)
  - (InfiniBand) RDMA over Converged (Enhanced) Ethernet (RoCE)

Comparing InfiniBand with Traditional Networking Stack

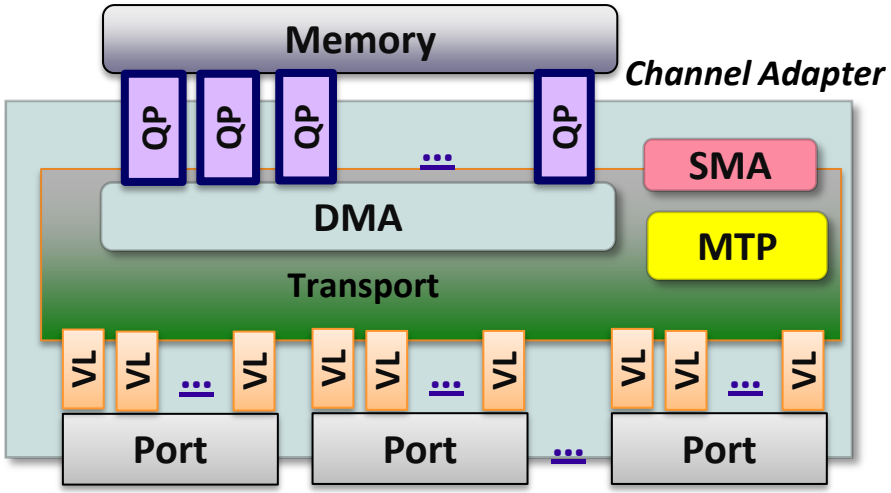


IB Overview

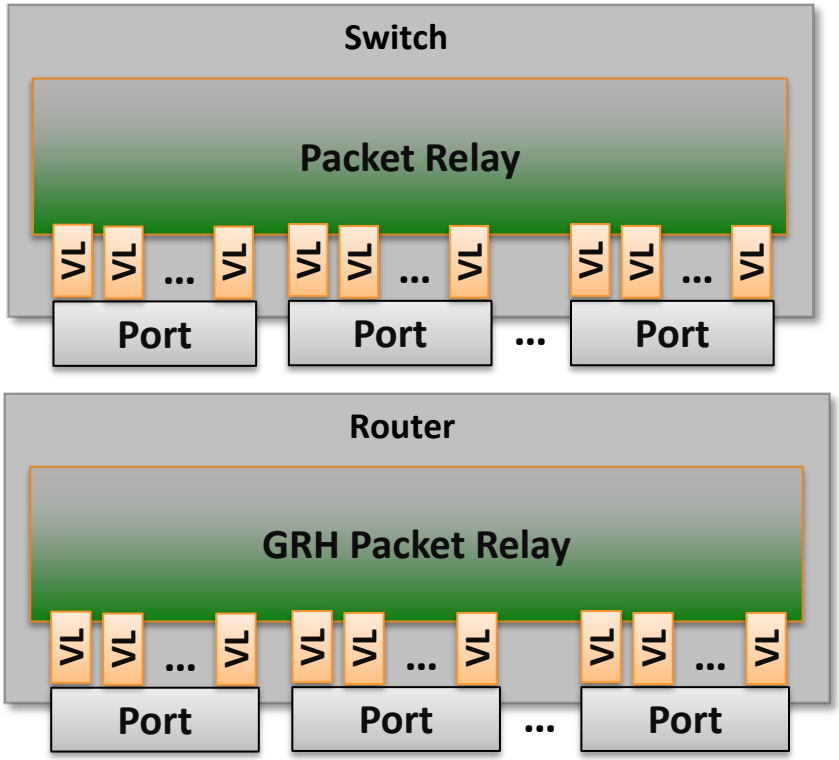
- **InfiniBand**
  - **Architecture and Basic Hardware Components**
  - Communication Model and Semantics
    - Communication Model
    - Memory registration and protection
    - Channel and memory semantics
  - Novel Features
    - Hardware Protocol Offload
      - Link, network and transport layer features
  - Subnet Management and Services

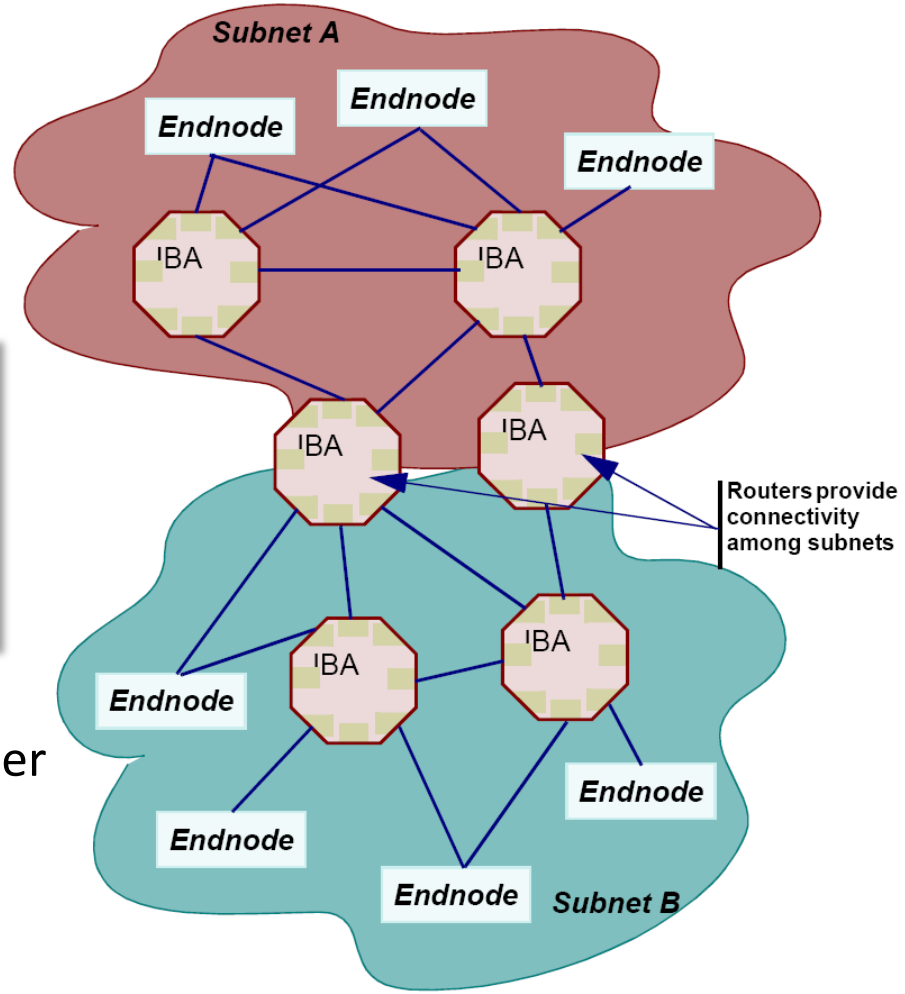
Components: Channel Adapters

- 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)

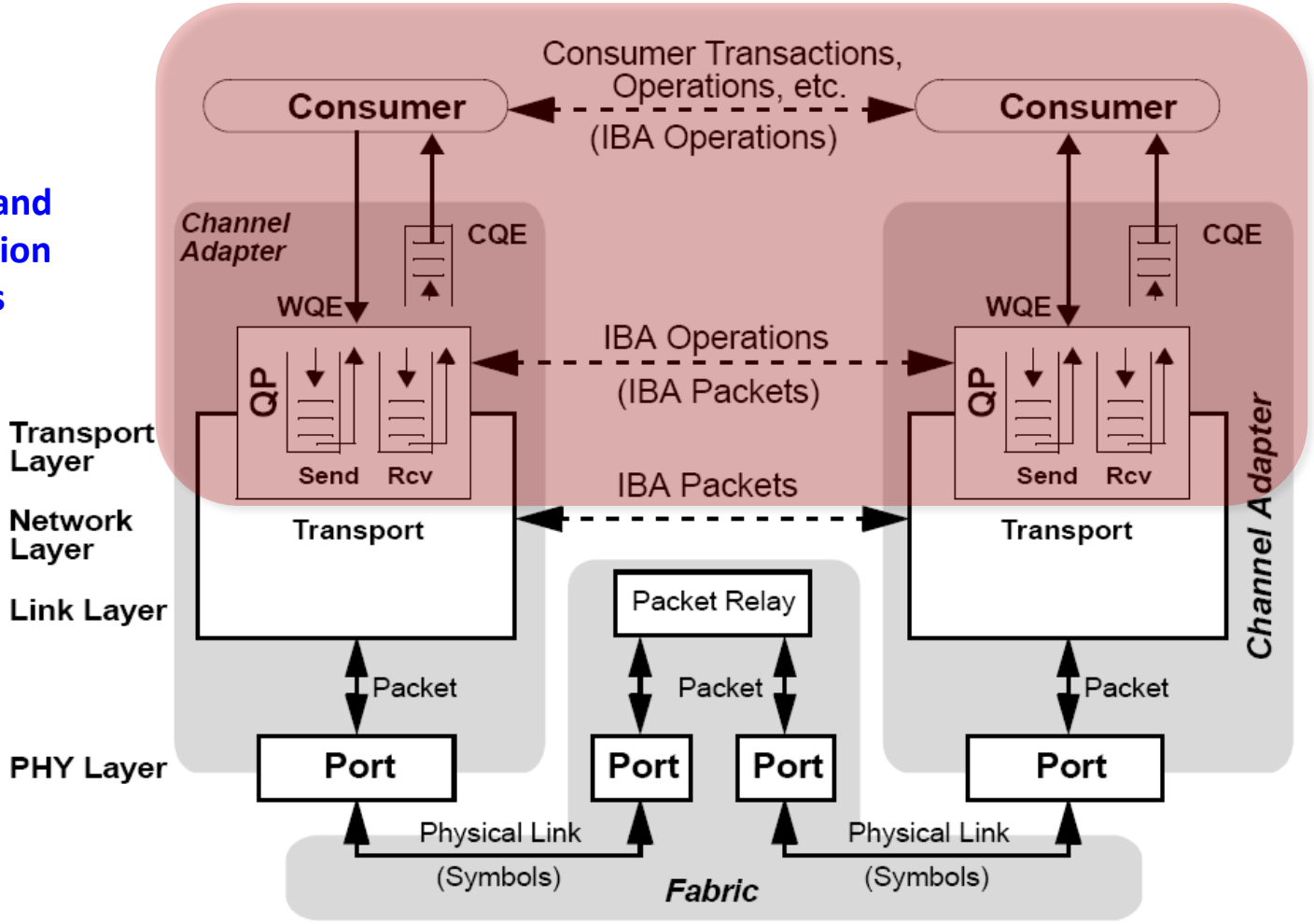
(Courtesy Intel)

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

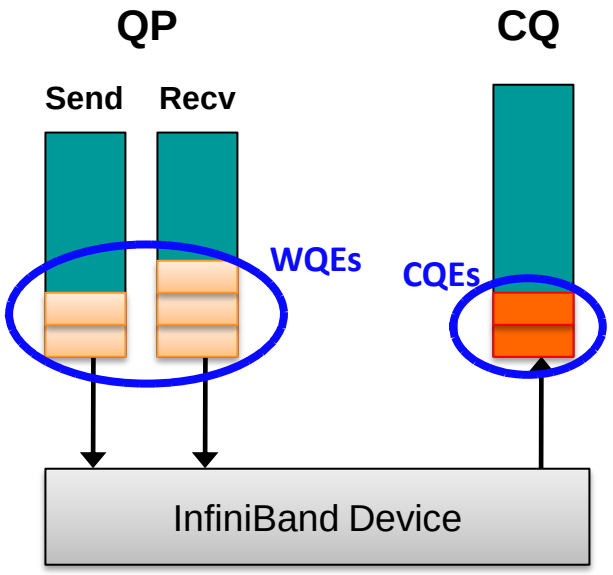
IB Communication Model

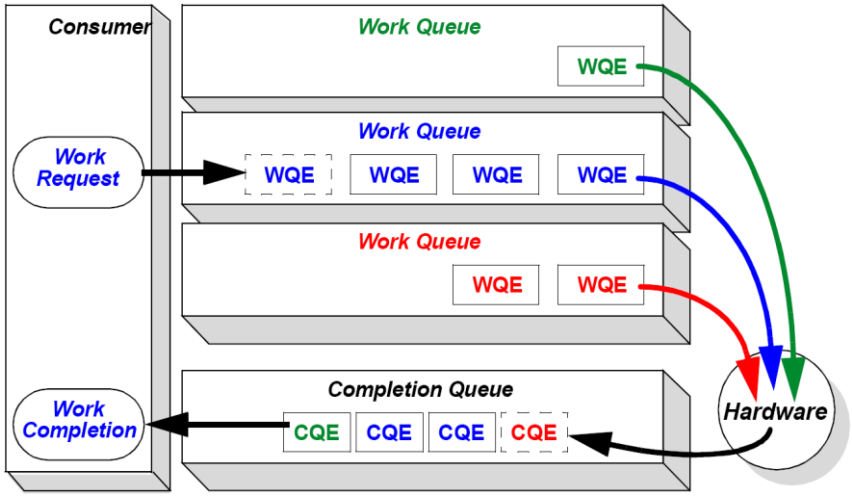
Basic InfiniBand Communication Semantics



Queue Pair Model

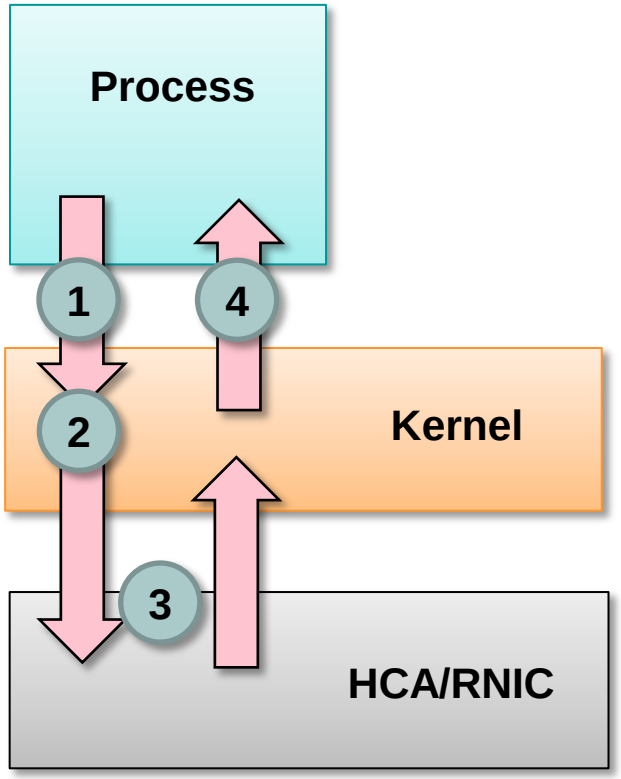
- Each QP has two queues
  - Send Queue (SQ)
  - Receive Queue (RQ)
  - Work requests are queued to the QP (WQEs: “Wookies”)
- QP to be linked to a Complete Queue (CQ)
  - Gives notification of operation completion from QPs
  - Completed WQEs are placed in the CQ with additional information (CQEs: “Cookies”)





Memory Registration

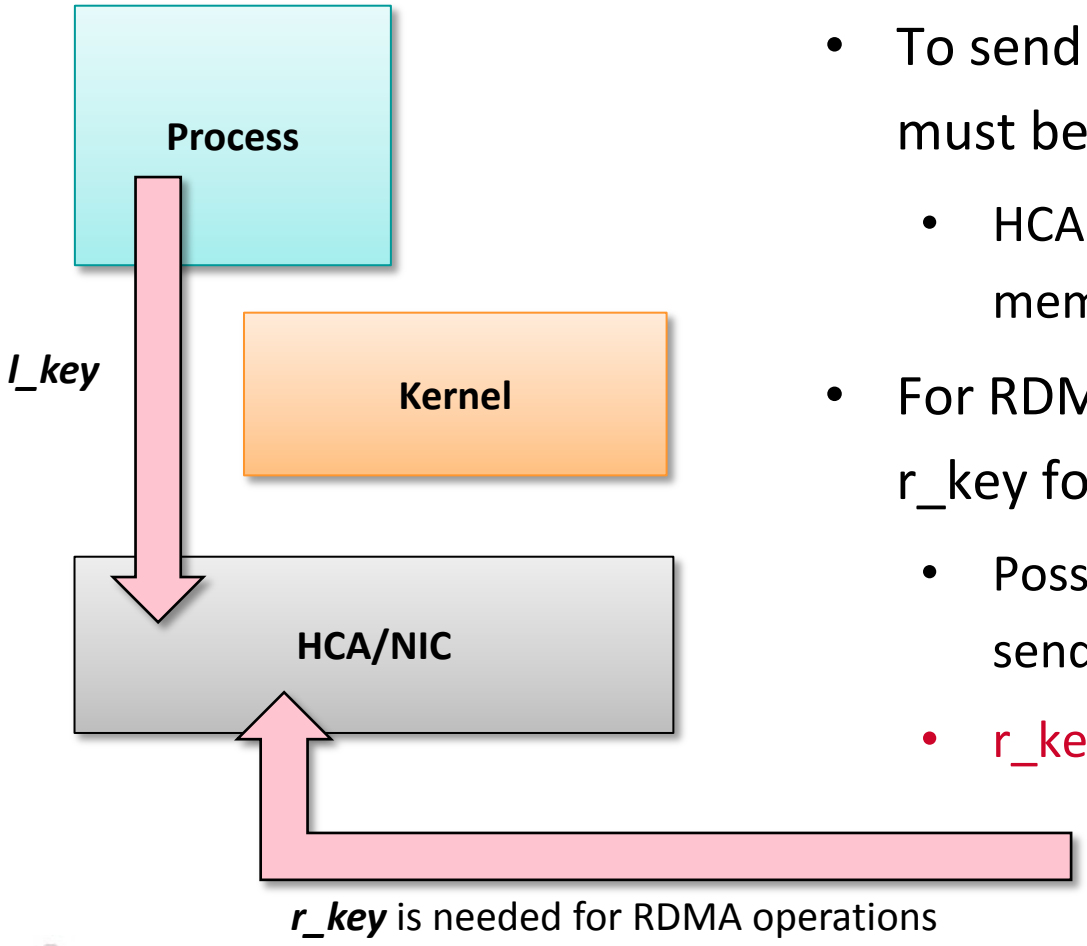
Before we do any communication:
All memory used for communication must be registered



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

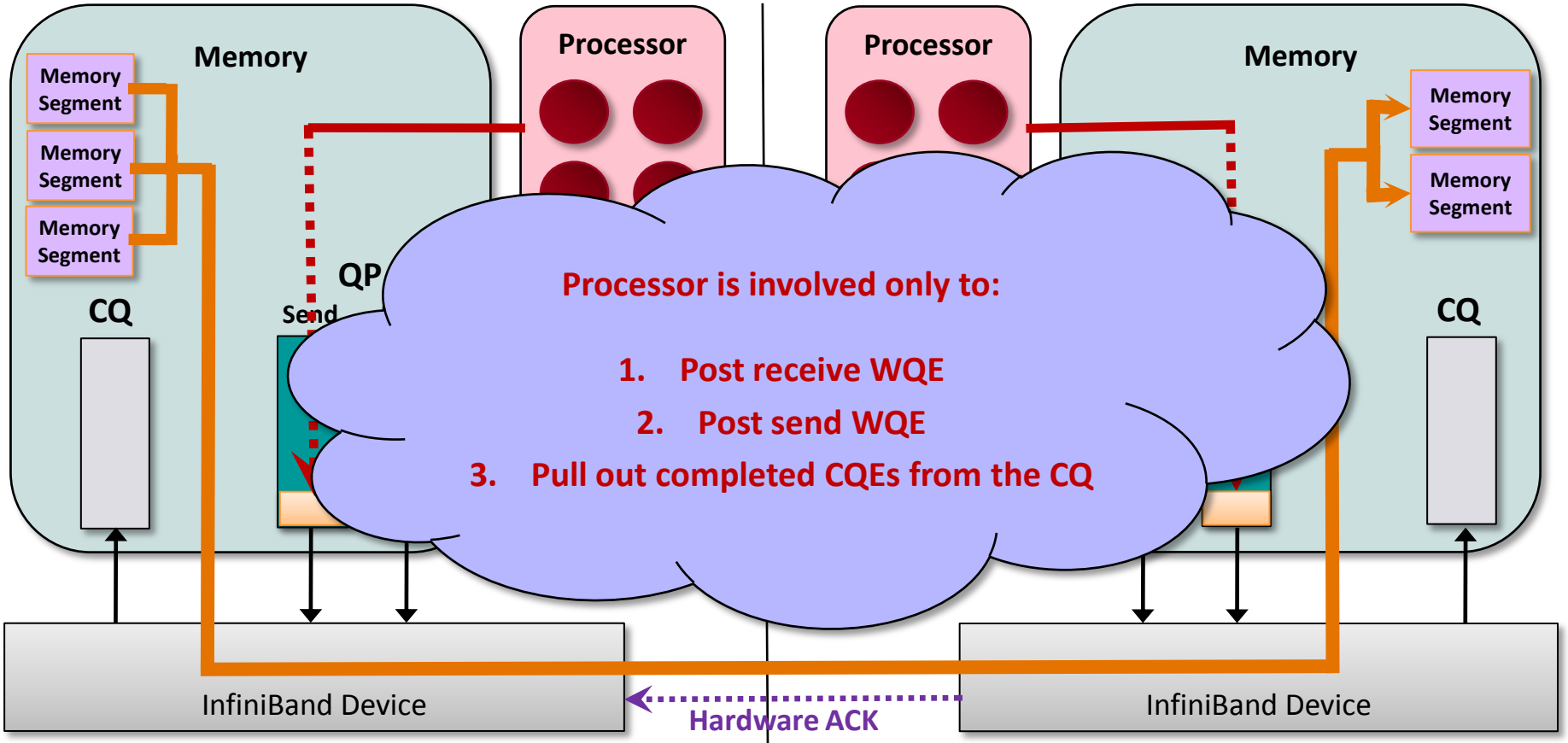
Memory Protection

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

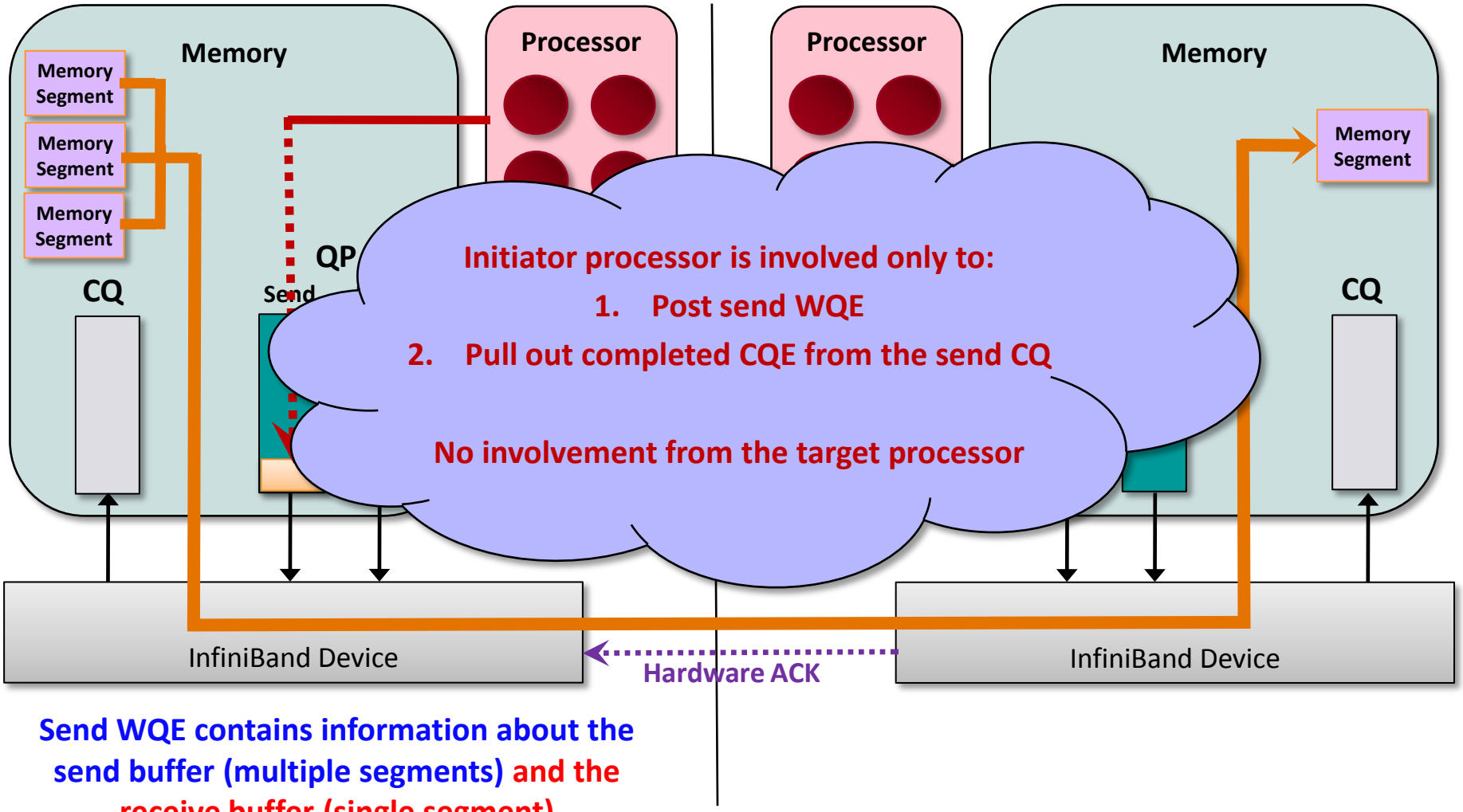
Communication in the Channel Semantics (Send/Receive Model)



Send WQE contains information about the send buffer (multiple non-contiguous segments)

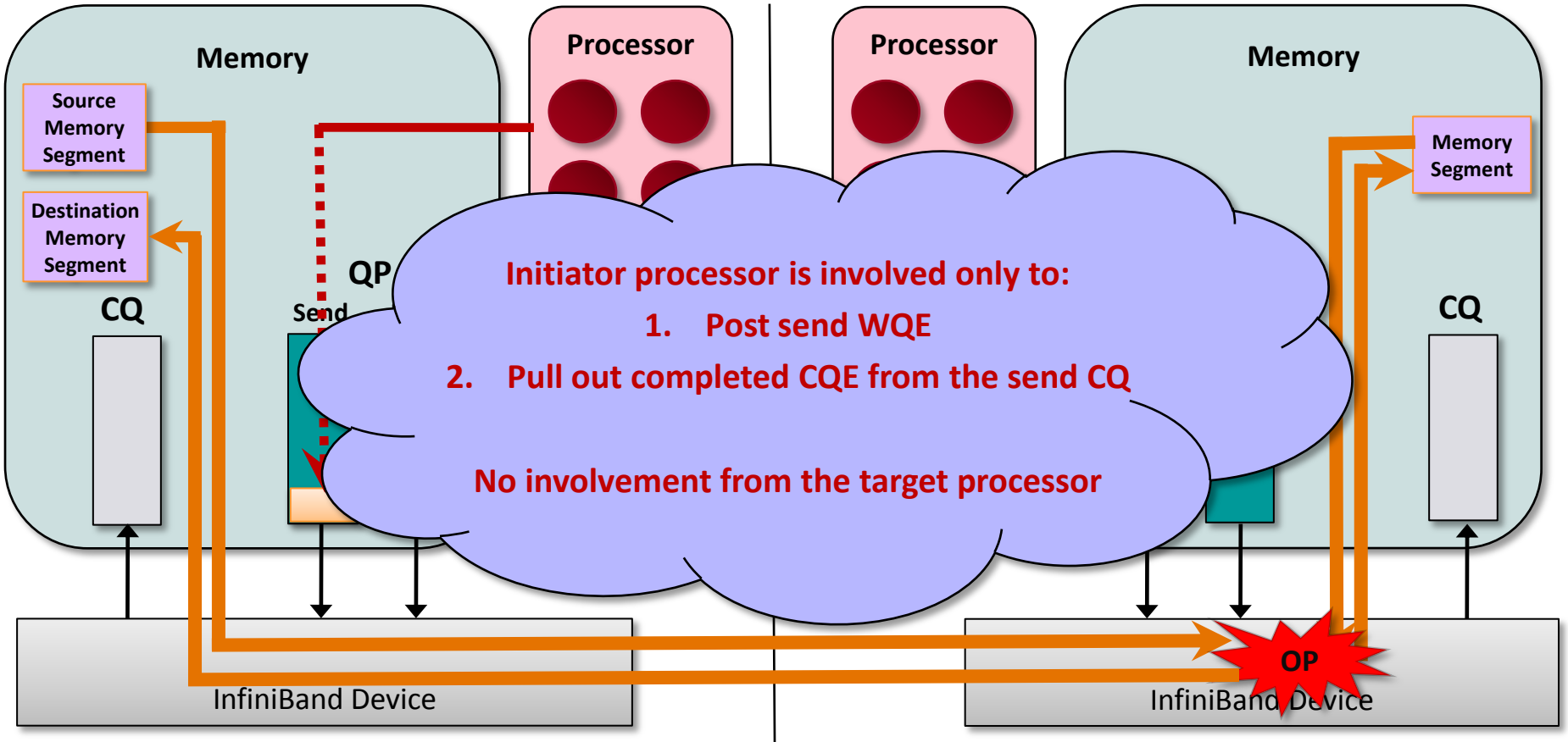
Receive WQE contains information on the receive buffer (multiple non-contiguous segments); Incoming messages have to be matched to a receive WQE to know where to place

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)



Send WQE contains information about the send buffer (single 64-bit segment) and the receive buffer (single 64-bit segment)

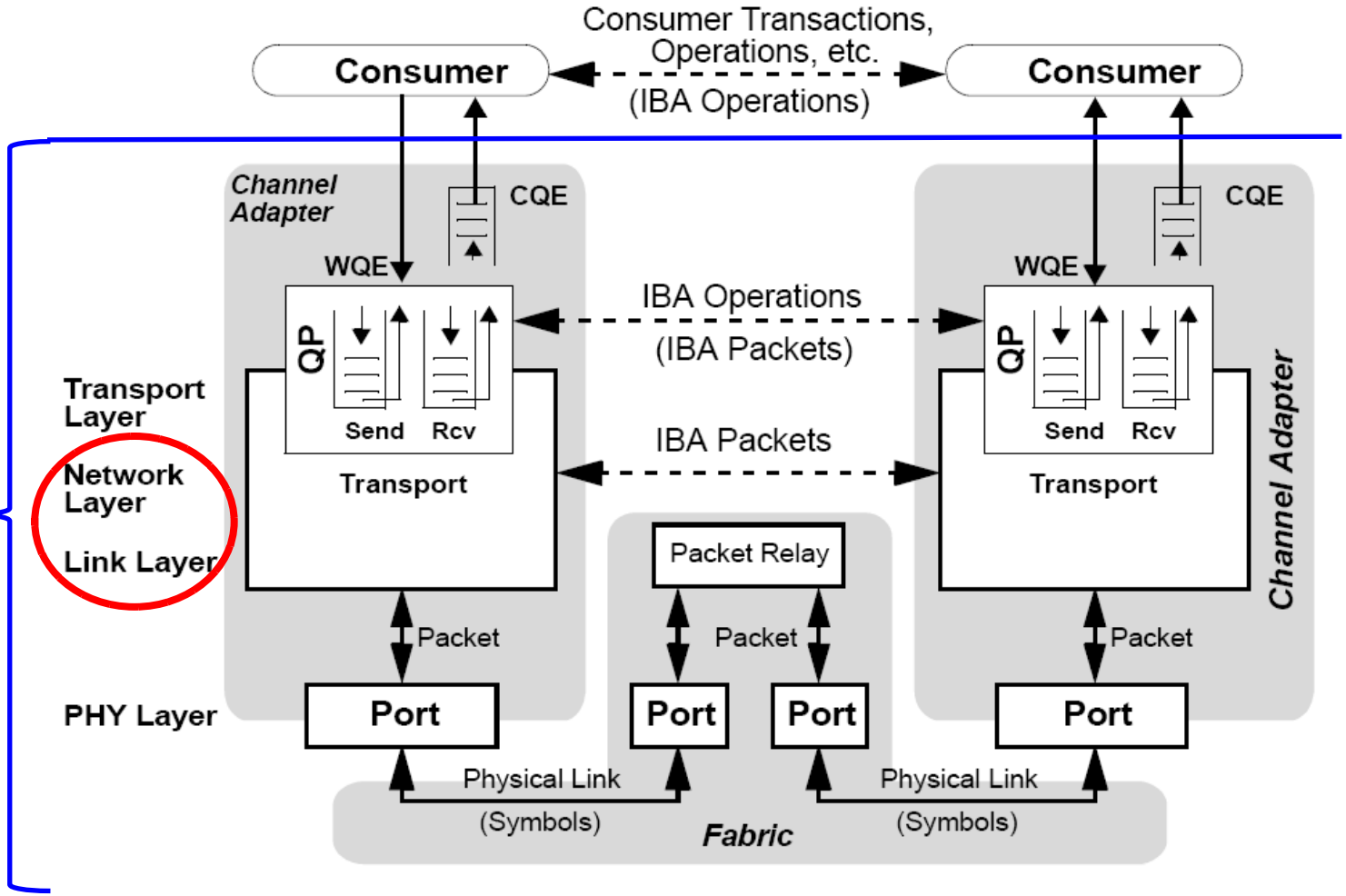
IB supports compare-and-swap and fetch-and-add atomic operations

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

Hardware Protocol Offload

Complete Hardware Implementations Exist



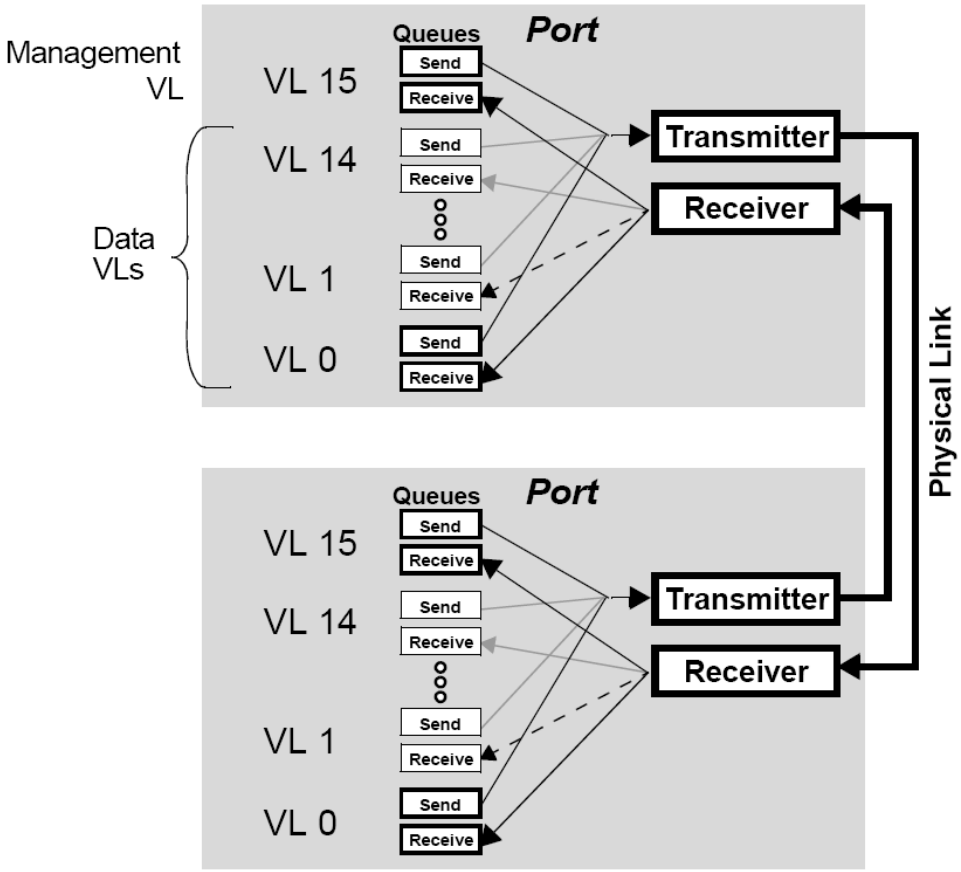
Link/Network Layer Capabilities

- Buffering and Flow Control
- Virtual Lanes, Service Levels and QoS
- Switching and Multicast

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

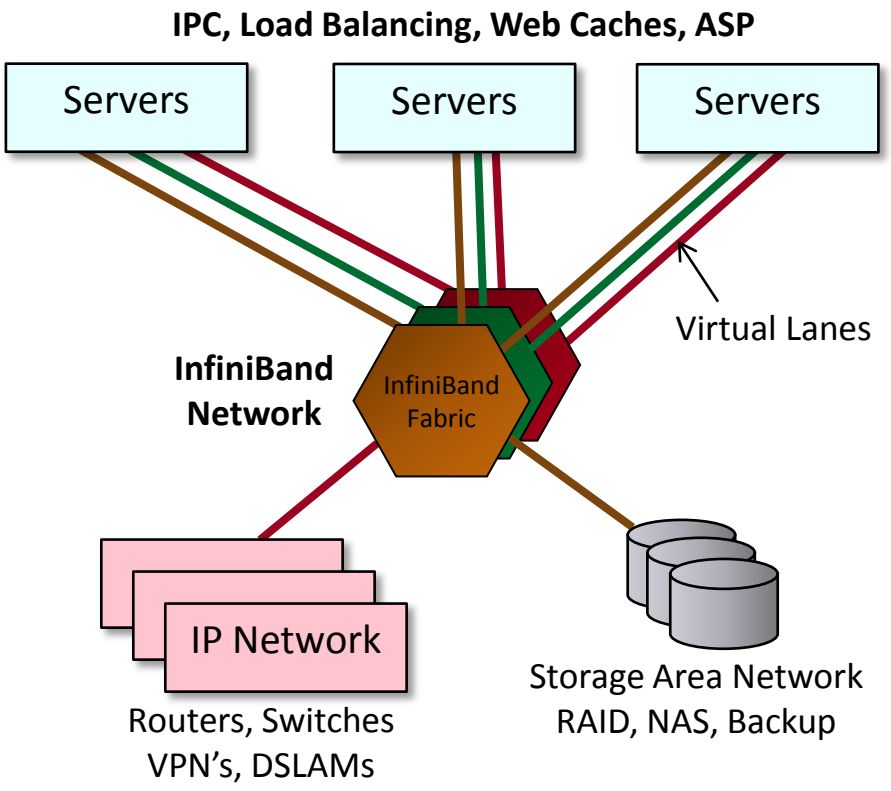


- Multiple virtual links within same physical link
  - Between 2 and 16
- Separate buffers and flow control
  - Avoids Head-of-Line Blocking
- VL15: reserved for management
- Each port supports one or more data VL

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



(Courtesy: Mellanox Technologies)

- 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

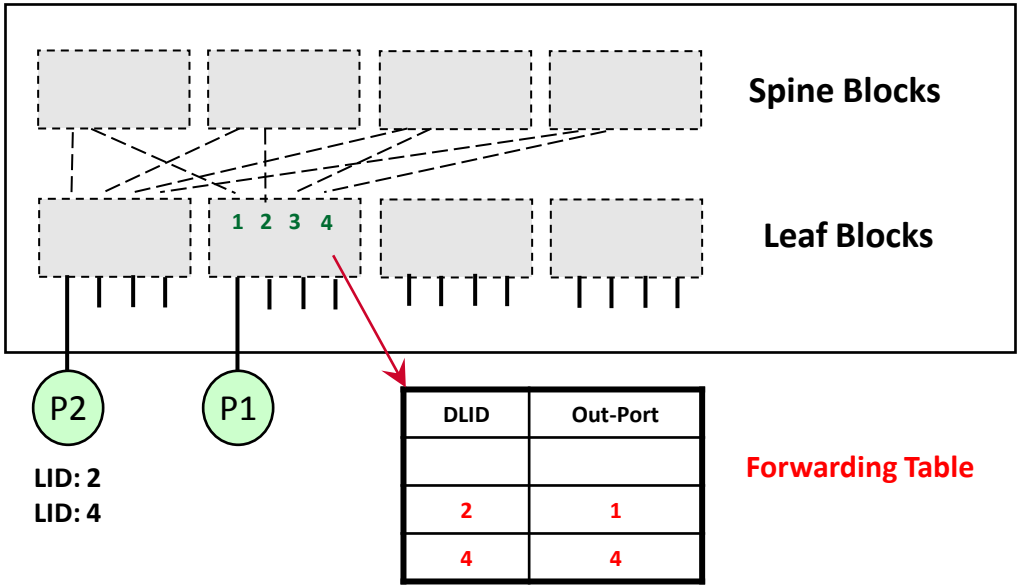
- Each port has one or more associated LIDs (Local Identifiers)
  - Switches look up which port to forward a packet to based on its destination LID (DLID)
  - This information is maintained at the switch
- For multicast packets, the switch needs to maintain multiple output ports to forward the packet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

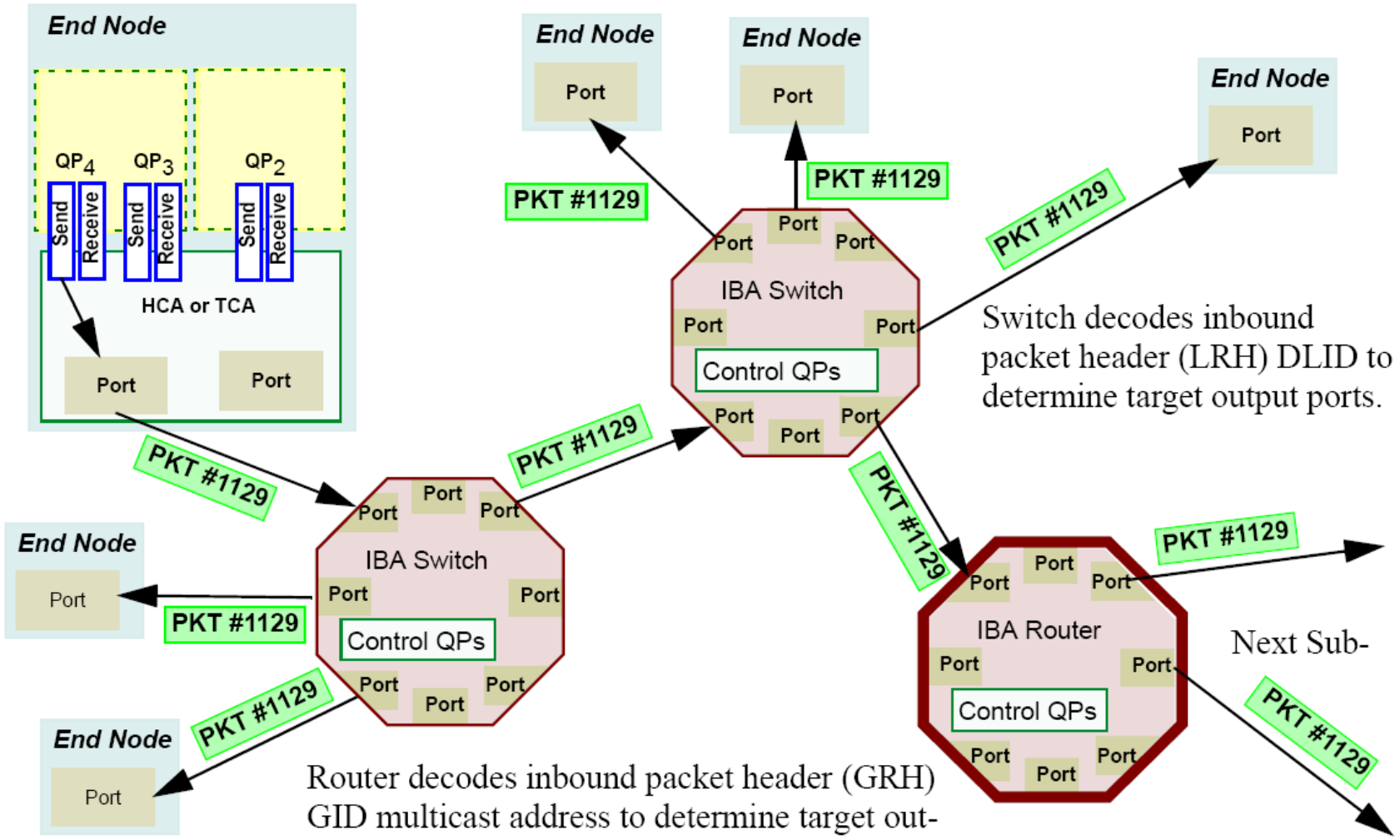
- Someone has to setup the forwarding tables and give every port an LID
  - "Subnet Manager" does this work
- Different routing algorithms give different paths

- Fat-Tree is a popular topology for IB Cluster
  - Different over-subscription ratio may be used
- Other topologies are also being used
  - 3D Torus (Sandia Red Sky) and SGI Altix (Hypercube)

More on Multipathing

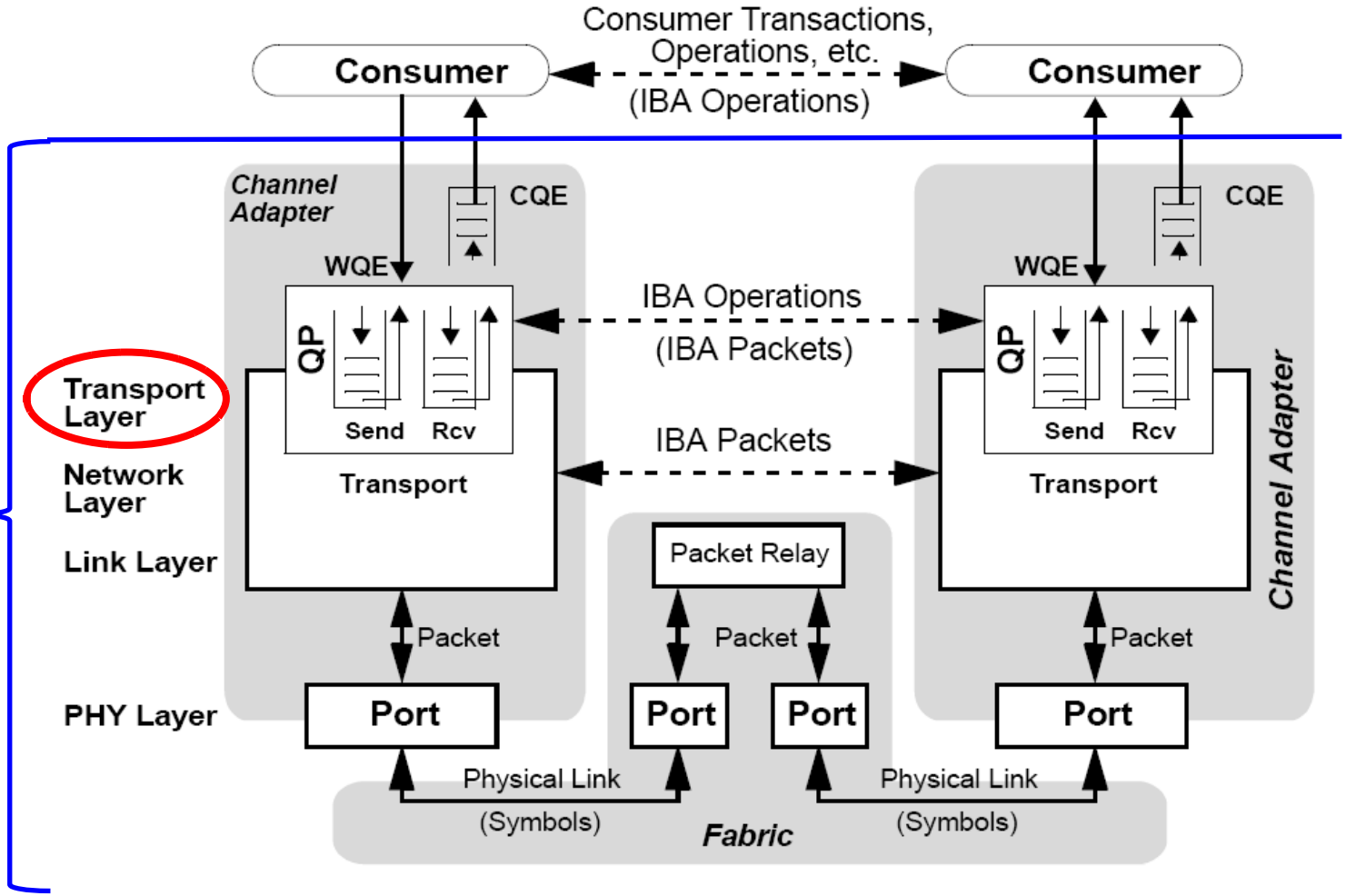
- Similar to basic switching, except…
  - … sender can utilize multiple LIDs associated to the same destination port
    - Packets sent to one DLID take a fixed path
    - Different packets can be sent using different DLIDs
    - Each DLID can have a different path (switch can be configured differently 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



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

- 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**	**No**
	Data Loss Detected	**Yes**				**No**	**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**	**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

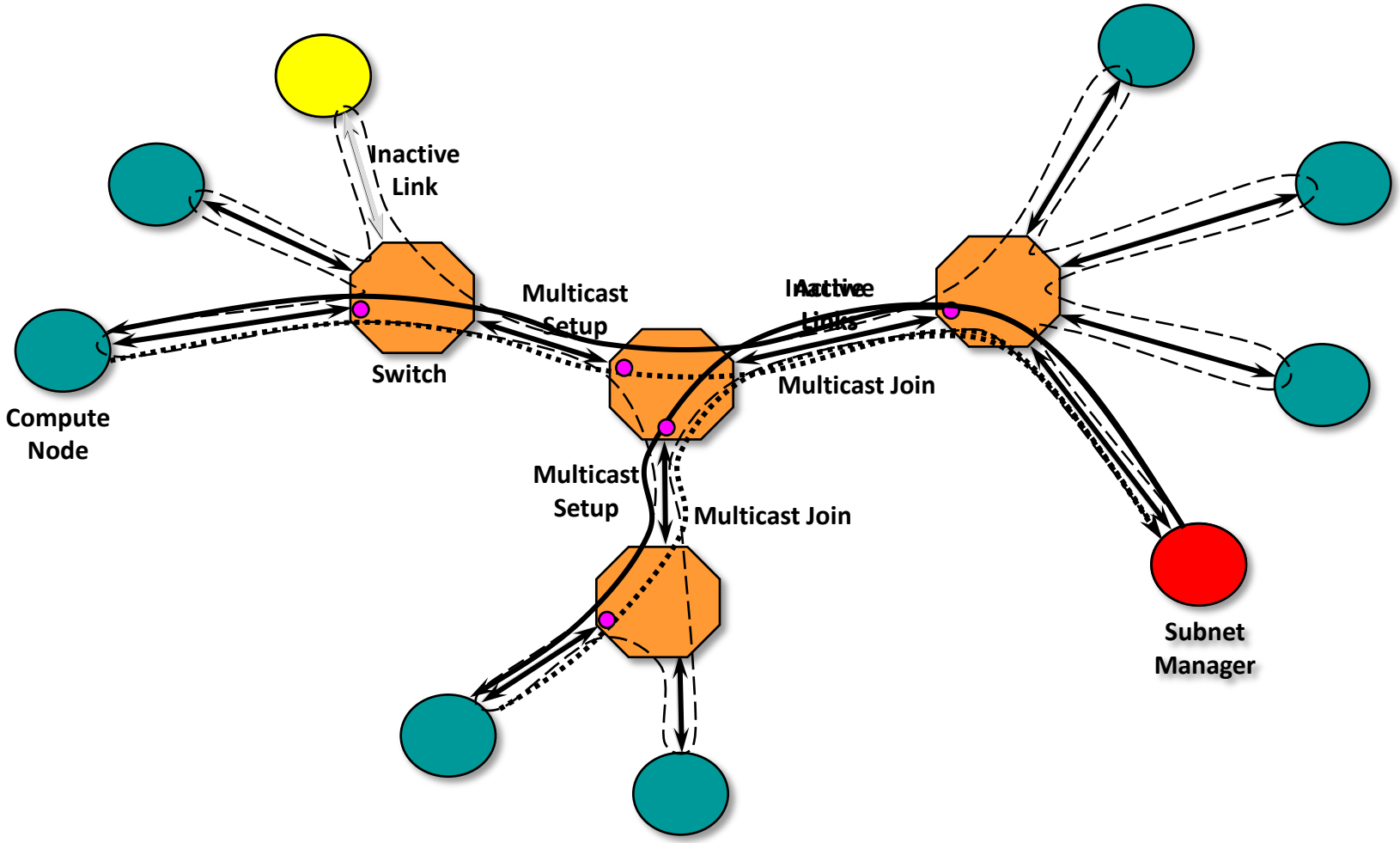
IB Overview

- **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**

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, 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 Ethernet (RoE)
  - (InfiniBand) RDMA over Converged (Enhanced) Ethernet (RoCE)

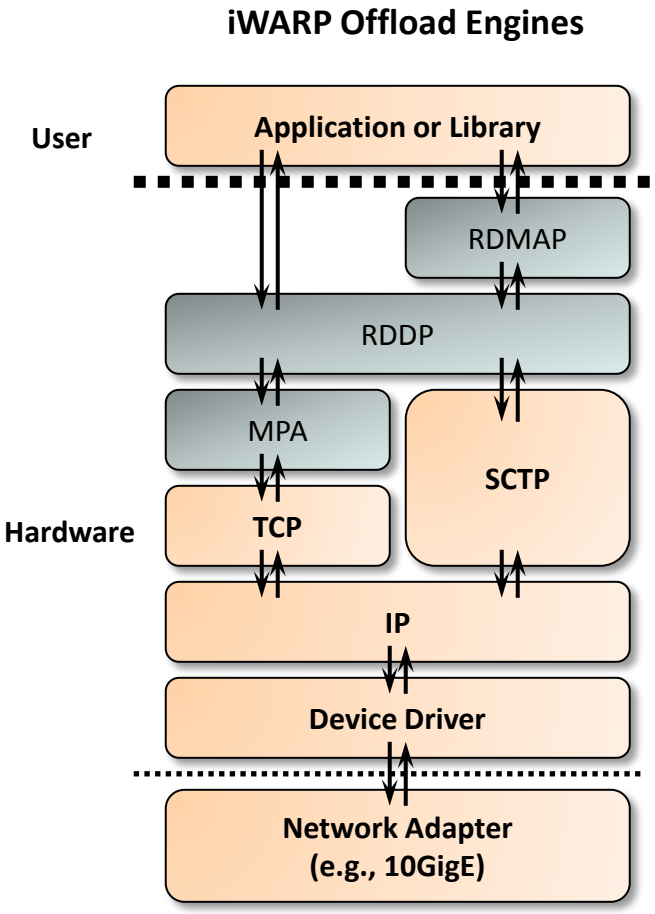
HSE Overview

- **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
      - Multipathing using VLANs
    - 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 HSE RDMA Models: Commonalities and Differences

	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



(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)

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 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

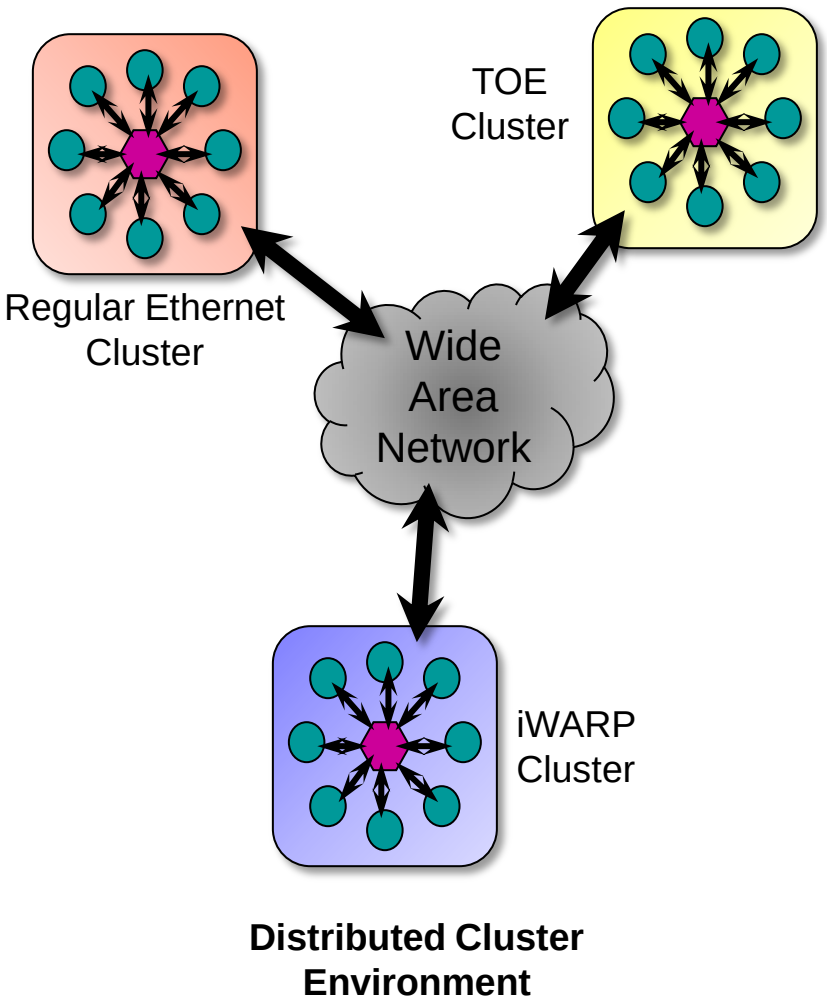
Prioritization and Fixed Bandwidth QoS

- 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)

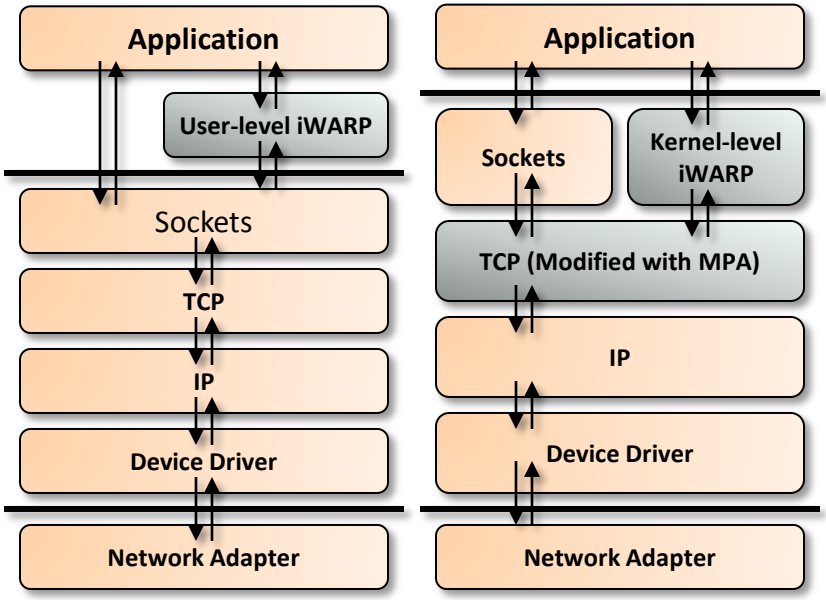
Software iWARP based Compatibility



- 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

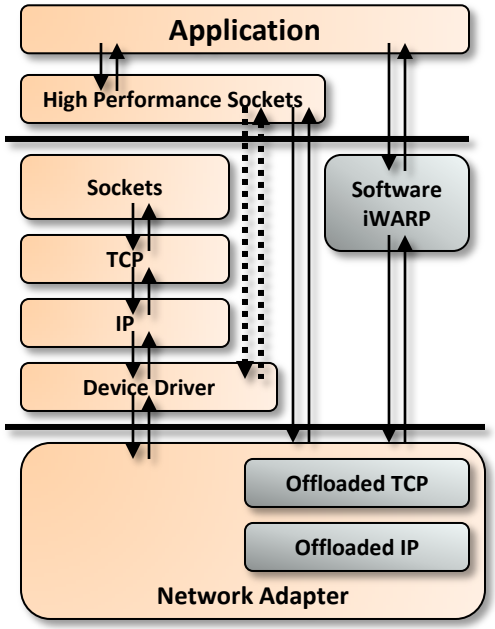
Different iWARP Implementations

OSU, OSC, IBM



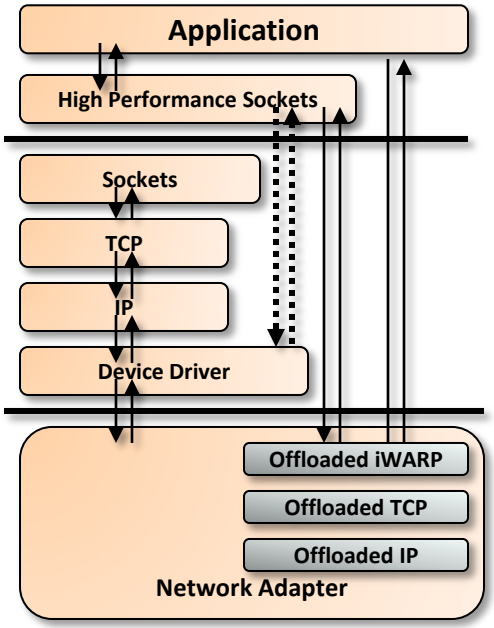
Regular Ethernet Adapters

OSU, ANL



TCP Offload Engines

Chelsio, NetEffect (Intel)



iWARP compliant Adapters

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
      - Multipathing using VLANs
    - 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)**

Myrinet Express (MX)

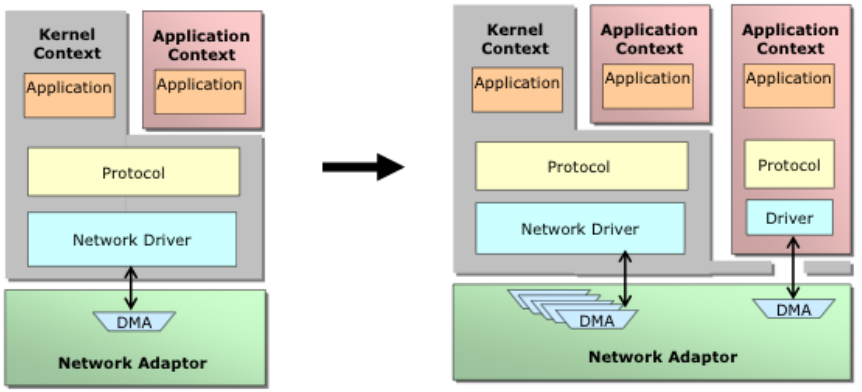
- Proprietary communication layer developed by Myricom for their Myrinet adapters
  - Third generation communication layer (after FM and GM)
  - Supports Myrinet-2000 and the newer Myri-10G adapters
- Low-level "MPI-like" messaging layer
  - Almost one-to-one match with MPI semantics (including connection-less 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*

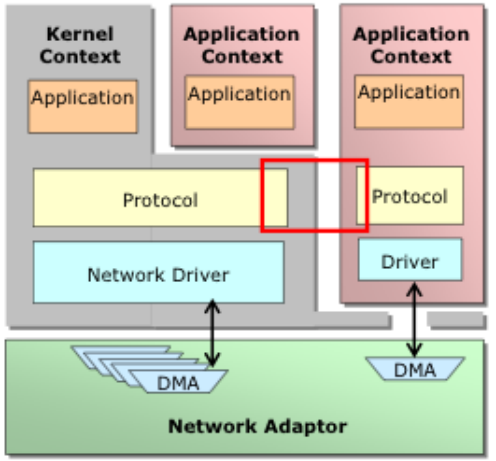
Solarflare Communications: OpenOnload Stack

- 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)



Typical Commodity Networking Stack *Typical HPC Networking Stack*

- 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



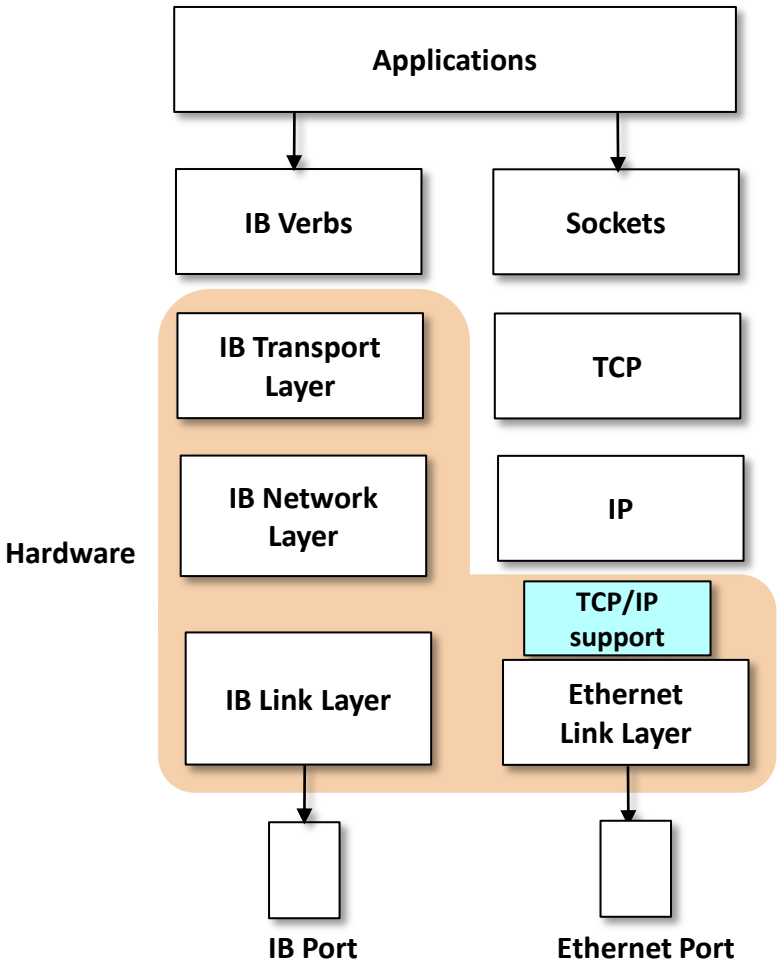
Solarflare approach to networking stack

Courtesy Solarflare communications (www.openonload.org/openonload-google-talk.pdf)

IB, HSE and their Convergence

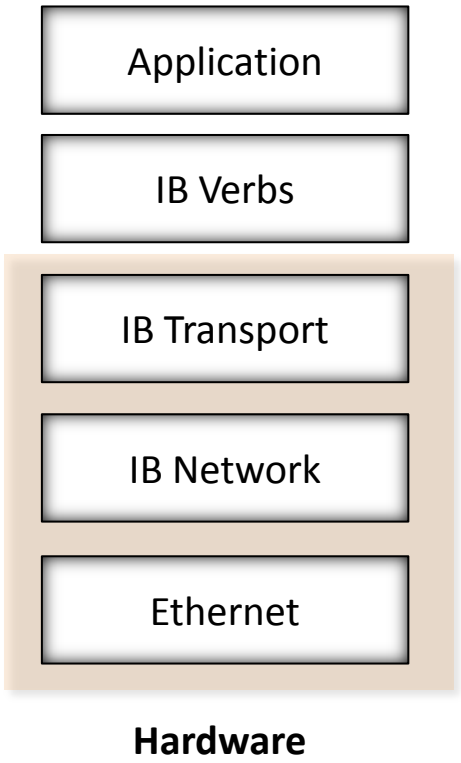
- InfiniBand
  - Architecture and Basic Hardware Components
  - Communication Model and Semantics
  - Novel Features
  - Subnet Management and Services
- High-speed Ethernet Family
  - Internet Wide Area RDMA Protocol (iWARP)
  - Alternate vendor-specific protocol stacks
- **InfiniBand/Ethernet Convergence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

Virtual Protocol Interconnect (VPI)



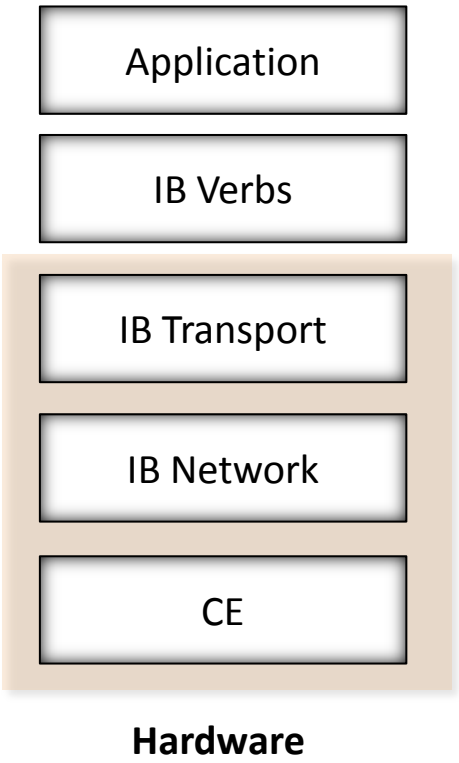
- Single network firmware to support both IB and Ethernet
- Autosensing of layer-2 protocol
  - Can be configured to automatically work with either IB or Ethernet networks
- Multi-port adapters can use one port on IB and another on Ethernet
- Multiple use modes:
  - Datacenters with IB inside the cluster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

(InfiniBand) RDMA over Ethernet (IBoE or RoE)



- Native convergence of IB network and transport layers with Ethernet link layer
- IB packets encapsulated in Ethernet frames
- IB network layer already uses IPv6 frames
- Pros:
  - Works natively in Ethernet environments (entire Ethernet management ecosystem is available)
  - Has all the benefits of IB verbs
- Cons:
  - Network bandwidth might be limited to Ethernet switches: 10GE switches available; 40GE yet to arrive; 32 Gbps IB available
  - Some IB native link-layer features are optional in (regular) Ethernet
- Approved by OFA board to be included into OFED

(InfiniBand) RDMA over Converged (Enhanced) Ethernet (RoCE)



- Very similar to IB over Ethernet
  - Often used interchangeably with IBoE
  - Can be used to explicitly specify 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: 10GE switches available; 40GE yet to arrive; 32 Gbps IB available

IB and HSE: Feature Comparison

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

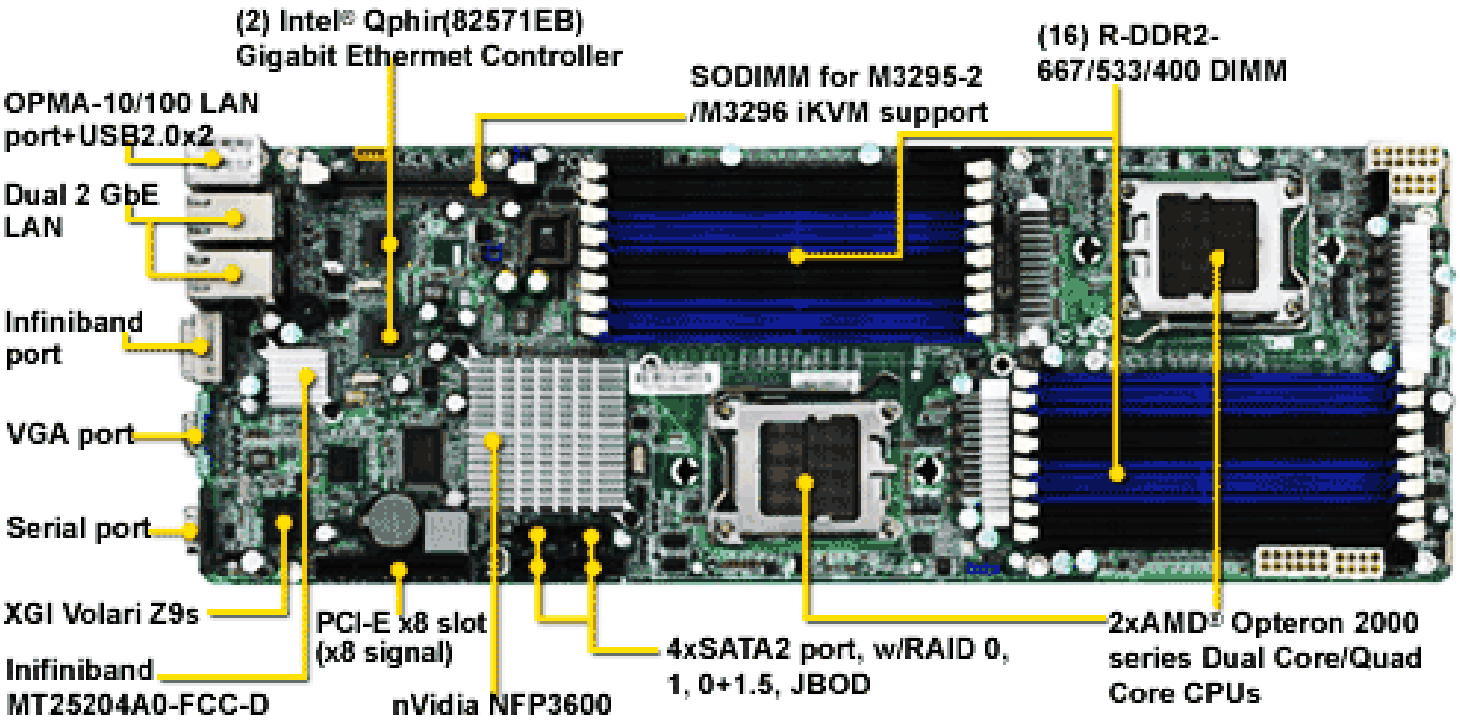
Presentation Overview

- 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
  - 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)
- New ConnectX-2 adapter 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.)

- Customized adapters to work with IB switches
  - Cray XD1 (formerly by Octigabay), Cray CX1
- 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
  - New FDR (56 Gbps) switch silicon (Bridge-X) has been announced by Mellanox in May '11
- Switch Routers with Gateways
  - IB-to-FC; IB-to-IP

10G, 40G and 100G Ethernet Products

- 10GE adapters: Intel, Myricom, Mellanox (ConnectX)
- 10GE/iWARP adapters: Chelsio, NetEffect (now owned by Intel)
- 40GE adapters: Mellanox ConnectX2-EN 40G
- 10GE switches
  - Fulcrum Microsystems
    - Low latency switch based on 24-port silicon
    - FM4000 switch with IP routing, and TCP/UDP support
  - Fujitsu, Myricom (512 ports), Force10, Cisco, Arista (formerly Arastra)
- 40GE and 100GE switches
  - Nortel Networks
    - 10GE downlinks with 40GE and 100GE uplinks
  - Broadcom has announced 40GE switch in early 2010

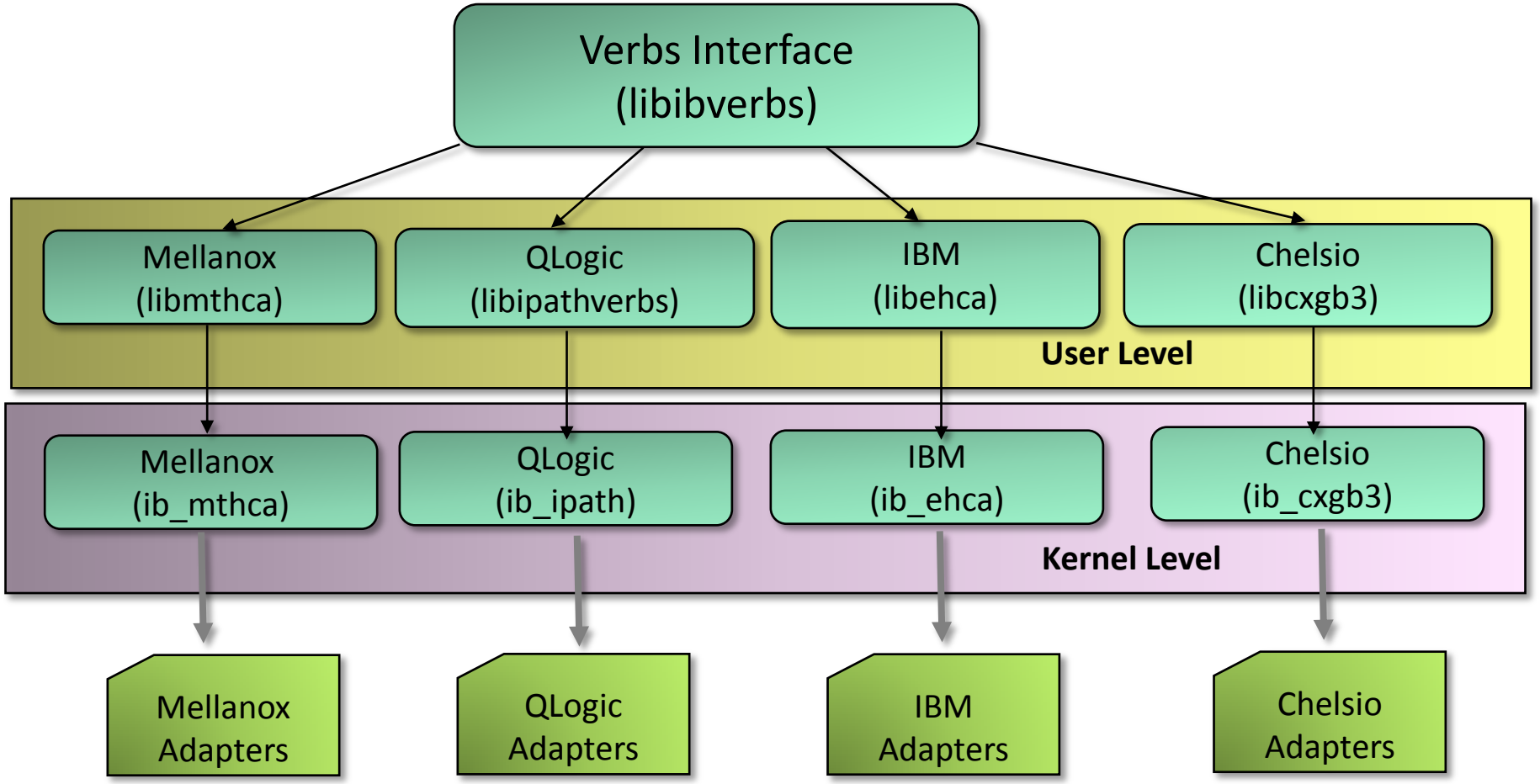
Products Providing IB and HSE Convergence

- 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) and 32Gbps (QDR) rates available for IB
  - 10GE rate available for RoCE
  - 40GE rate for RoCE is expected to be available in near future

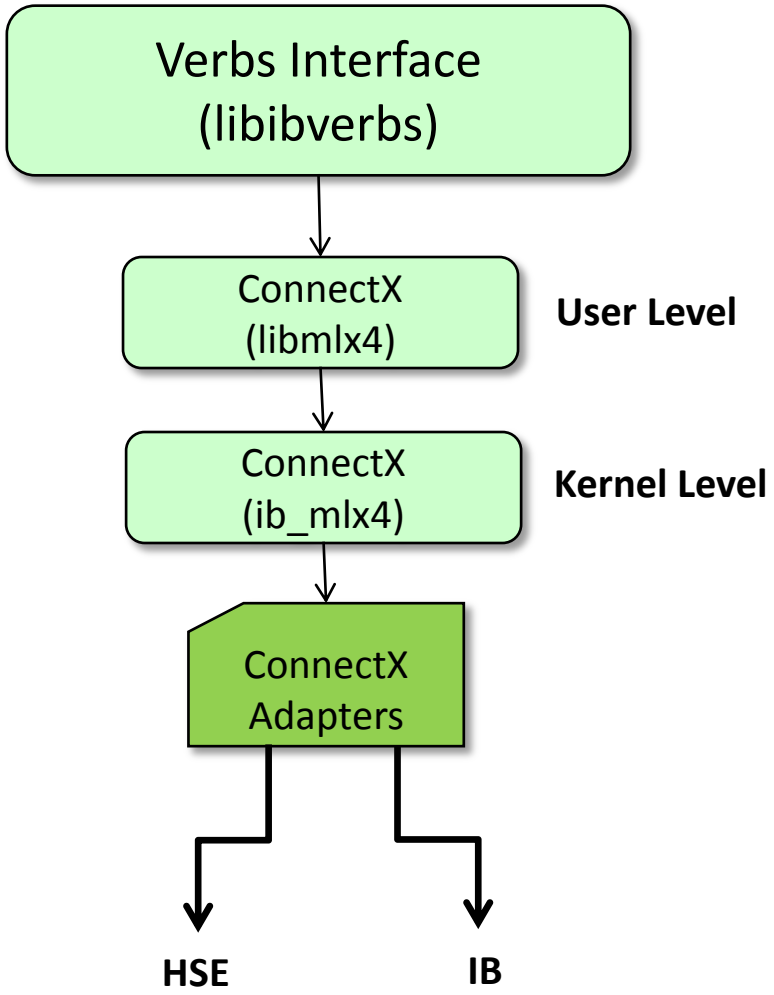
Software Convergence with OpenFabrics

- Open source organization (formerly OpenIB)
  - www.openfabrics.org
- Incorporates both IB and iWARP in a unified manner
  - Support for Linux and Windows
  - Design of complete stack with `best of breed' components
    - Gen1
    - Gen2 (current focus)
- Users can download the entire stack and run
  - Latest release is OFED 1.5.3
  - OFED 1.6 is underway

OpenFabrics Stack with Unified Verbs Interface

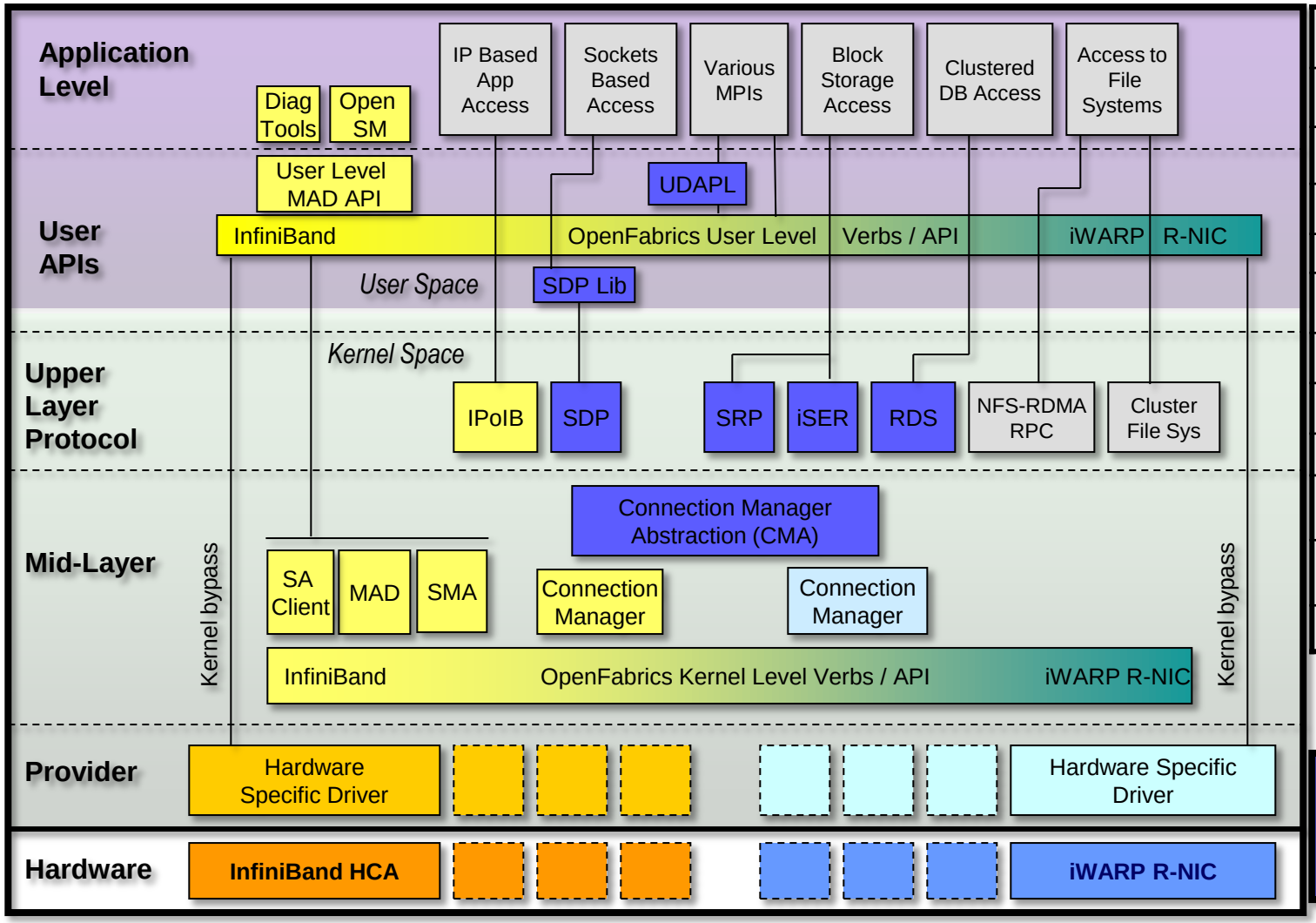


OpenFabrics on Convergent IB/HSE

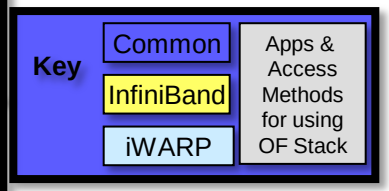


- For IBoE and RoCE, the upper-level stacks remain completely unchanged
- Within the hardware:
  - Transport and network layers remain completely unchanged
  - Both IB and Ethernet (or CEE) link layers are supported on the network adapter
- Note: The OpenFabrics stack is not valid for the Ethernet path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

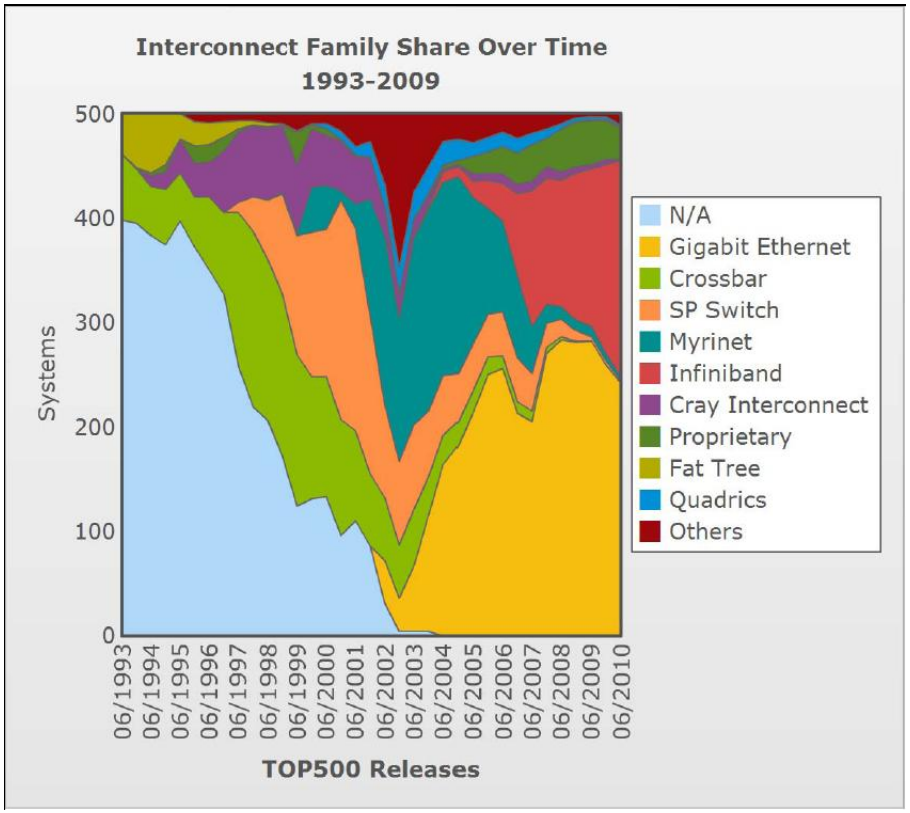
OpenFabrics Software Stack

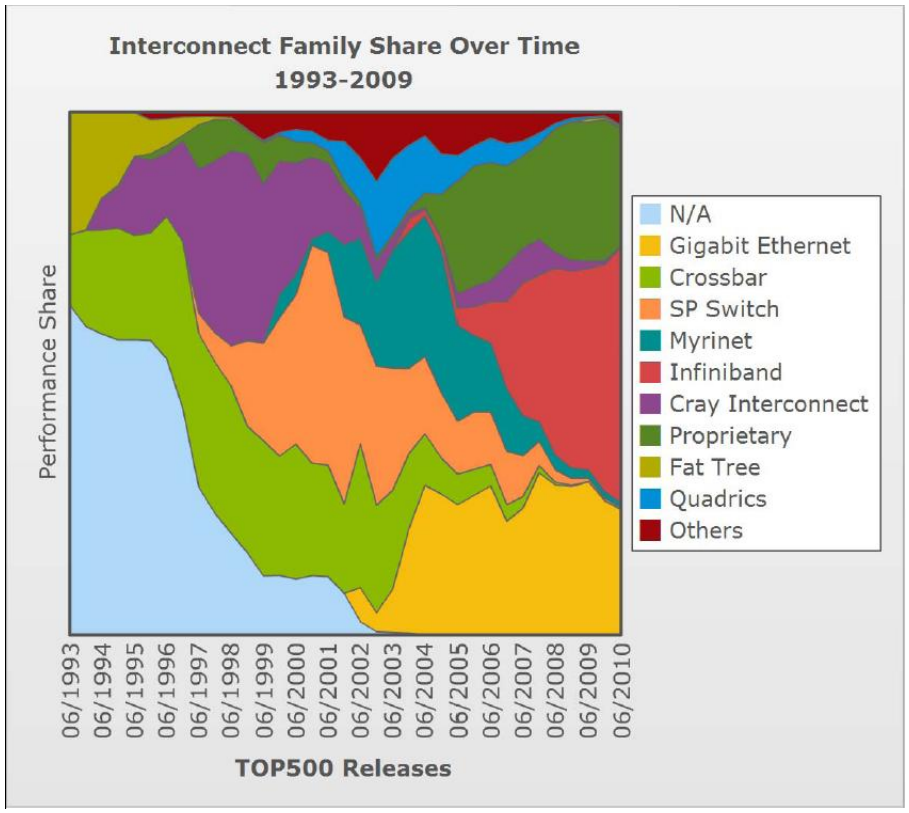


SA	Subnet Administrator
MAD	Management Datagram
SMA	Subnet Manager Agent
PMA	Performance Manager Agent
IPoIB	IP over InfiniBand
SDP	Sockets Direct Protocol
SRP	SCSI RDMA Protocol (Initiator)
iSER	iSCSI RDMA Protocol (Initiator)
RDS	Reliable Datagram Service
UDAPL	User Direct Access Programming Lib
HCA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



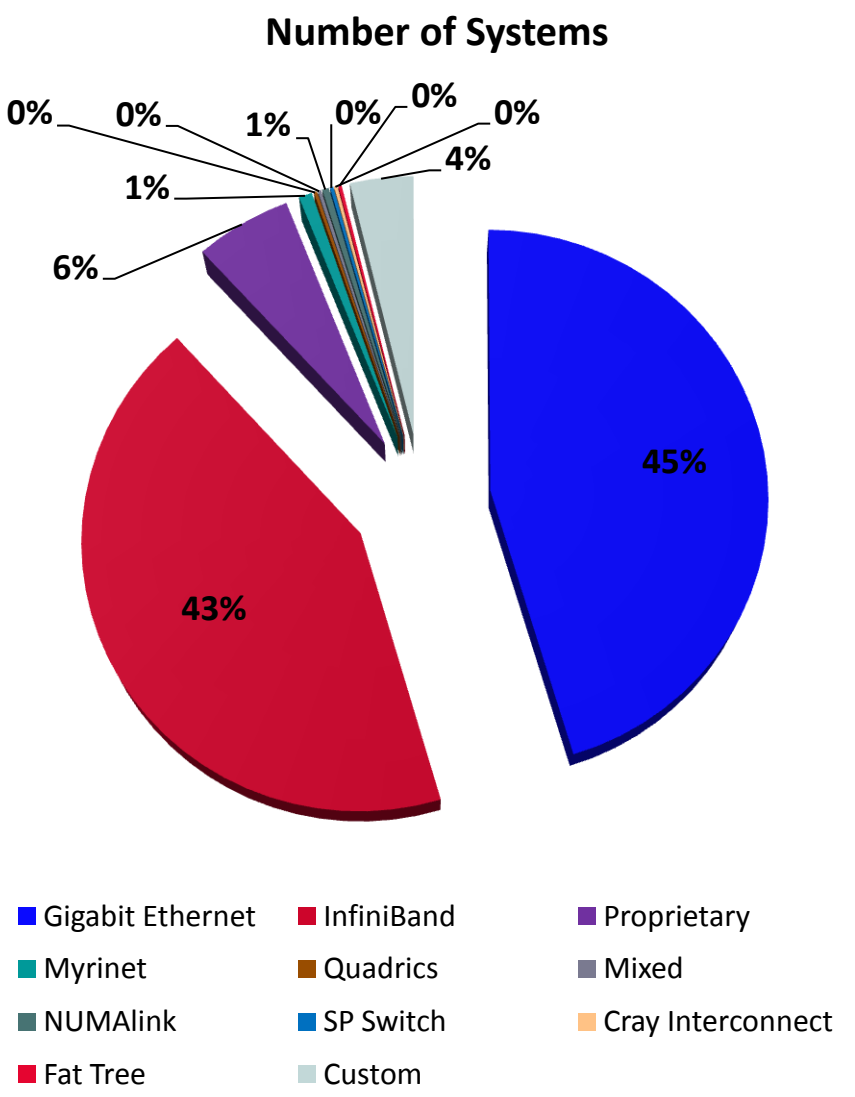
InfiniBand in the Top500

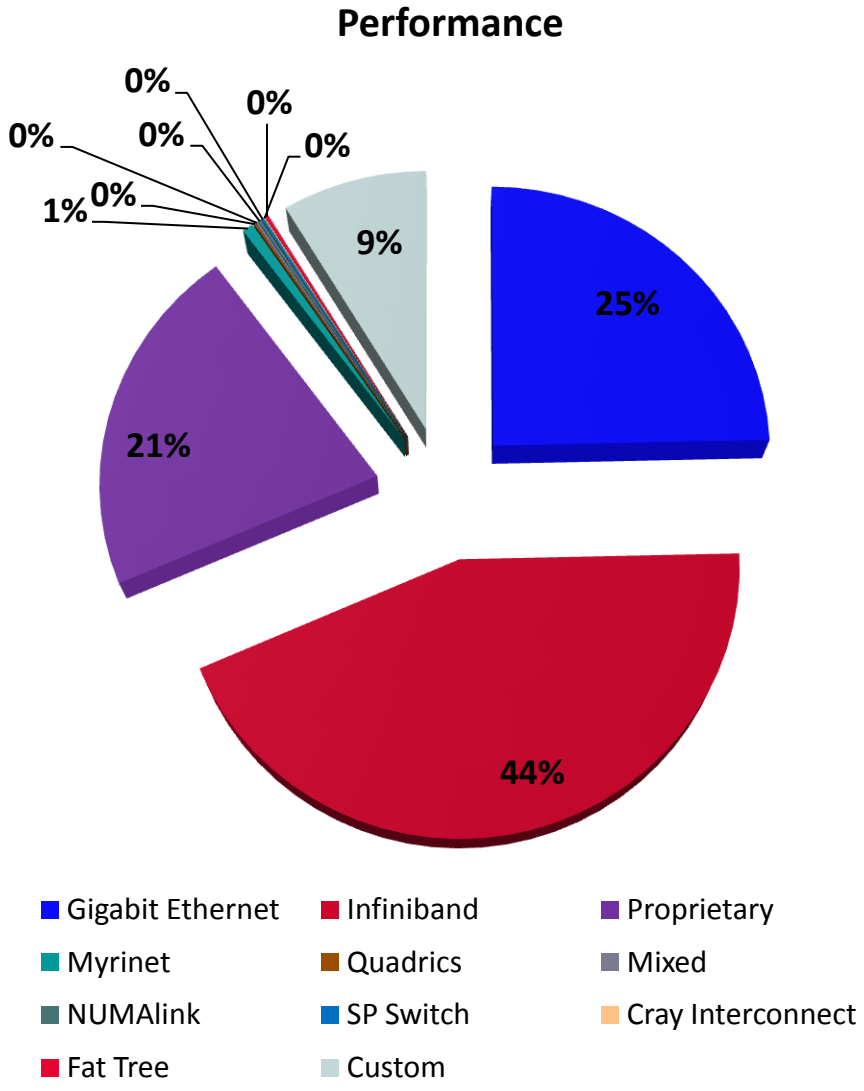




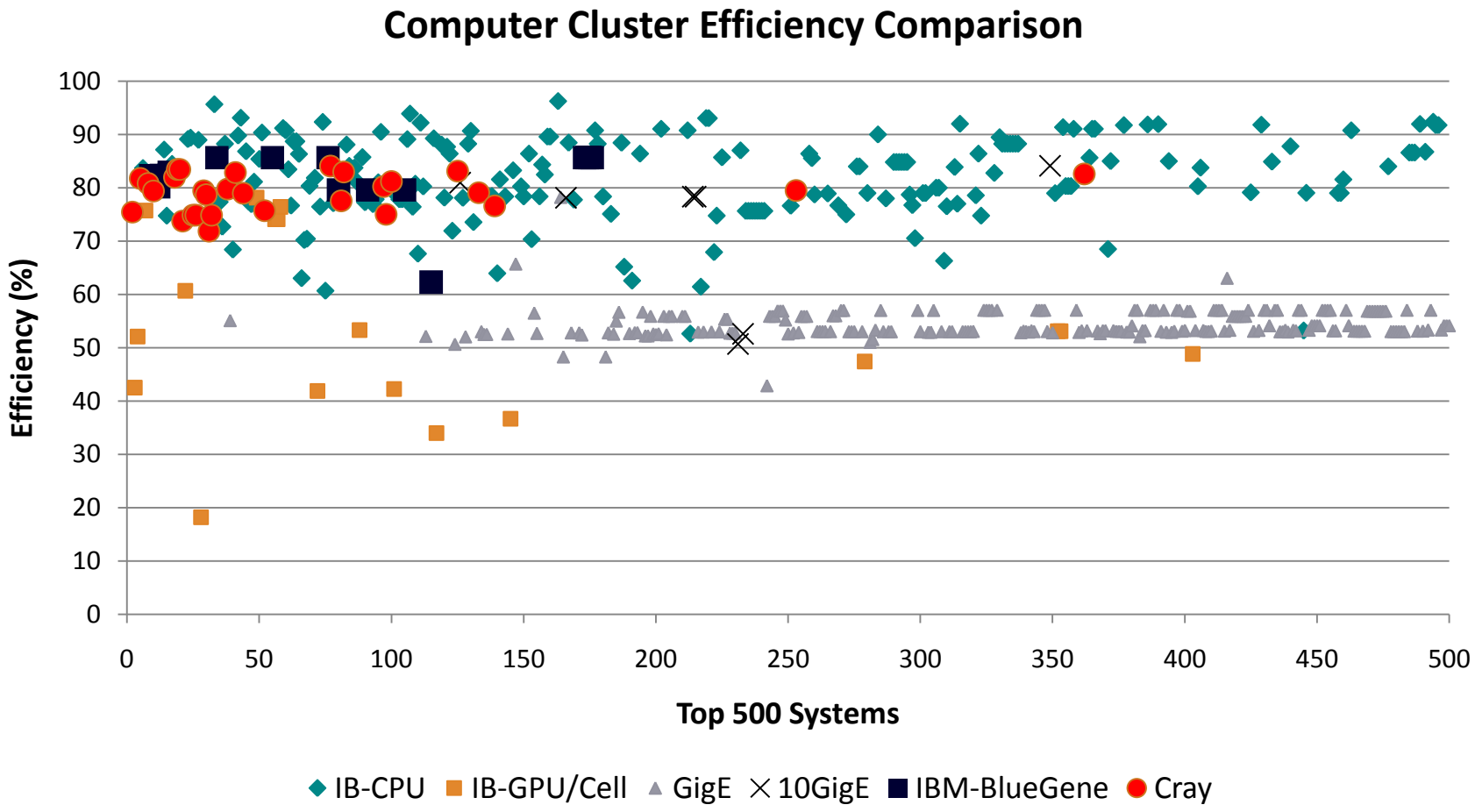
Percentage share of InfiniBand is steadily increasing

InfiniBand in the Top500 (Nov. 2010)





InfiniBand System Efficiency in the Top500 List



Large-scale InfiniBand Installations

- 214 IB Clusters (42.8%) in the Nov '10 Top500 list (http://www.top500.org)
- Installations in the Top 30 (13 systems):

120,640 cores (Nebulae) in China (3rd)	15,120 cores (Loewe) in Germany (22nd)
73,278 cores (Tsubame-2.0) at Japan (4th)	26,304 cores (Juropa) in Germany (23rd)
138,368 cores (Tera-100) at France (6th)	26,232 cores (Tachyonll) in South Korea (24th)
122,400 cores (RoadRunner) at LANL (7th)	23,040 cores (Jade) at GENCI (27th)
81,920 cores (Pleiades) at NASA Ames (11th)	33,120 cores (Mole-8.5) in China (28th)
42,440 cores (Red Sky) at Sandia (14th)	***More are getting installed !***
62,976 cores (Ranger) at TACC (15th)	
35,360 cores (Lomonosov) in Russia (17th)	

HSE Scientific Computing Installations

- HSE compute systems with ranking in the Nov 2010 Top500 list
  - 8,856-core installation in Purdue with ConnectX-EN 10GigE (#126)
  - 7,944-core installation in Purdue with 10GigE Chelsio/iWARP (#147)
  - 6,828-core installation in Germany (#166)
  - 6,144-core installation in Germany (#214)
  - 6,144-core installation in Germany (#215)
  - 7,040-core installation at the Amazon EC2 Cluster (#231)
  - 4,000-core installation at HHMI (#349)
- Other small clusters
  - 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 Argonne National Lab with Myri-10G
- Integrated Systems
  - BG/P uses 10GE for I/O (ranks 9, 13, 16, and 34 in the Top 50)

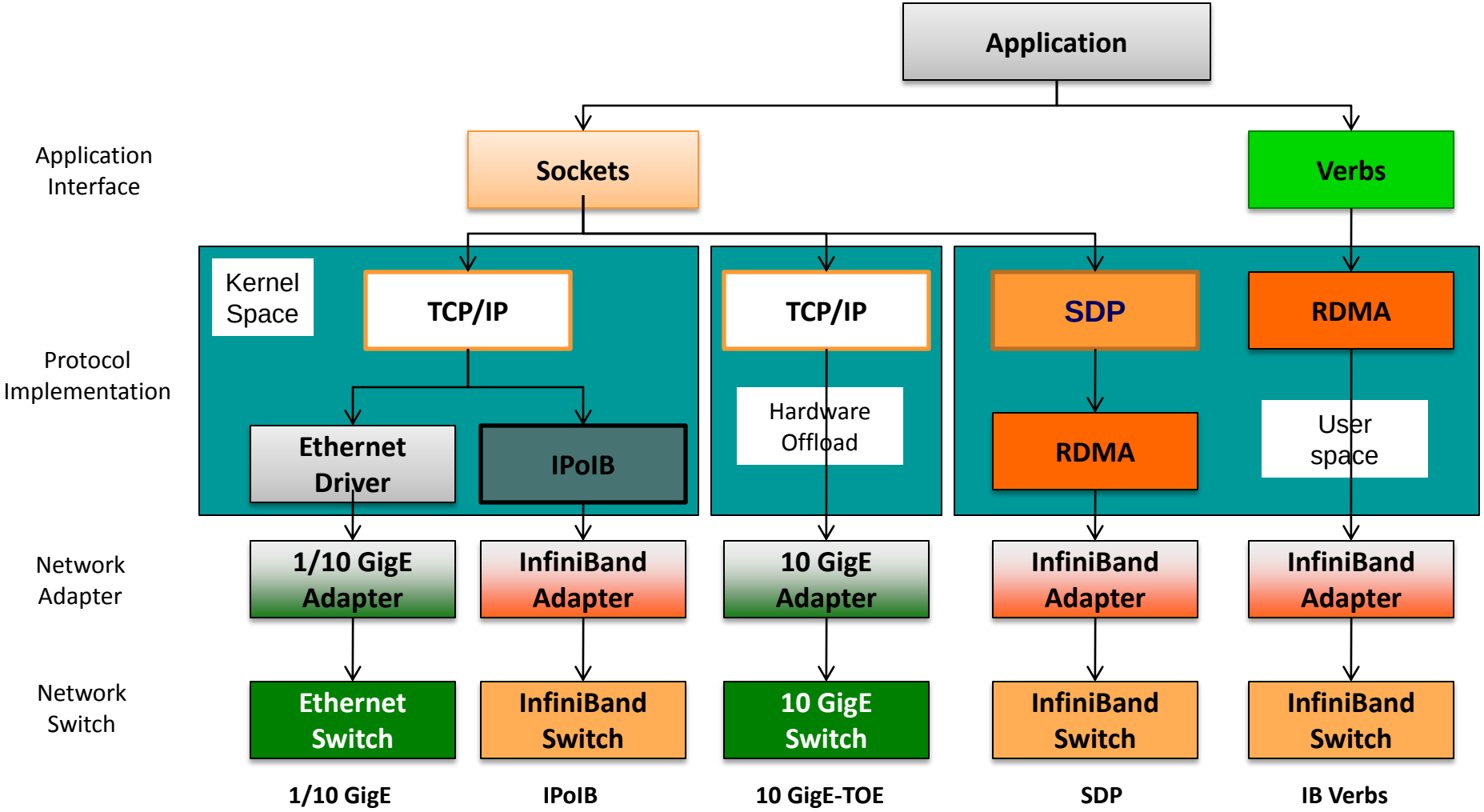
Other HSE Installations

- 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"), 20th 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 to install $62M 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

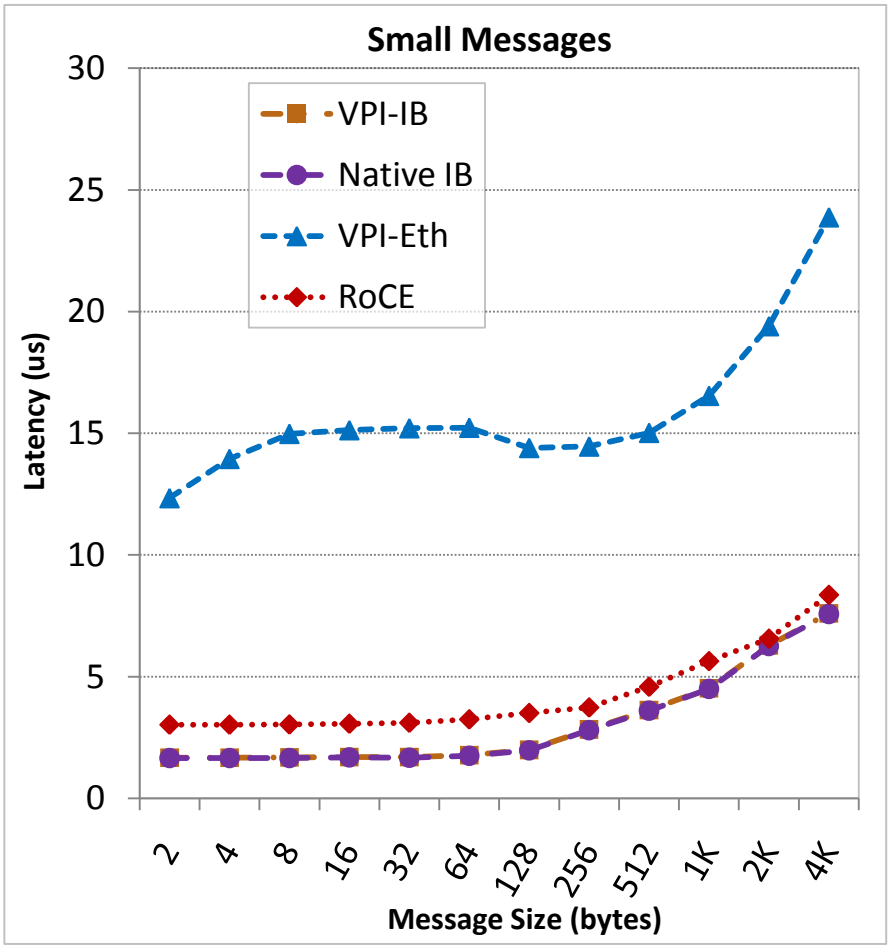
Modern Interconnects and Protocols

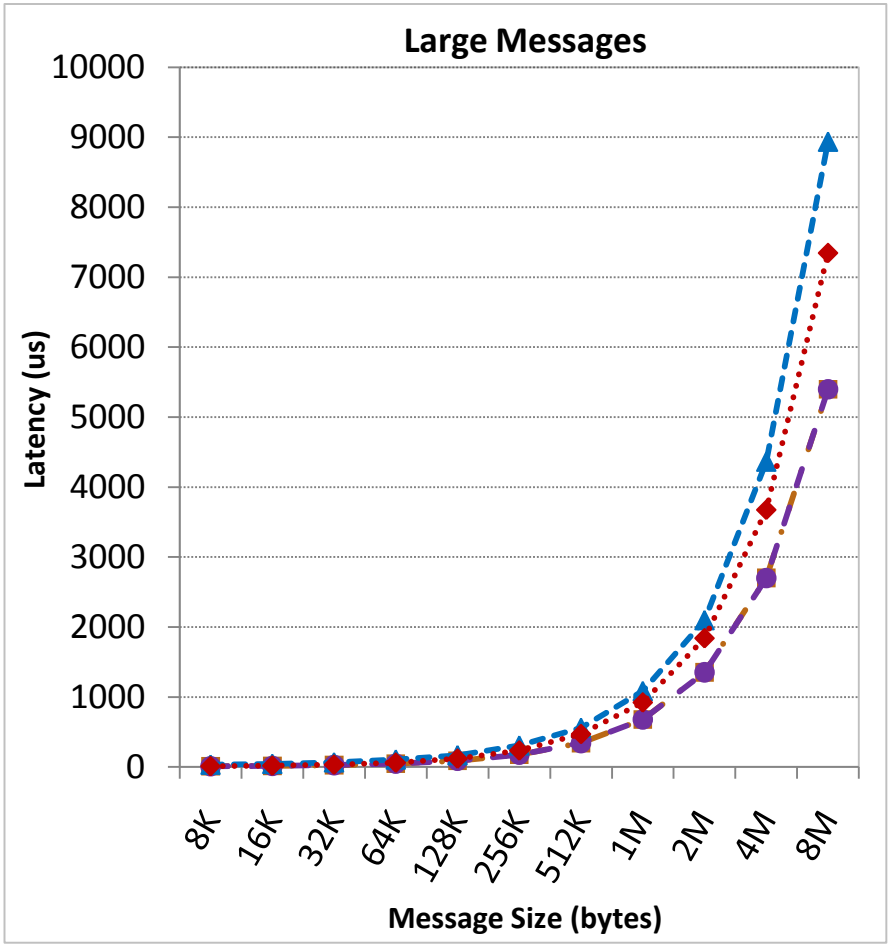


Case Studies

- **Low-level Network Performance**
- Clusters with Message Passing Interface (MPI)
- Datacenters with Sockets Direct Protocol (SDP) and TCP/IP (IPoIB)
- InfiniBand in WAN and Grid-FTP
- Cloud Computing: Hadoop and Memcached

Low-level Latency Measurements

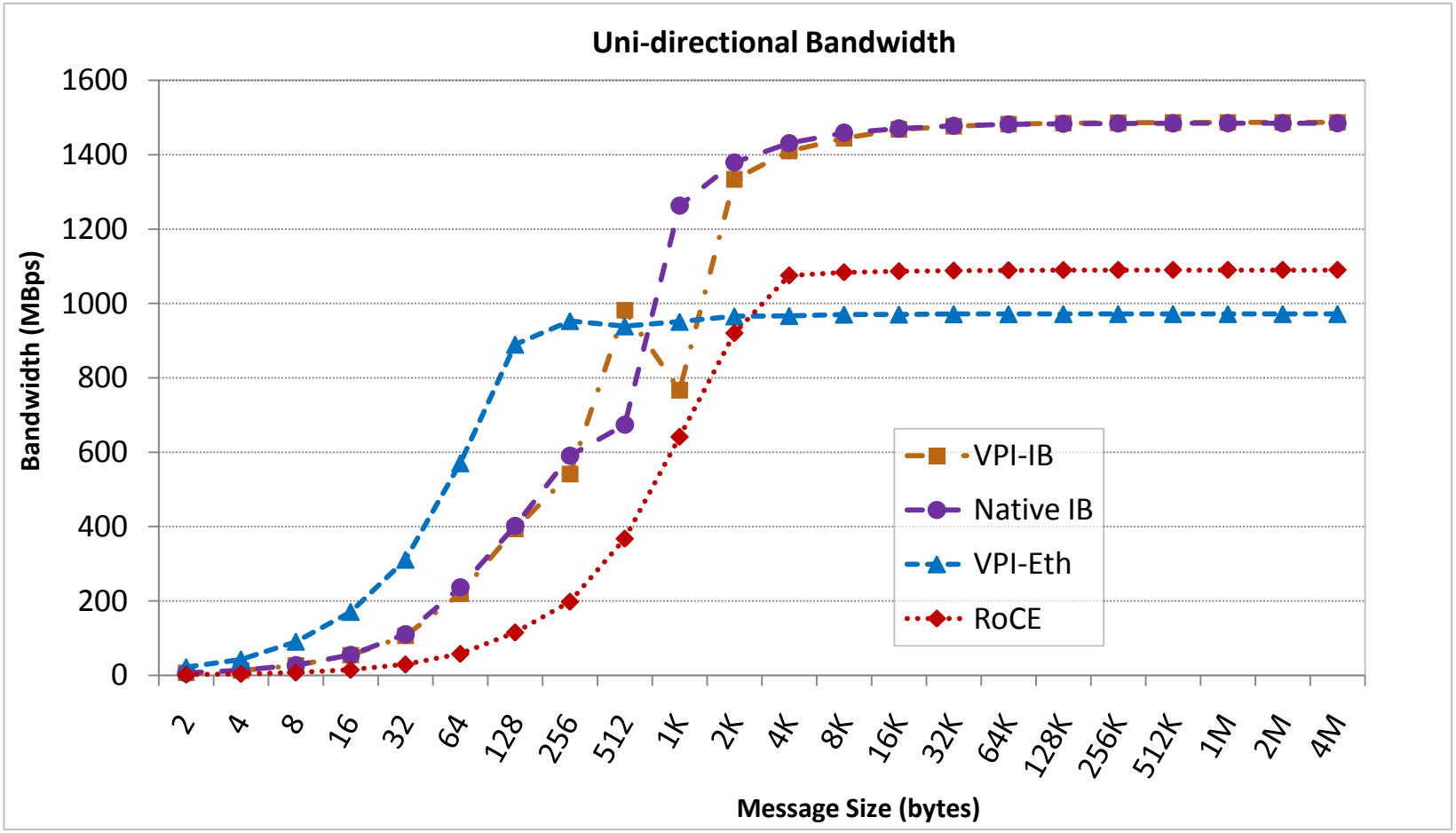




ConnectX-DDR: 2.4 GHz Quad-core (Nehalem) Intel with IB and 10GE switches

RoCE has a slight overhead compared to native IB because it operates at a slower clock rate (required to support only a 10Gbps link for Ethernet, as compared to a 32Gbps link for IB)

Low-level Uni-directional Bandwidth Measurements



ConnectX-DDR: 2.4 GHz Quad-core (Nehalem) Intel with IB and 10GE switches

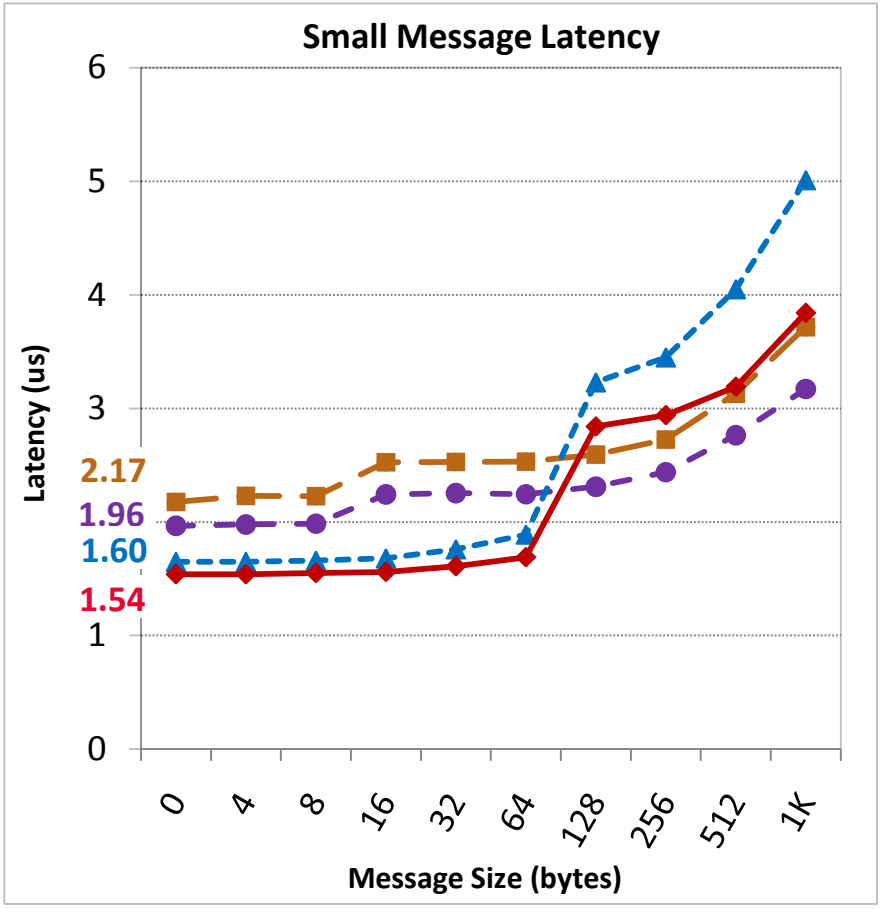
Case Studies

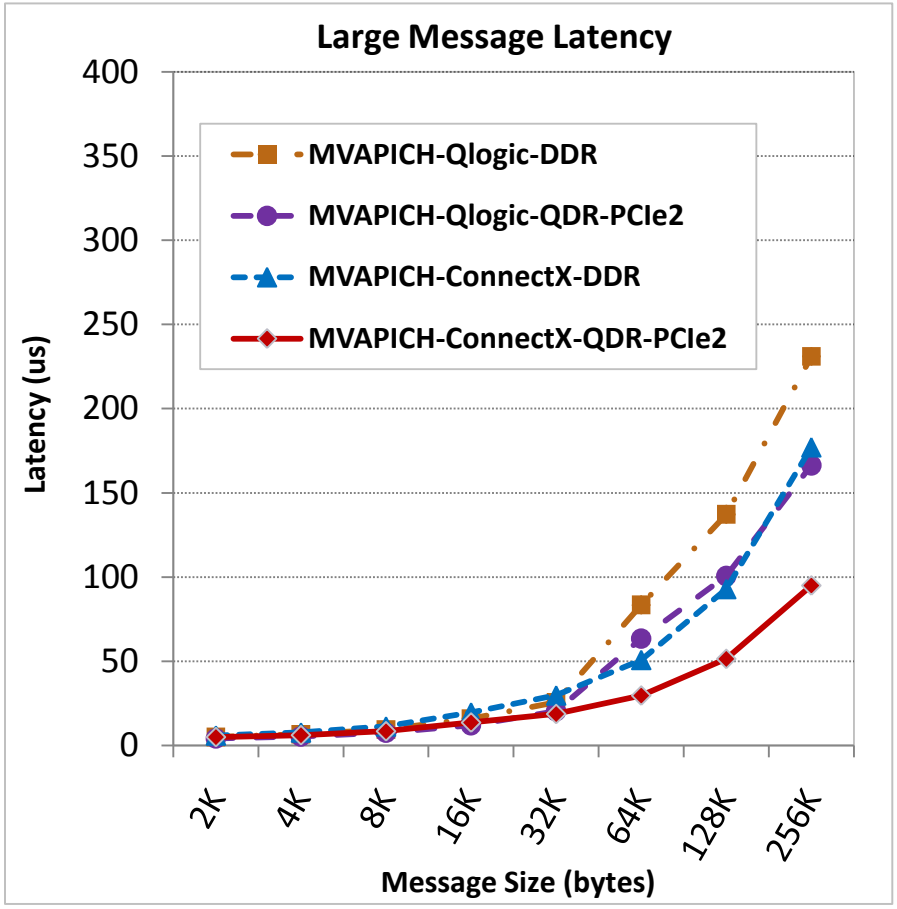
- Low-level Network Performance
- **Clusters with Message Passing Interface (MPI)**
- Datacenters with Sockets Direct Protocol (SDP) and TCP/IP (IPoIB)
- InfiniBand in WAN and Grid-FTP
- Cloud Computing: Hadoop and Memcached

MVAPICH/MVAPICH2 Software

- High Performance MPI Library for IB and HSE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2.2)
  - Used by more than 1,550 organizations in 60 countries
  - More than 66,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 11th ranked 81,920-core cluster (Pleiades) at NASA
    - 15th ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, HSE and server vendors including Open Fabrics Enterprise Distribution (OFED) and Linux Distros (RedHat and SuSE)
  - http://mvapich.cse.ohio-state.edu

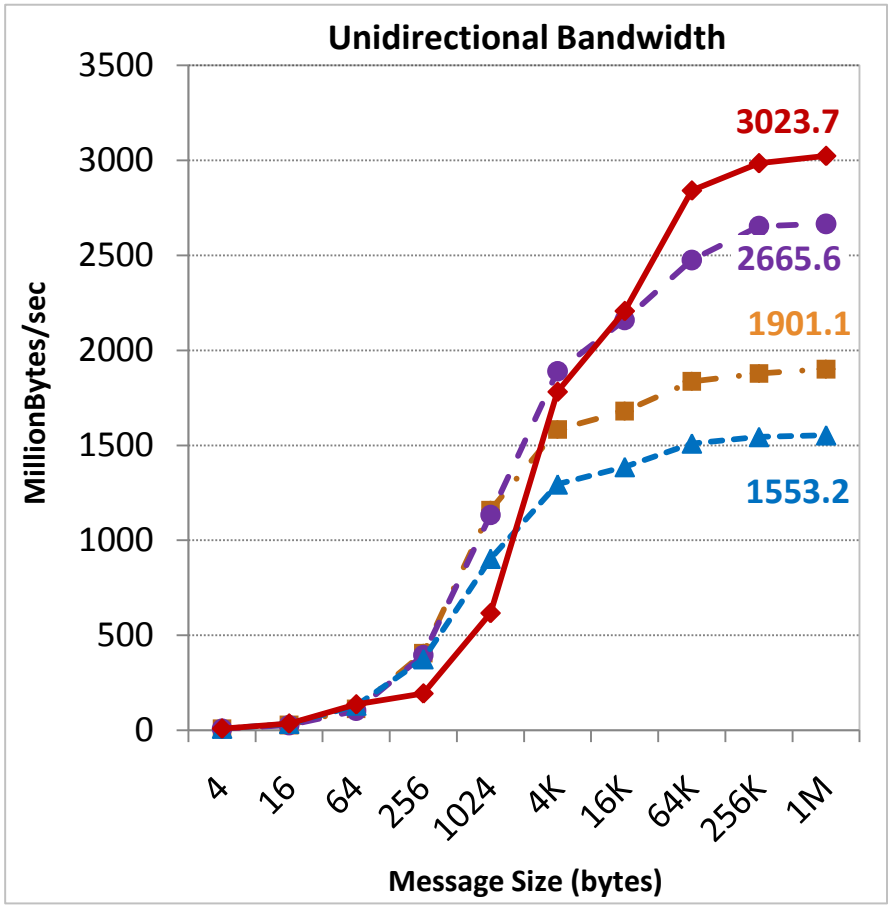
One-way Latency: MPI over IB

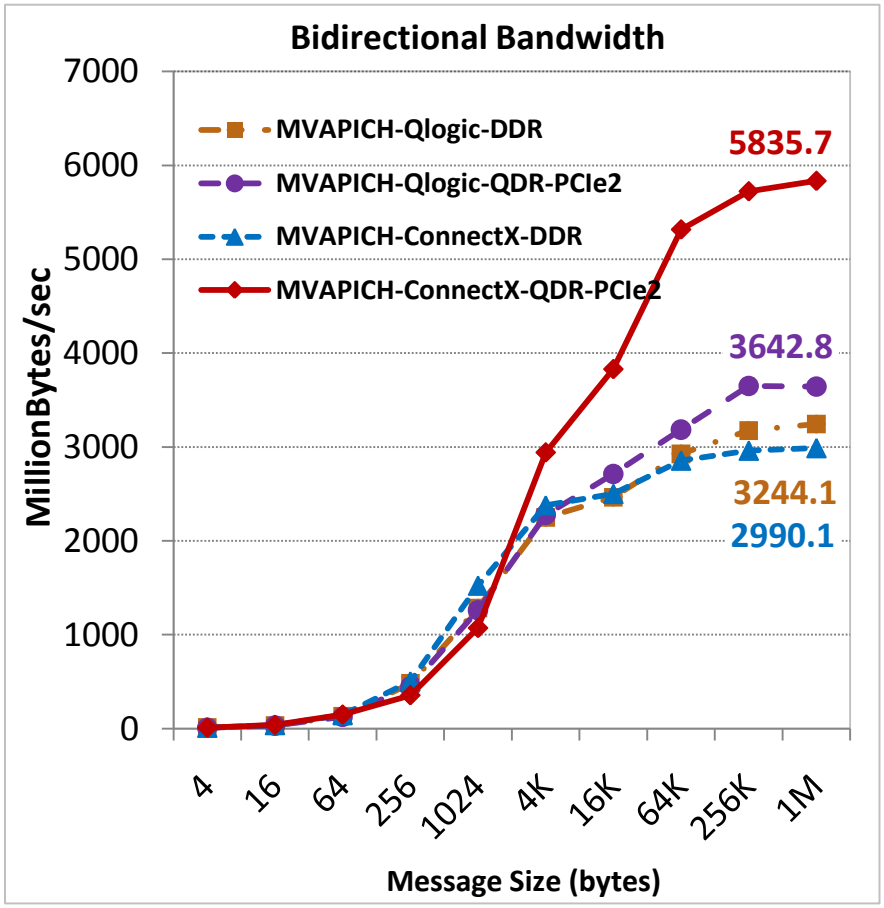




All numbers taken on 2.4 GHz Quad-core (Nehalem) Intel with IB switch

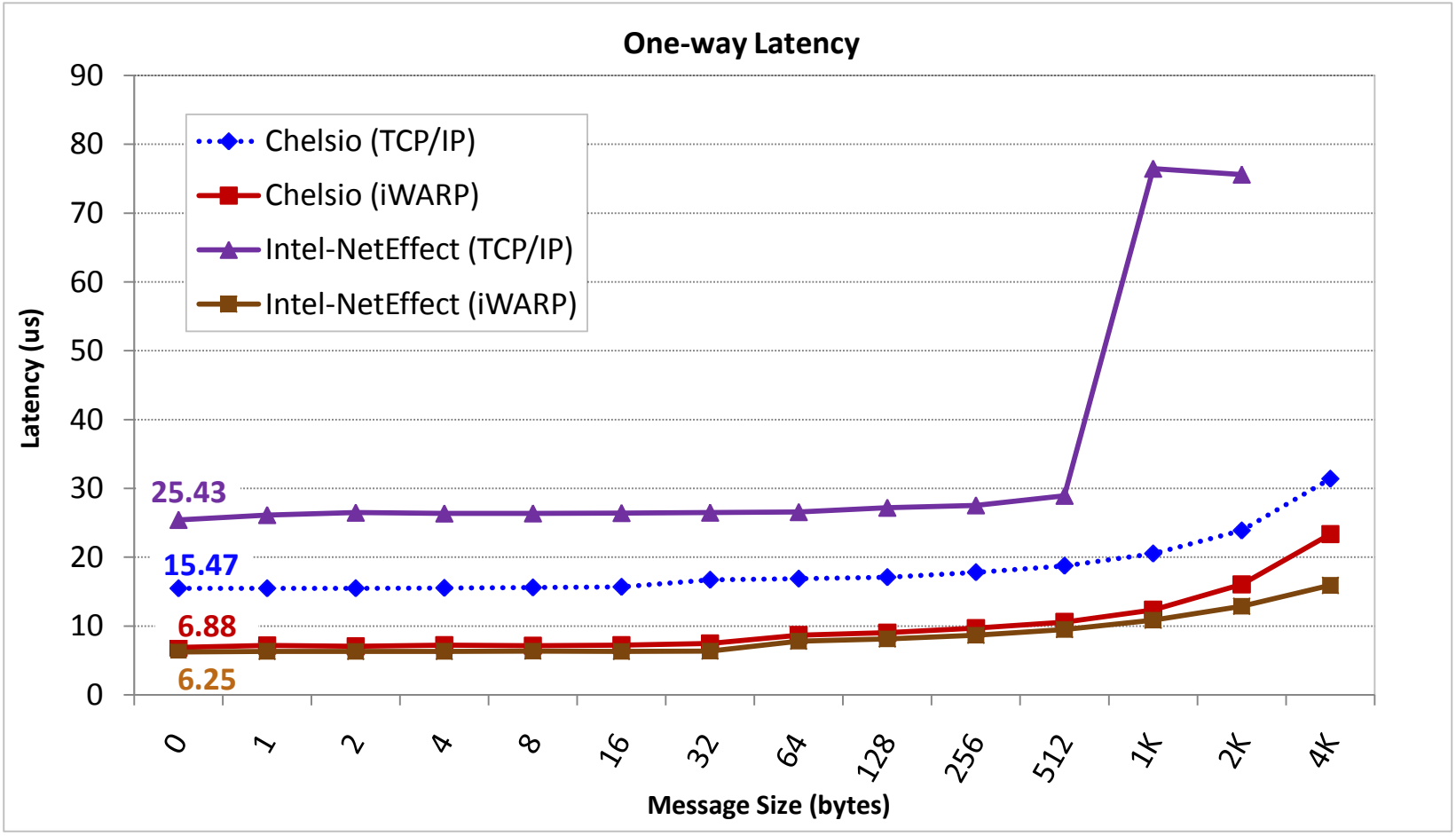
Bandwidth: MPI over IB





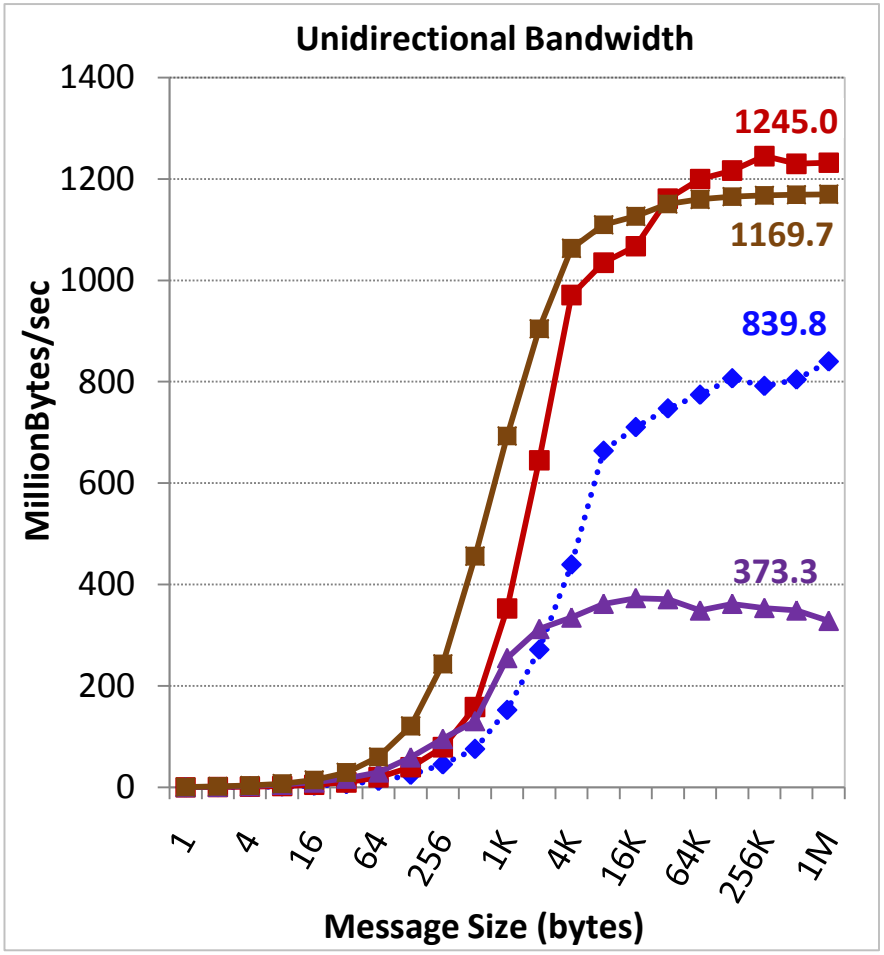
All numbers taken on 2.4 GHz Quad-core (Nehalem) Intel with IB switch

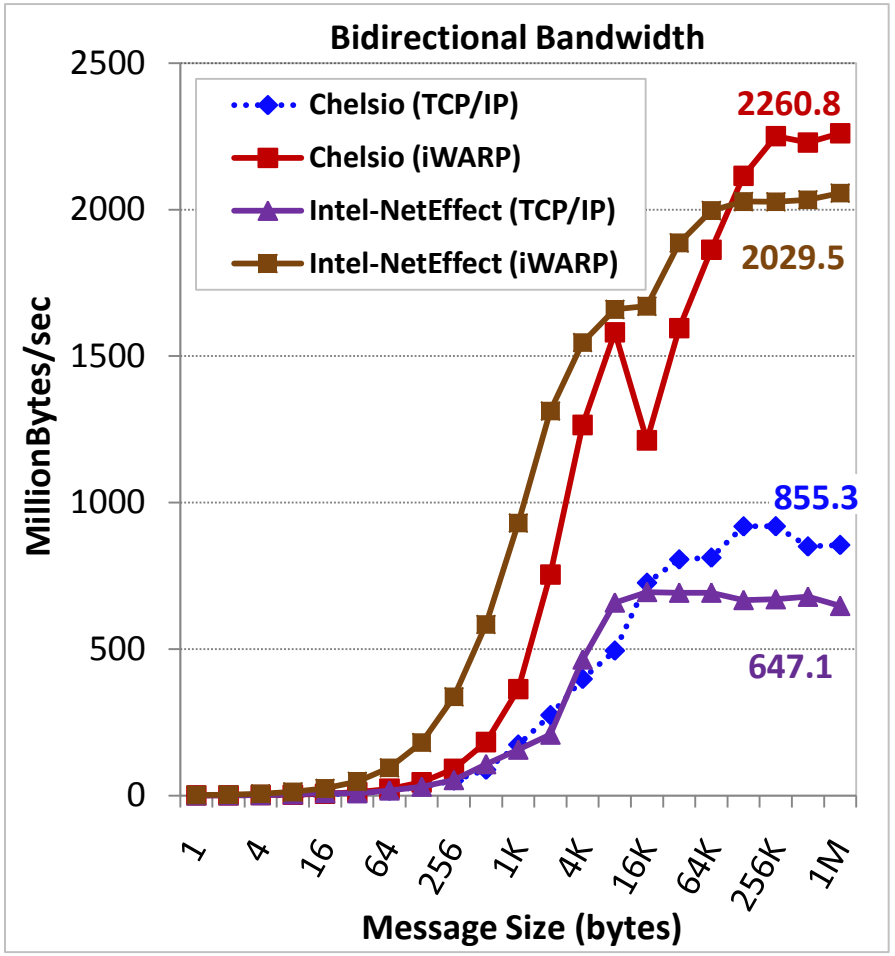
One-way Latency: MPI over iWARP



2.4 GHz Quad-core Intel (Clovertown) with 10GE (Fulcrum) Switch

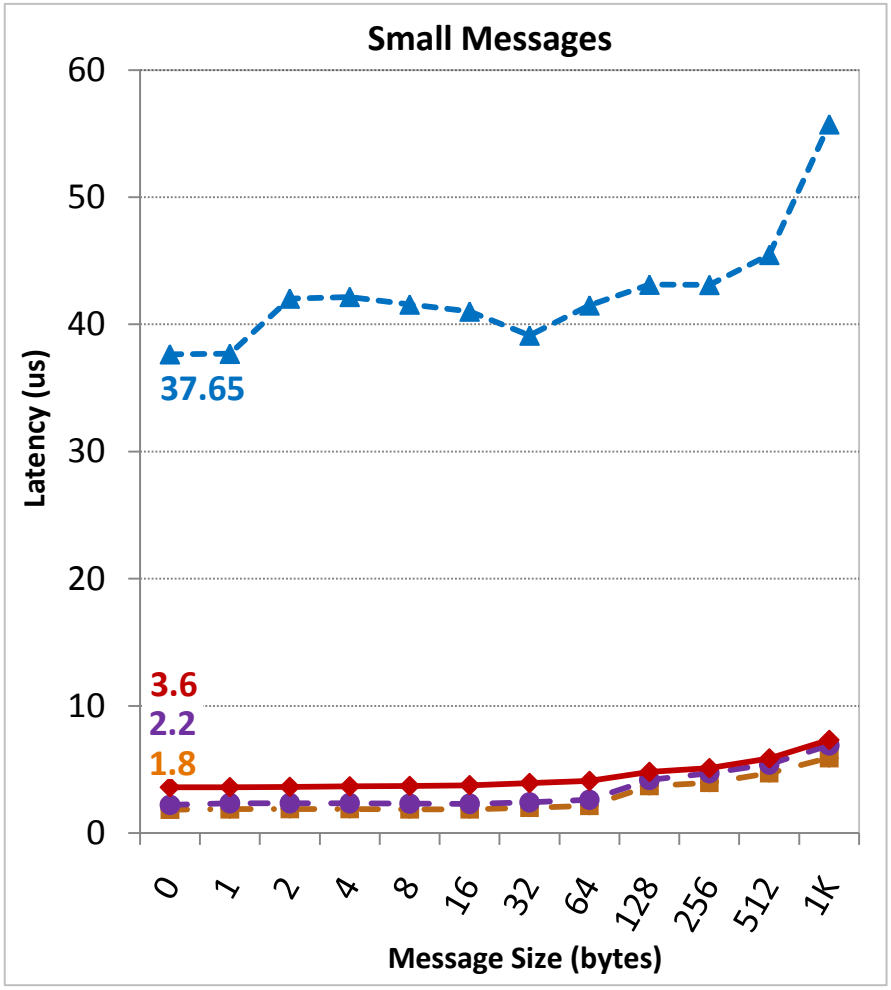
Bandwidth: MPI over iWARP

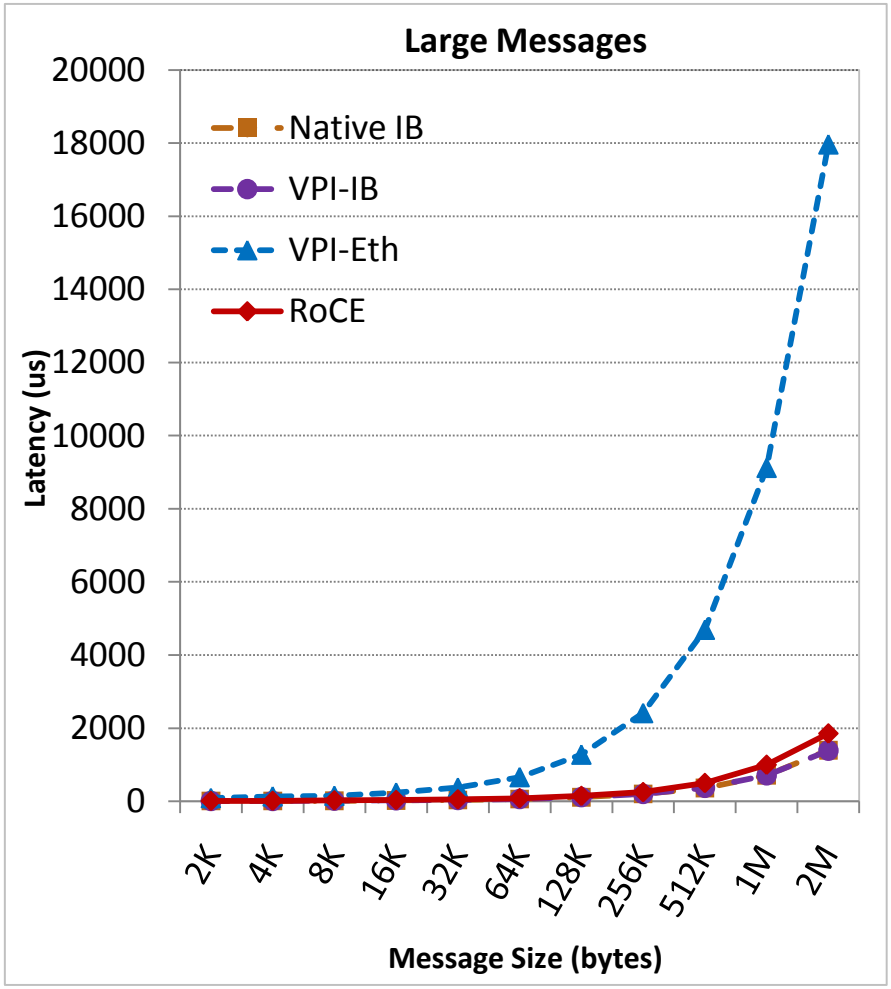




2.33 GHz Quad-core Intel (Clovertown) with 10GE (Fulcrum) Switch

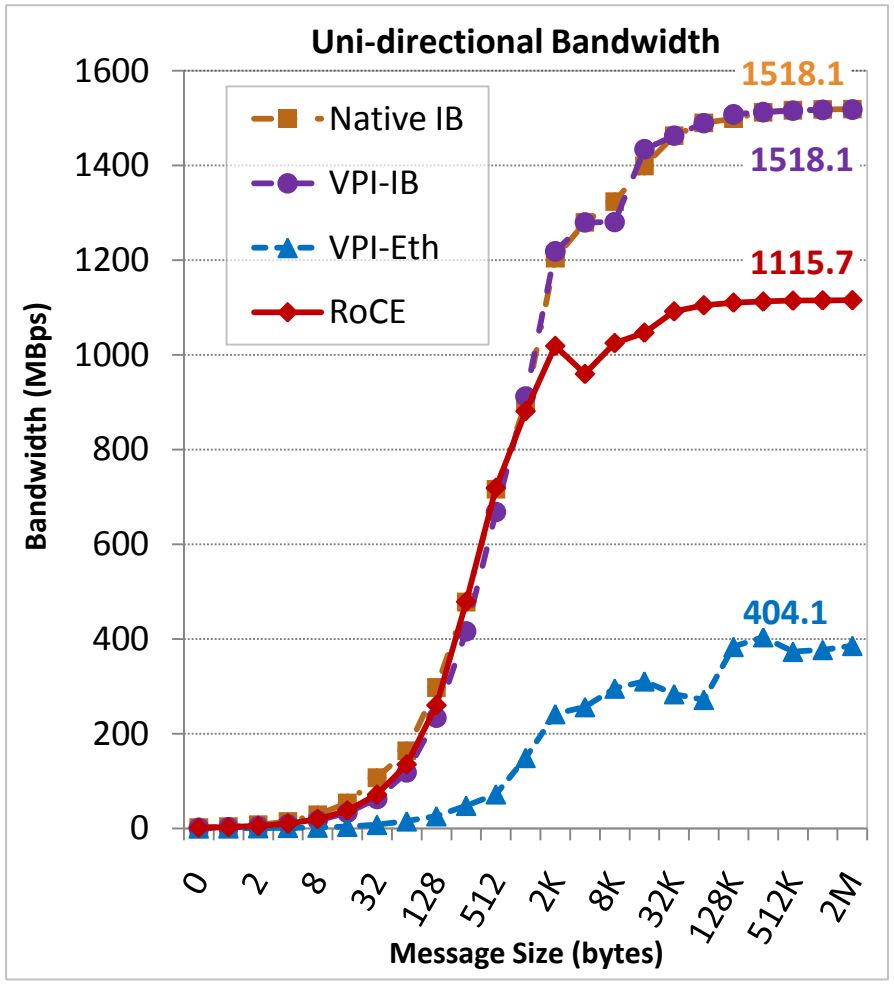
Convergent Technologies: MPI Latency

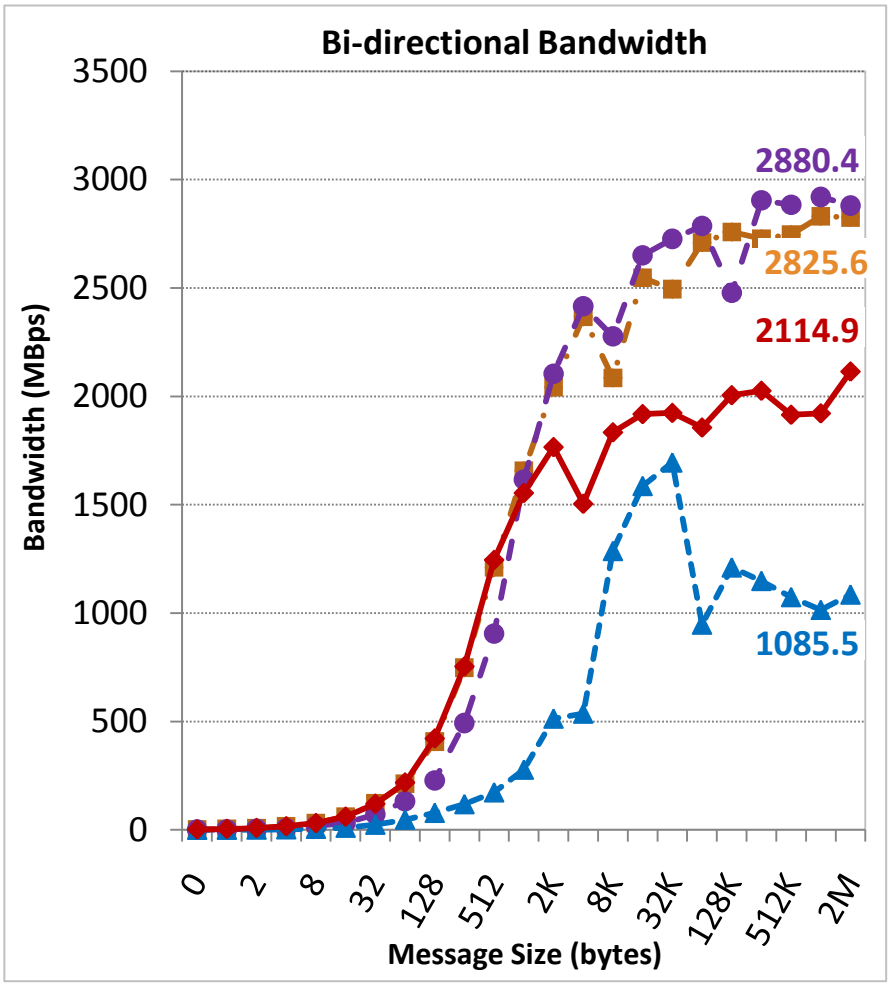




ConnectX-DDR: 2.4 GHz Quad-core (Nehalem) Intel with IB and 10GE switches

Convergent Technologies: MPI Uni- and Bi-directional Bandwidth



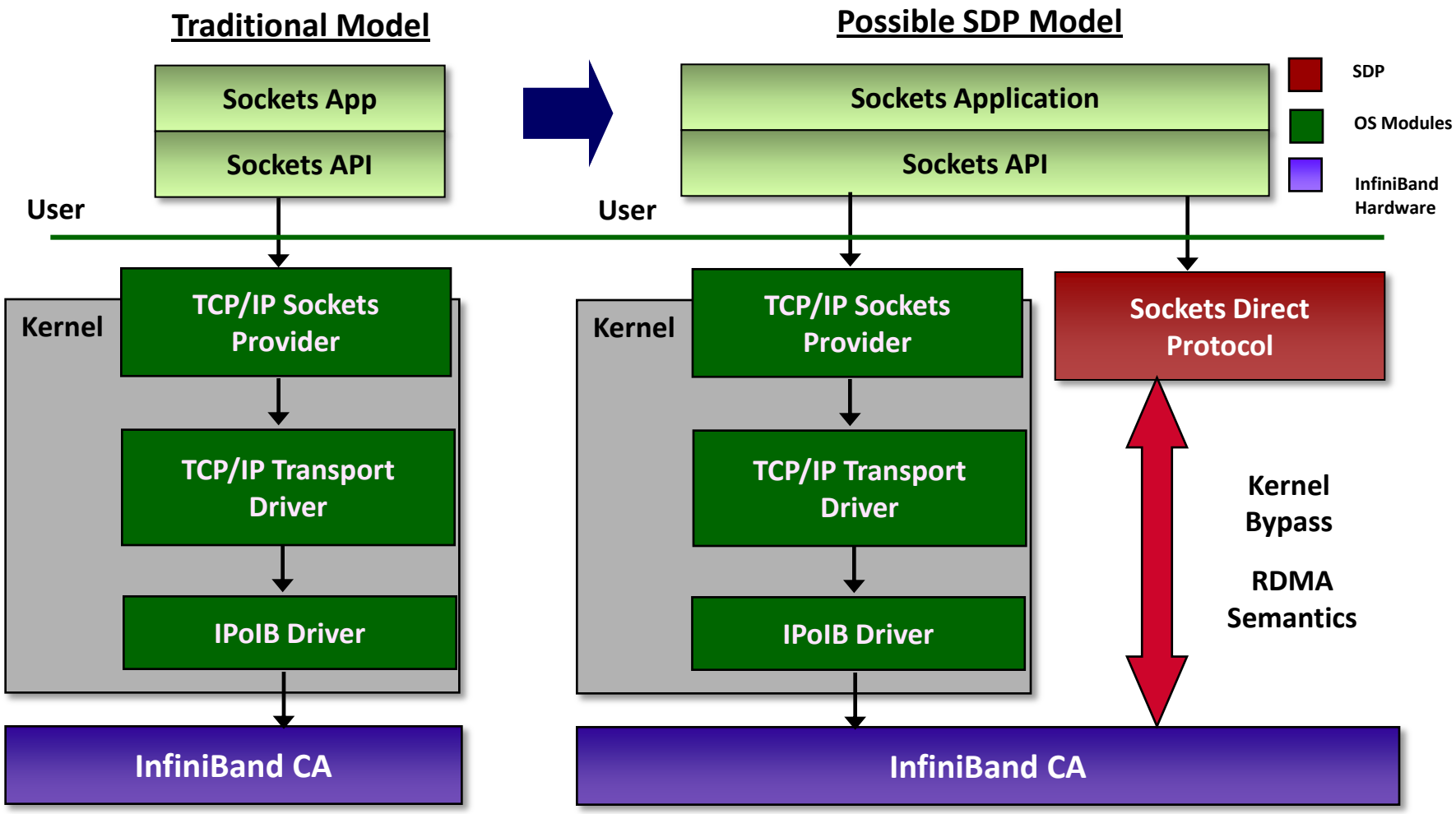


ConnectX-DDR: 2.4 GHz Quad-core (Nehalem) Intel with IB and 10GE switches

Case Studies

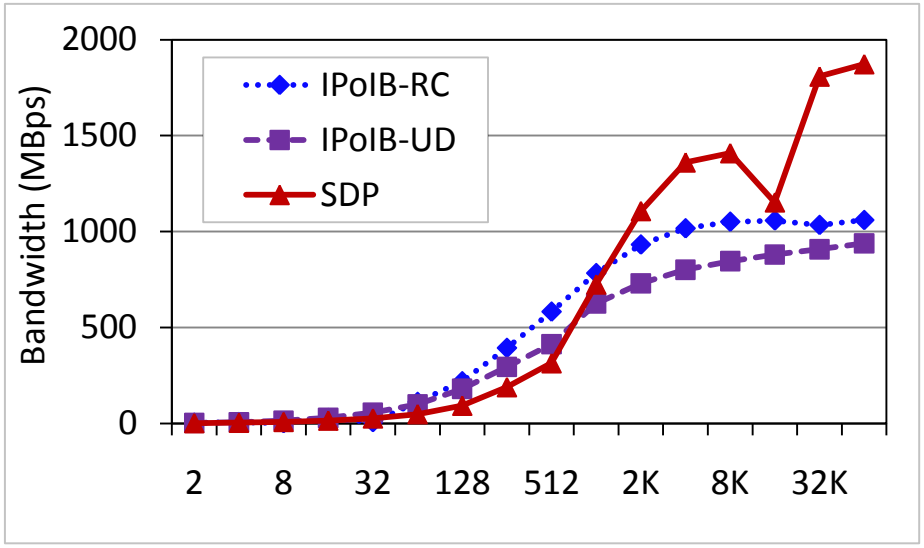
- Low-level Network Performance
- Clusters with Message Passing Interface (MPI)
- **Datacenters with Sockets Direct Protocol (SDP) and TCP/IP (IPoIB)**
- InfiniBand in WAN and Grid-FTP
- Cloud Computing: Hadoop and Memcached

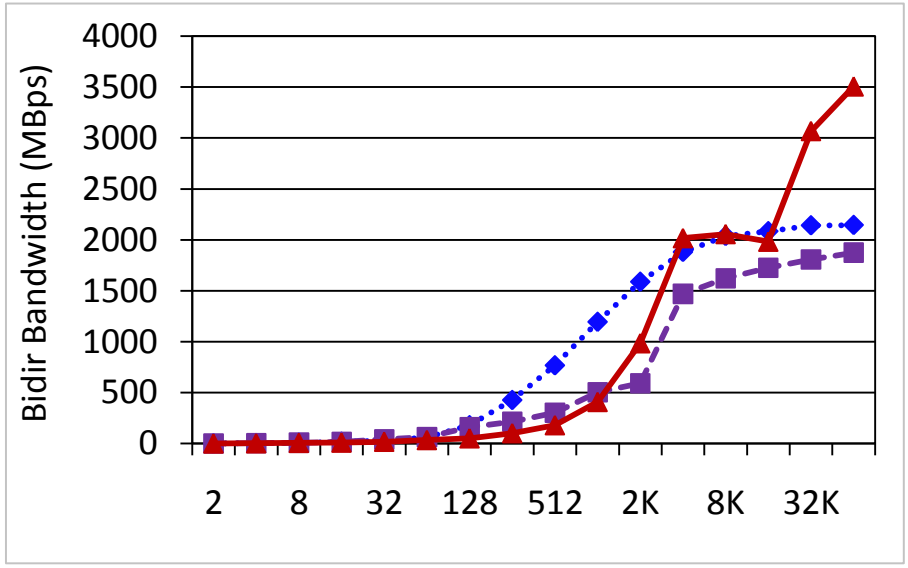
IPoIB vs. SDP Architectural Models

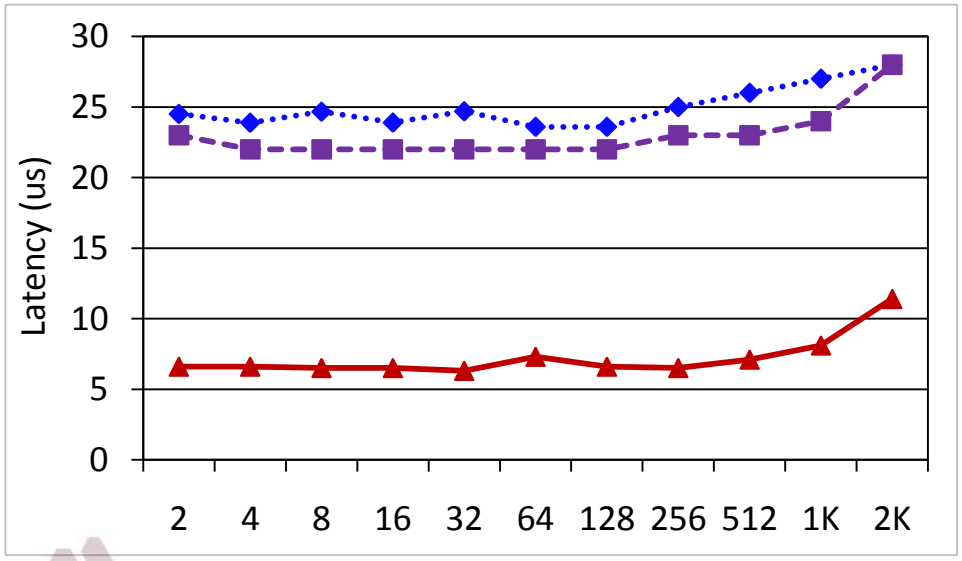


(Source: InfiniBand Trade Association)

SDP vs. IPoIB (IB QDR)







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

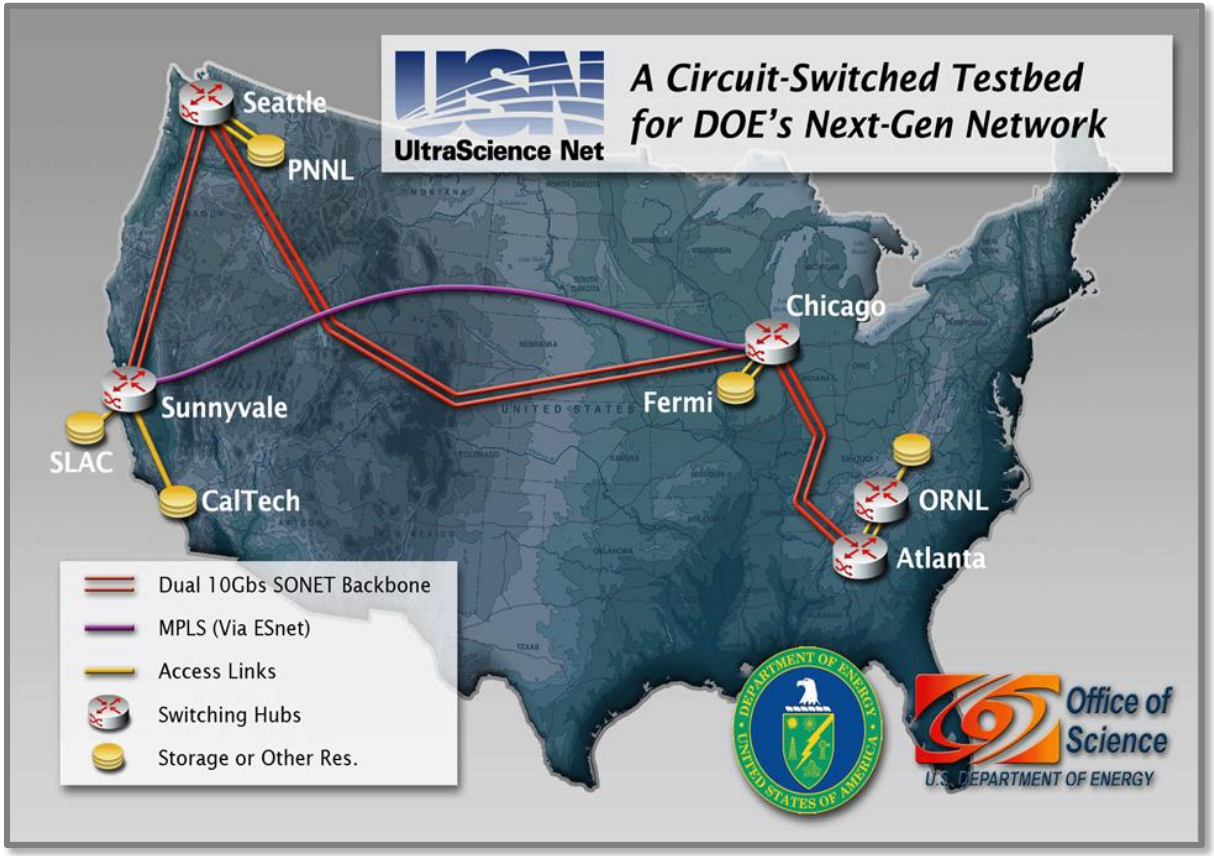
Case Studies

- Low-level Network Performance
- Clusters with Message Passing Interface (MPI)
- Datacenters with Sockets Direct Protocol (SDP) and TCP/IP (IPoIB)
- **InfiniBand in WAN and Grid-FTP**
- Cloud Computing: Hadoop and Memcached

IB on the WAN

- Option 1: Layer-1 Optical networks
  - IB standard specifies link, network and transport layers
  - Can use any layer-1 (though the standard says copper and optical)
- Option 2: Link Layer Conversion Techniques
  - InfiniBand-to-Ethernet conversion at the link layer: switches available from multiple companies (e.g., Obsidian)
    - Technically, it's not conversion; it's just tunneling (L2TP)
  - InfiniBand's network layer is IPv6 compliant

UltraScience Net: Experimental Research Network Testbed



Since 2005

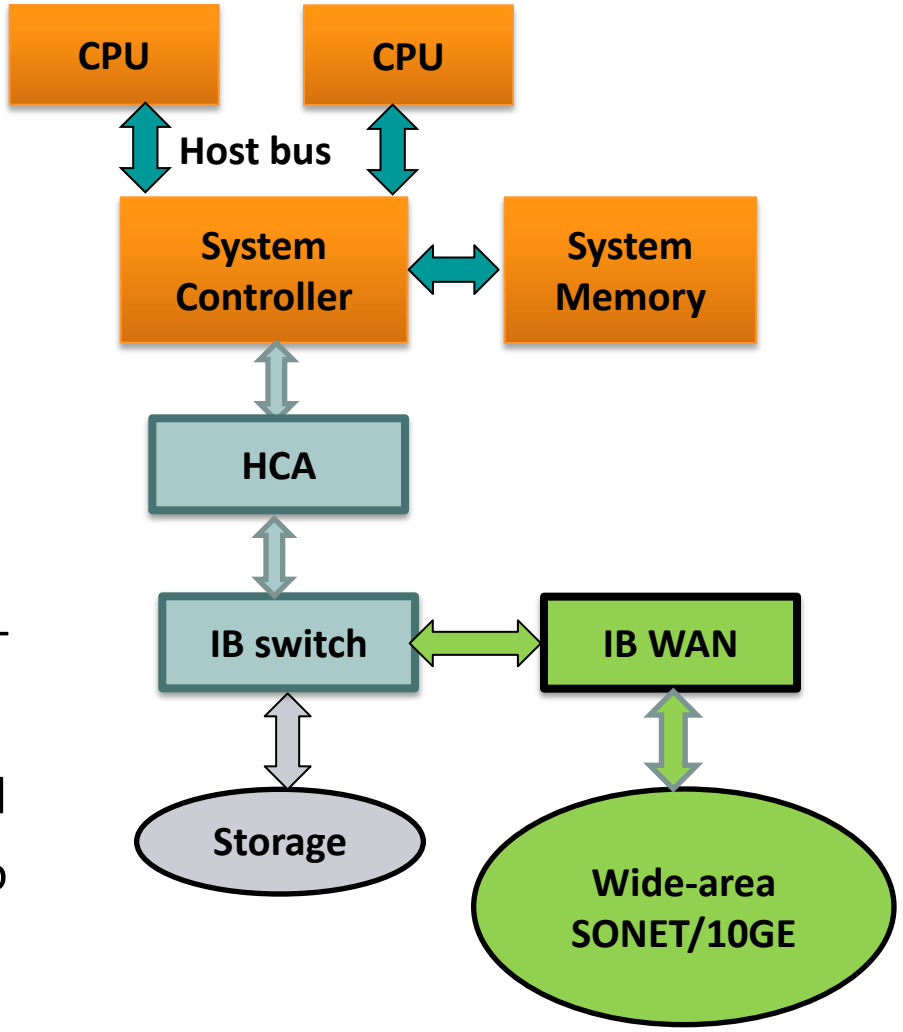
This and the following IB WAN slides are courtesy Dr. Nagi Rao (ORNL)

Features

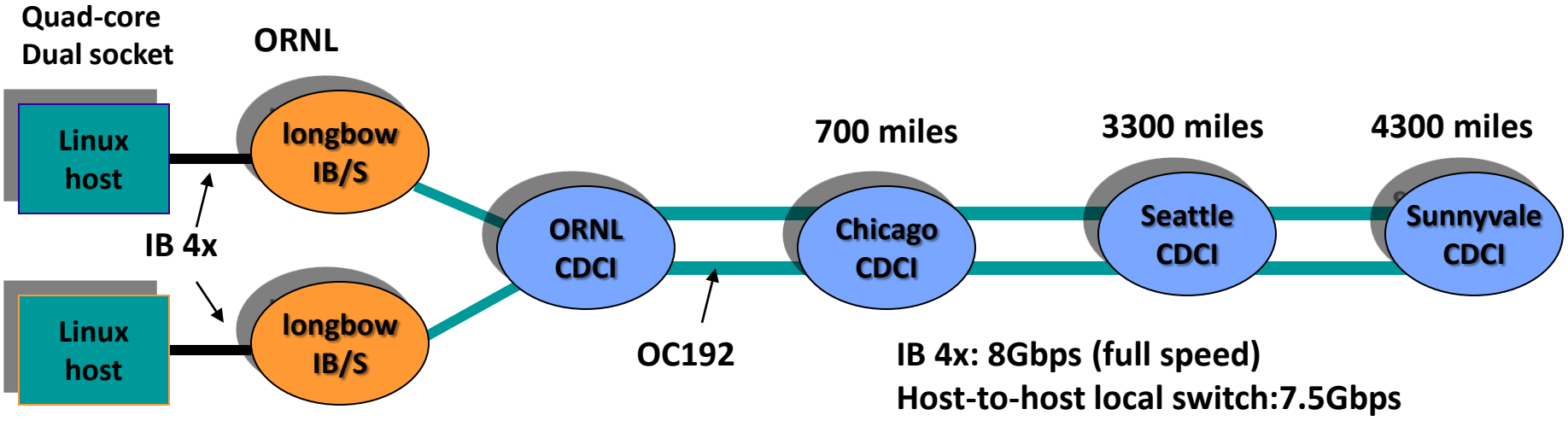
- End-to-end guaranteed bandwidth channels
- Dynamic, in-advance, reservation and provisioning of fractional/full lambdas
- Secure control-plane for signaling
- Peering with ESnet, National Science Foundation CHEETAH, and other networks

IB-WAN Connectivity with Obsidian Switches

- Supports SONET OC-192 or 10GE LAN-PHY/WAN-PHY
- Idea is to make remote storage "appear" local
- IB-WAN switch does frame conversion
  - IB standard allows per-hop credit-based flow control
  - IB-WAN switch uses large internal buffers to allow enough credits to fill the wire



InfiniBand Over SONET: Obsidian Longbows RDMA throughput measurements over USN



ORNL loop -0.2 mile: 7.48Gbps

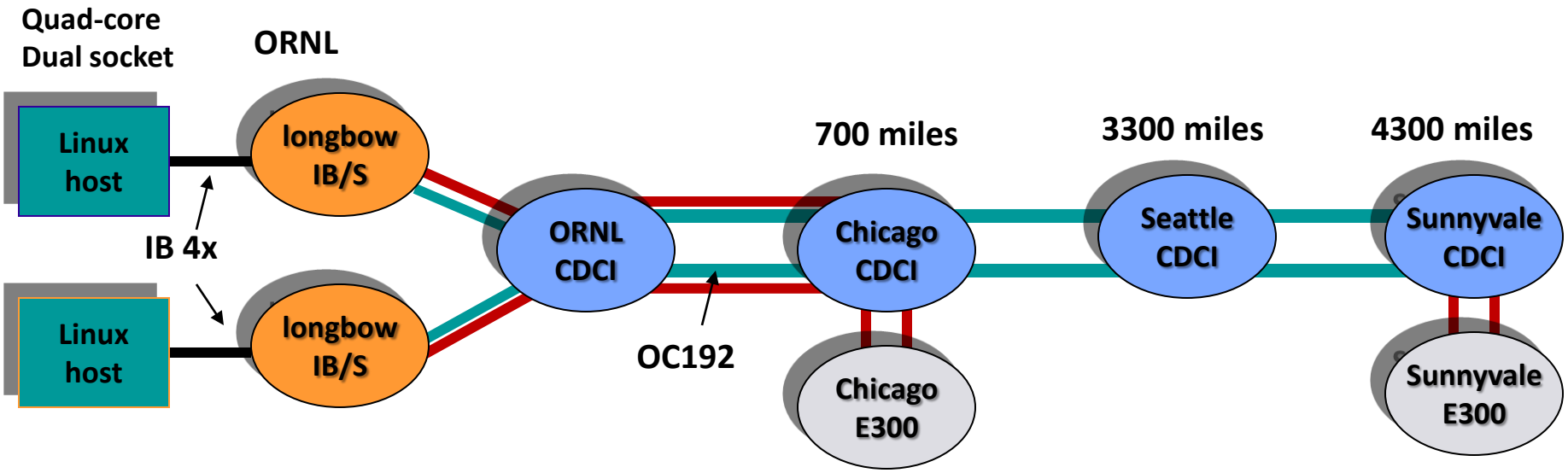
ORNL-Chicago loop – 1400 miles: 7.47Gbps

ORNL- Chicago - Seattle loop – 6600 miles: 7.37Gbps

ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles: 7.34Gbps

Hosts:
Dual-socket Quad-core 2 GHz AMD Opteron, 4GB memory, 8-lane PCI-Express slot, Dual-port Voltaire 4x SDR HCA

IB over 10GE LAN-PHY and WAN-PHY



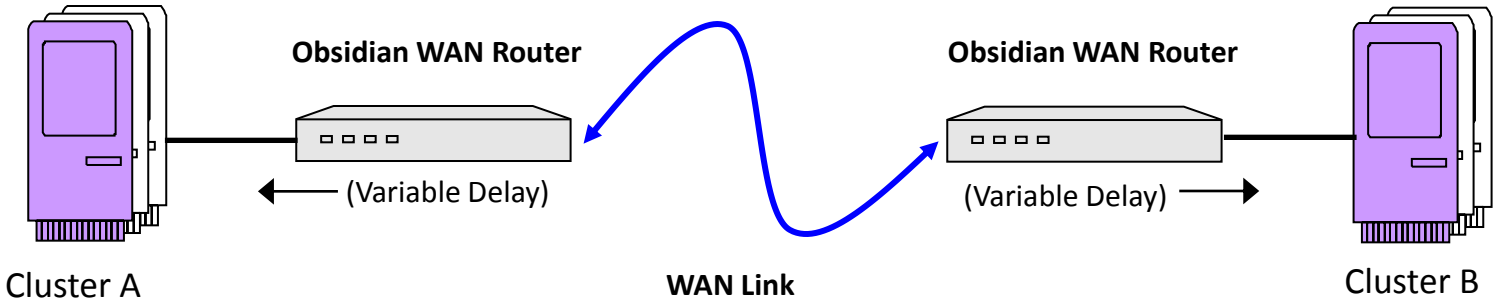
ORNL loop -0.2 mile: 7.5Gbps

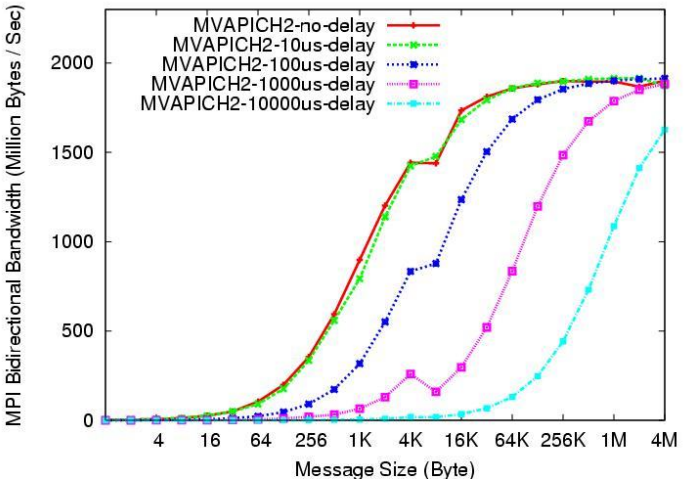
ORNL-Chicago loop – 1400 miles: 7.49Gbps

ORNL- Chicago - Seattle loop – 6600 miles: 7.39Gbps

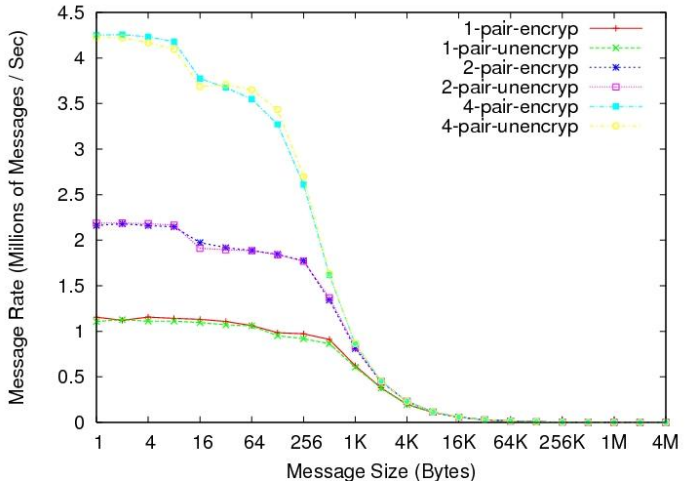
ORNL – Chicago – Seattle - Sunnyvale loop – 8600 miles: 7.36Gbps

MPI over IB-WAN: Obsidian Routers





MPI Bidirectional Bandwidth



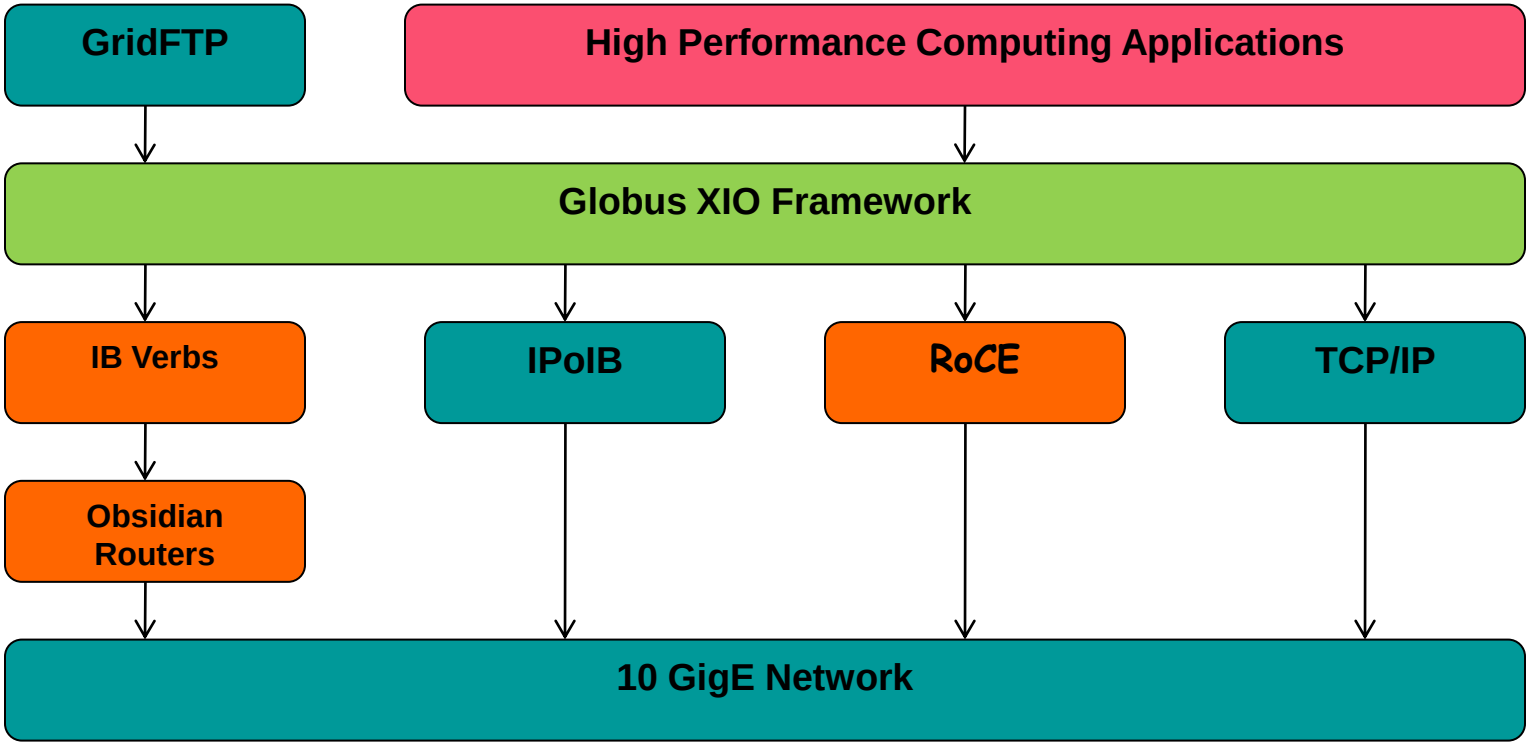
Impact of Encryption on Message Rate (Delay 0 ms)

Delay (us)	Distance (km)
10	2
100	20
1000	200
10000	2000

Hardware encryption has no impact on performance for less communicating streams

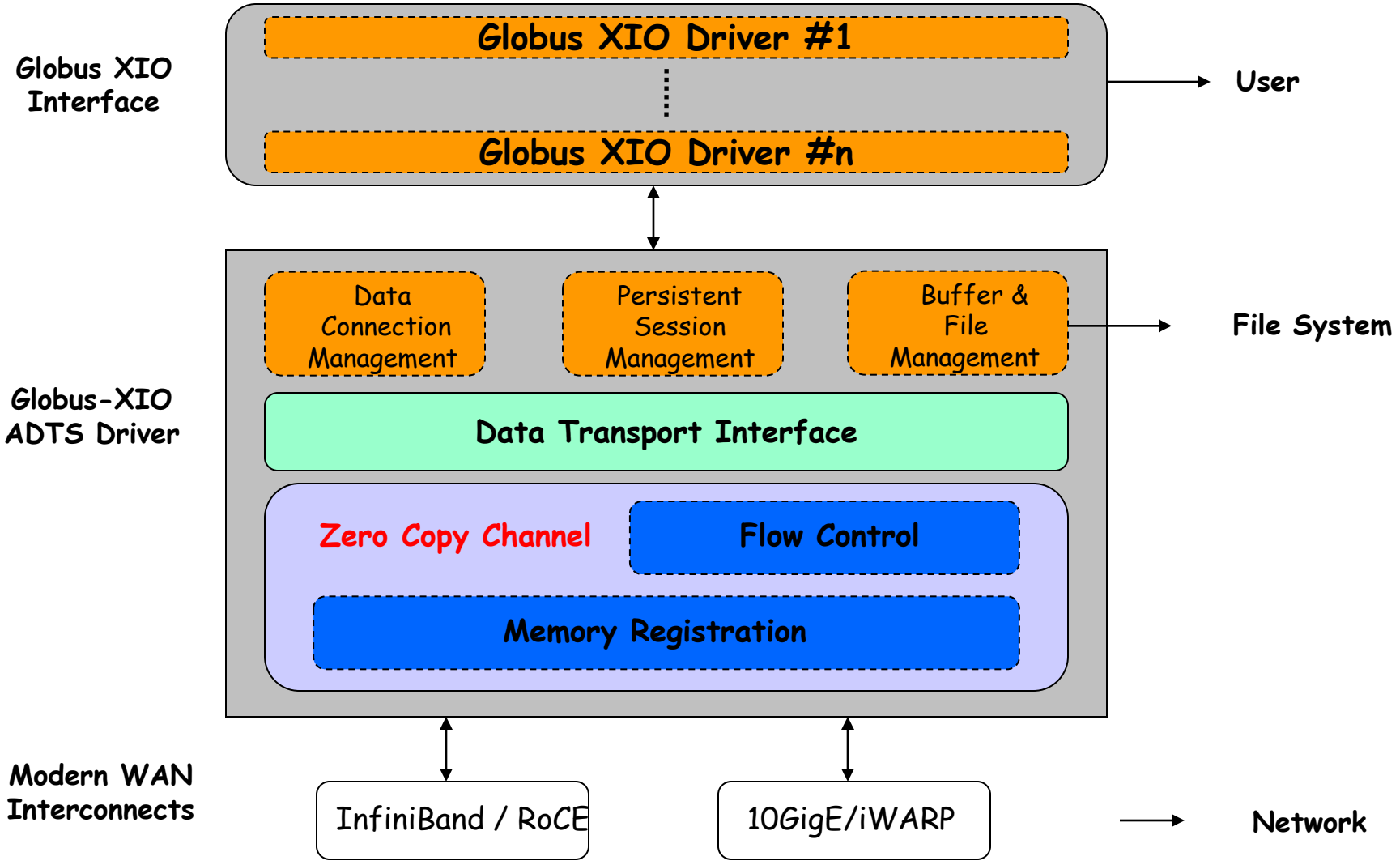
S. Narravula, H. Subramoni, P. Lai, R. Noronha and D. K. Panda, Performance of HPC Middleware over InfiniBand WAN, Int'l Conference on Parallel Processing (ICPP '08), September 2008.

Communication Options in Grid

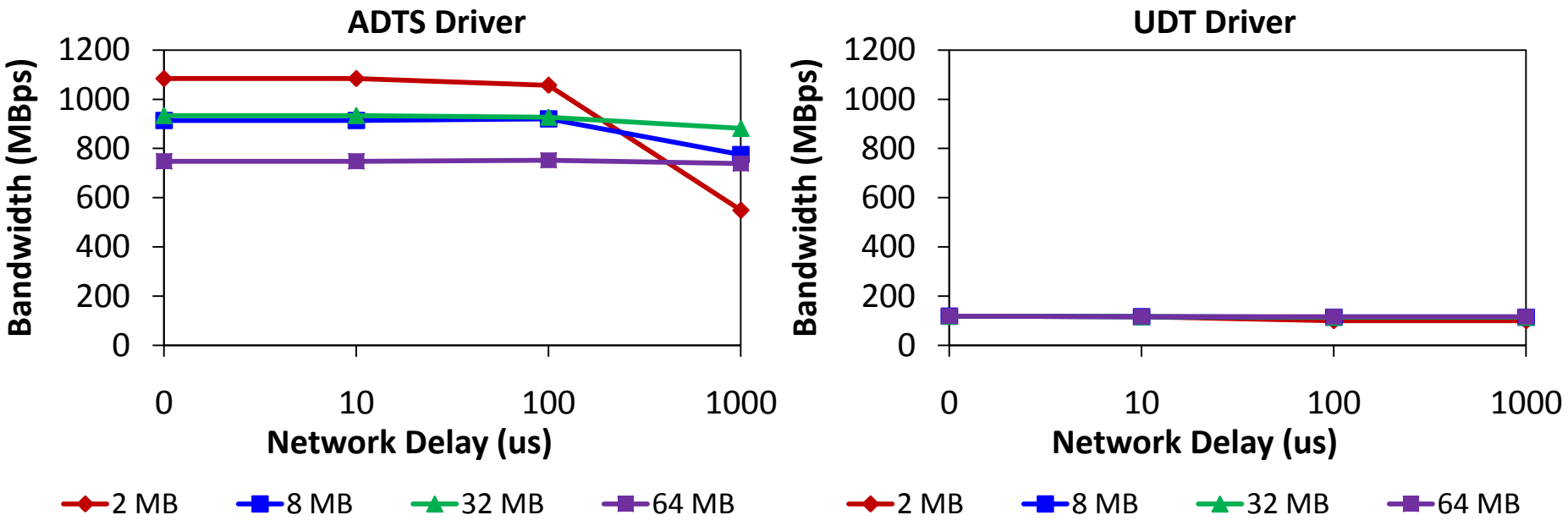


- Multiple options exist to perform data transfer on Grid
- Globus-XIO framework currently does not support IB natively
- We create the Globus-XIO ADTS driver and add native IB support to GridFTP

Globus-XIO Framework with ADTS Driver

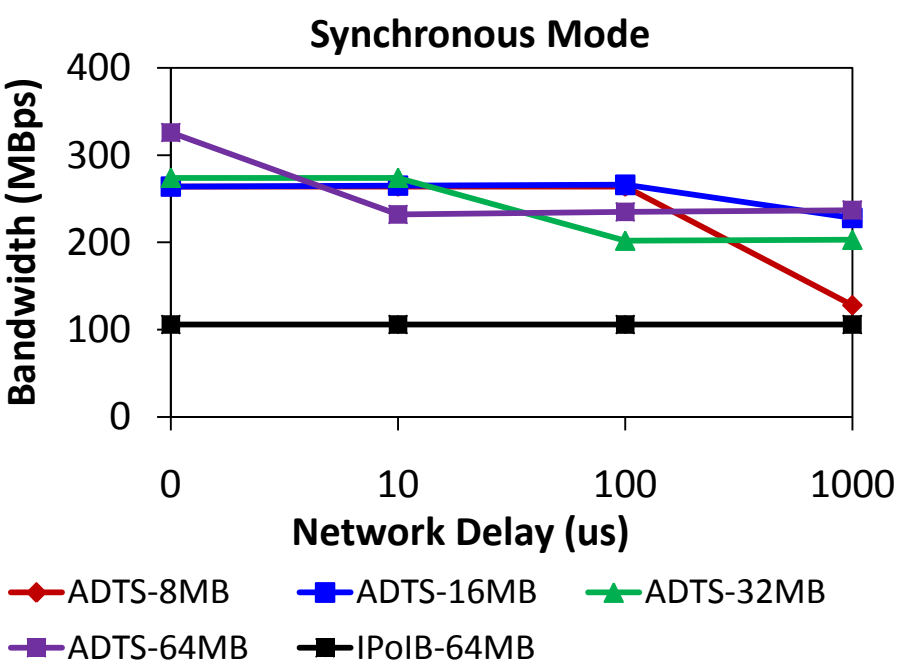


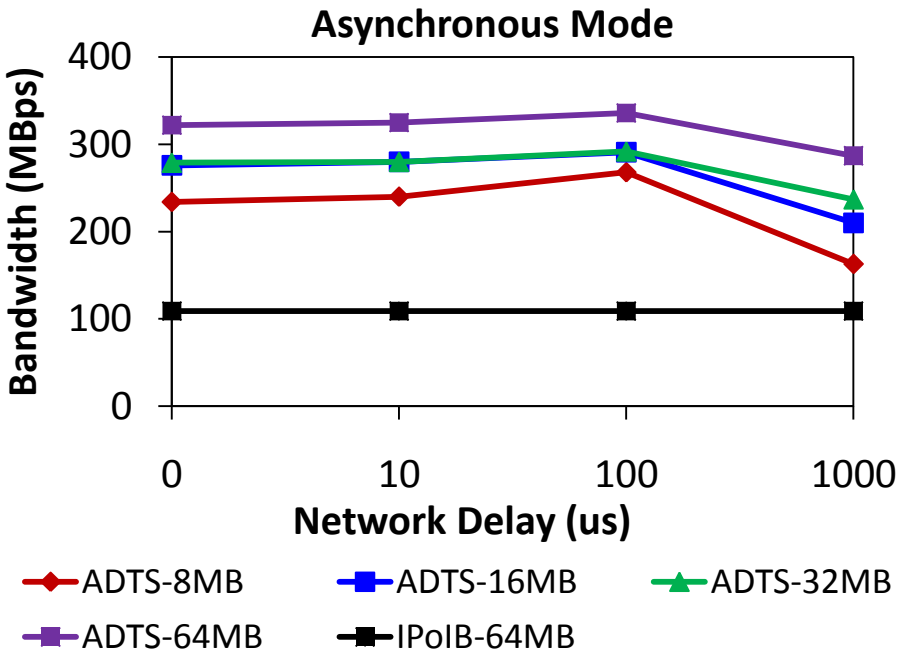
Performance of Memory Based Data Transfer



- Performance numbers obtained while transferring 128 GB of aggregate data in chunks of 256 MB files
- ADTS based implementation is able to saturate the link bandwidth
- Best performance for ADTS obtained when performing data transfer with a network buffer of size 32 MB

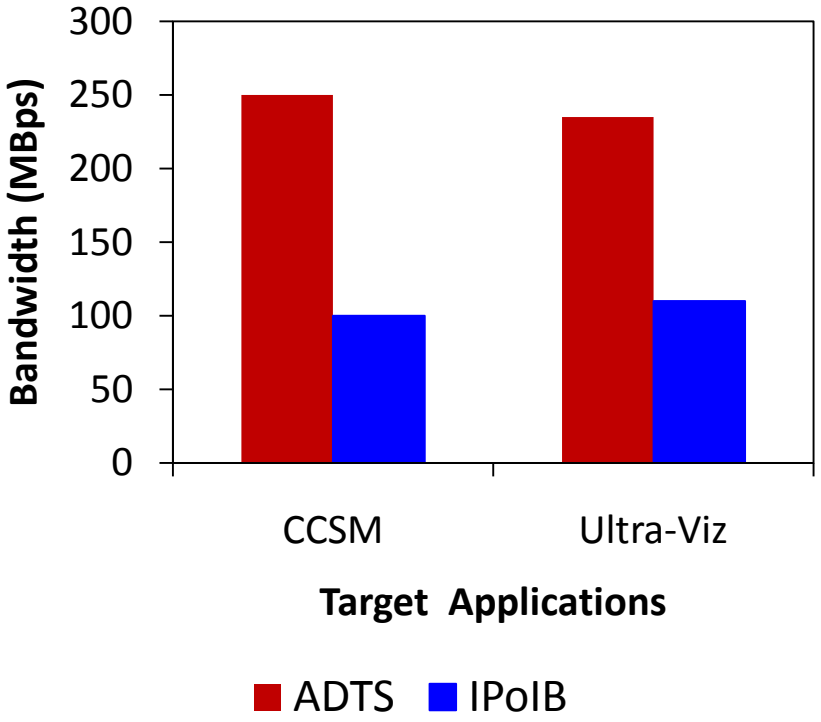
Performance of Disk Based Data Transfer





- Performance numbers obtained while transferring 128 GB of aggregate data in chunks of 256 MB files
- Predictable as well as better performance when Disk-IO threads assist network thread (Asynchronous Mode)
- Best performance for ADTS obtained with a circular buffer with individual buffers of size 64 MB

Application Level Performance



- Application performance for FTP *get* operation for disk based transfers
- Community Climate System Model (CCSM)
  - Part of Earth System Grid Project
  - Transfers 160 TB of total data in chunks of 256 MB
  - Network latency - 30 ms
- Ultra-Scale Visualization (Ultra-Viz)
  - Transfers files of size 2.6 GB
  - Network latency - 80 ms

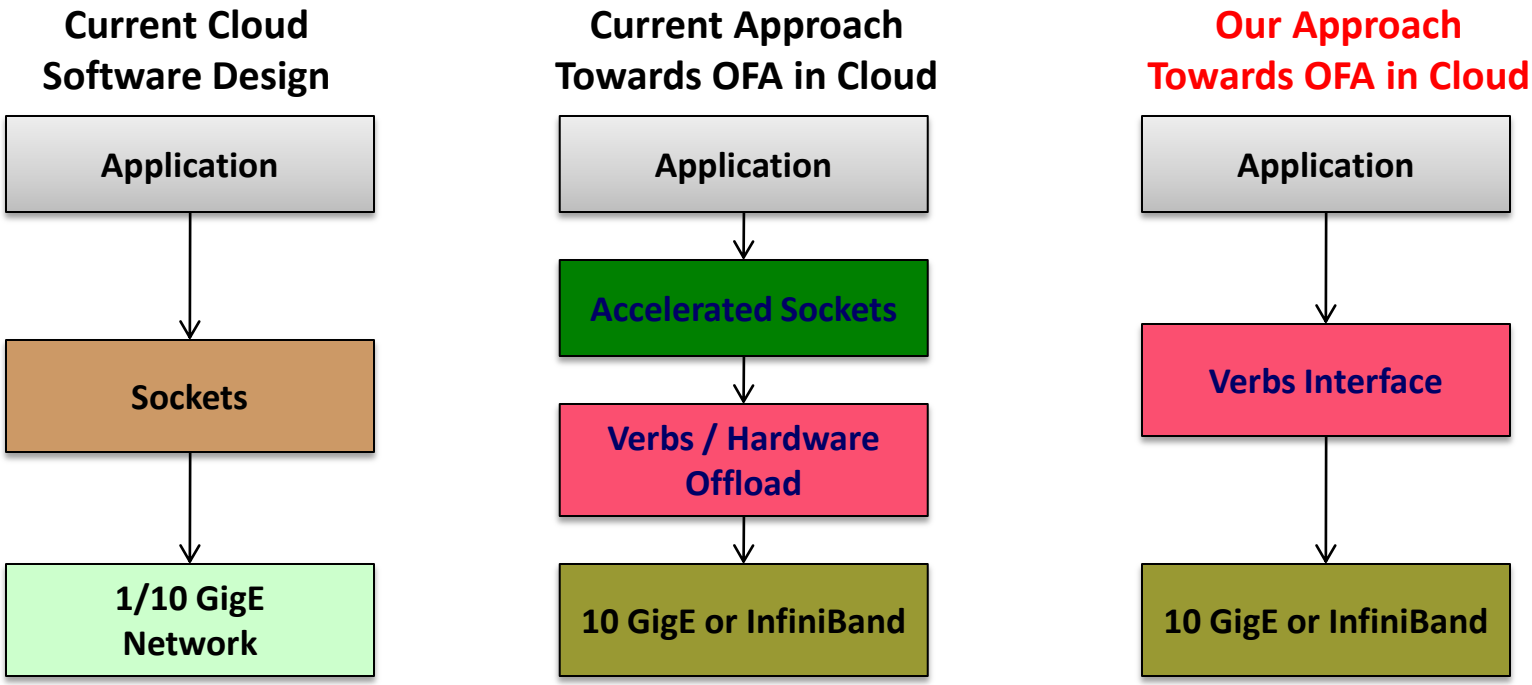
The ADTS driver out performs the UDT driver using IPoIB by more than 100%

H. Subramoni, P. Lai, R. Kettimuthu and D. K. Panda, High Performance Data Transfer in Grid Environment Using GridFTP over InfiniBand, Int'l Symposium on Cluster Computing and the Grid (CCGrid), May 2010.

Case Studies

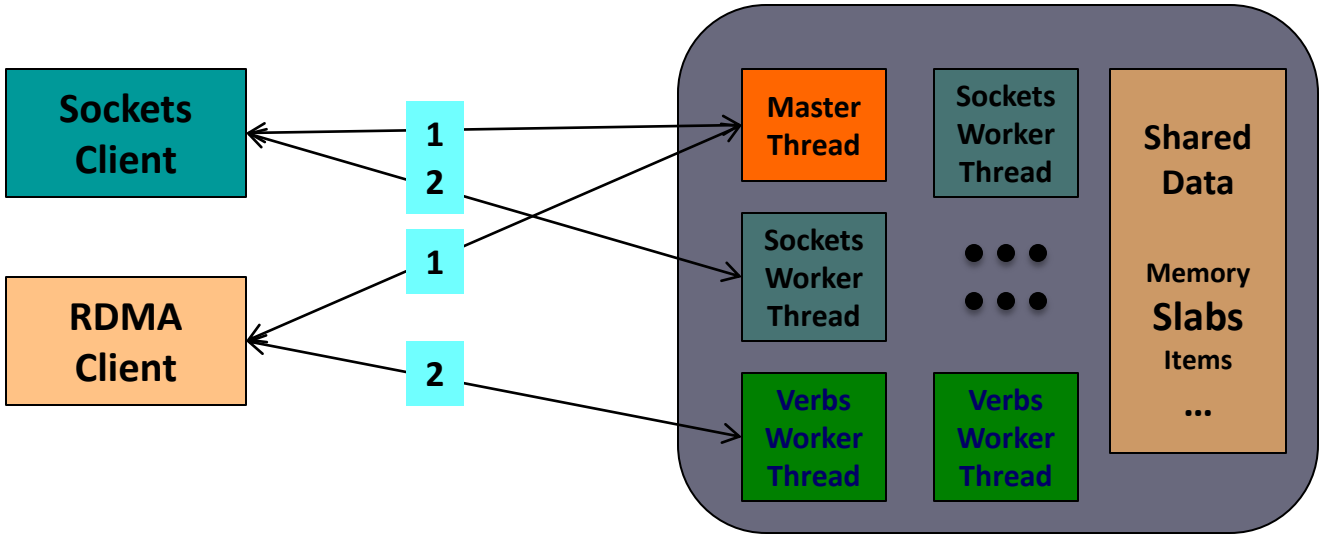
- Low-level Network Performance
- Clusters with Message Passing Interface (MPI)
- Datacenters with Sockets Direct Protocol (SDP) and TCP/IP (IPoIB)
- InfiniBand in WAN and Grid-FTP
- **Cloud Computing: Hadoop and Memcached**

A New Approach towards OFA in Cloud



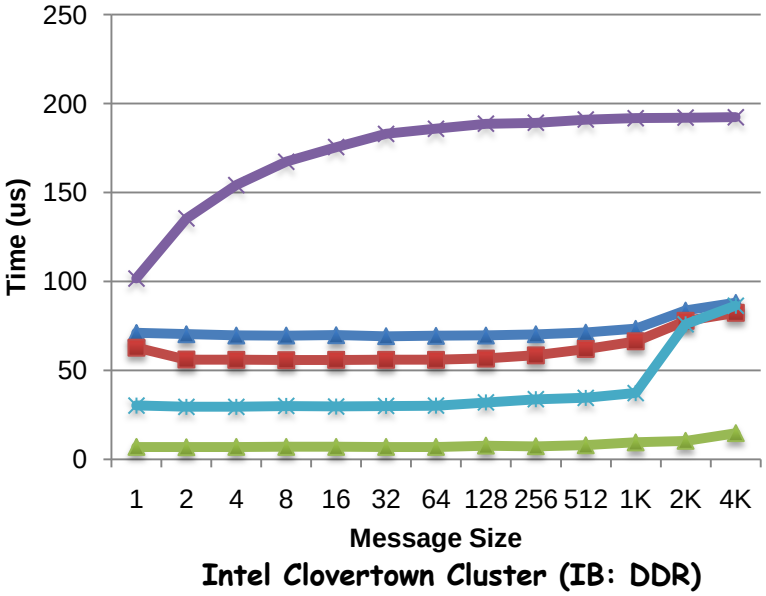
- Sockets not designed for high-performance
  - Stream semantics often mismatch for upper layers (Memcached, Hadoop)
  - Zero-copy not available for non-blocking sockets (Memcached)
- Significant consolidation in cloud system software
  - Hadoop and Memcached are developer facing APIs, not sockets
  - Improving Hadoop and Memcached will benefit many applications immediately!

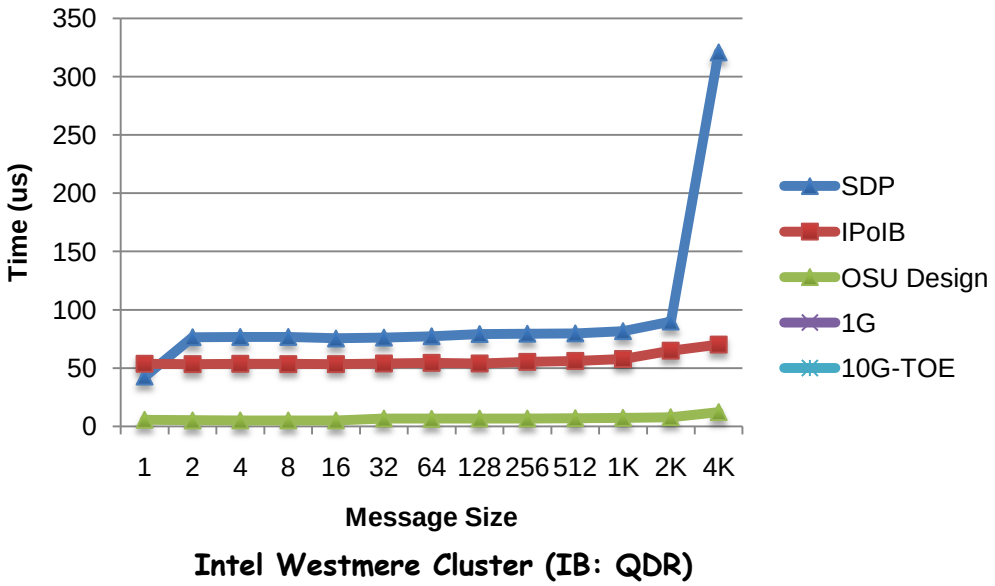
Memcached Design Using Verbs



- Server and client perform a negotiation protocol
  - Master thread assigns clients to appropriate worker thread
- Once a client is assigned a verbs worker thread, it can communicate directly and is “bound” to that thread
- All other Memcached data structures are shared among RDMA and Sockets worker threads

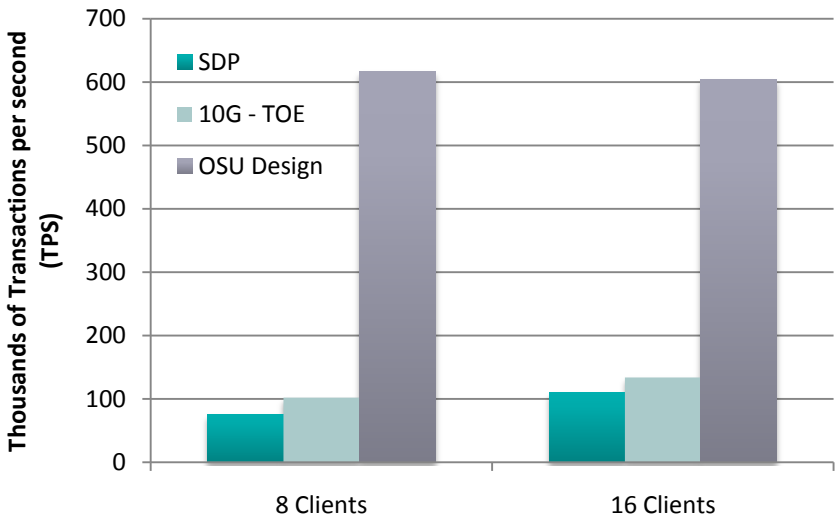
Memcached Get Latency



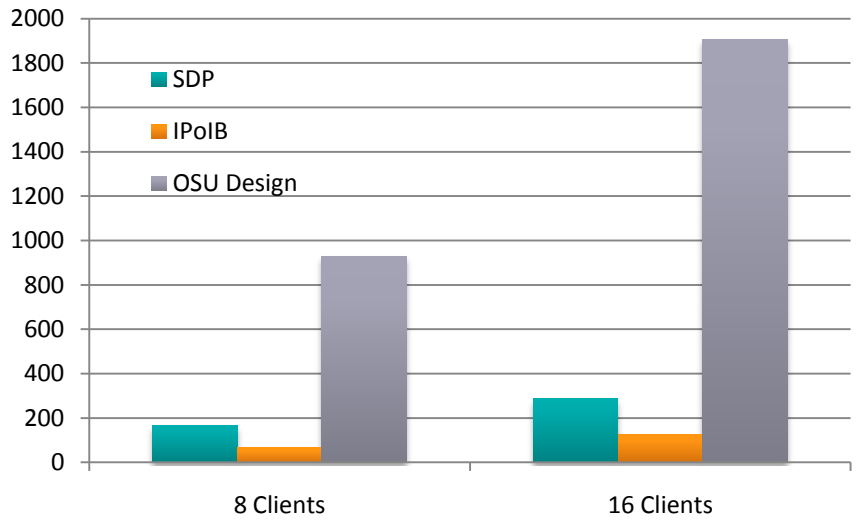


- Memcached Get latency
  - 4 bytes – DDR: 6 us; QDR: 5 us
  - 4K bytes -- DDR: 20 us; QDR:12 us
- Almost factor of *four* improvement over 10GE (TOE) for 4K
- We are in the process of evaluating iWARP on 10GE

Memcached Get TPS



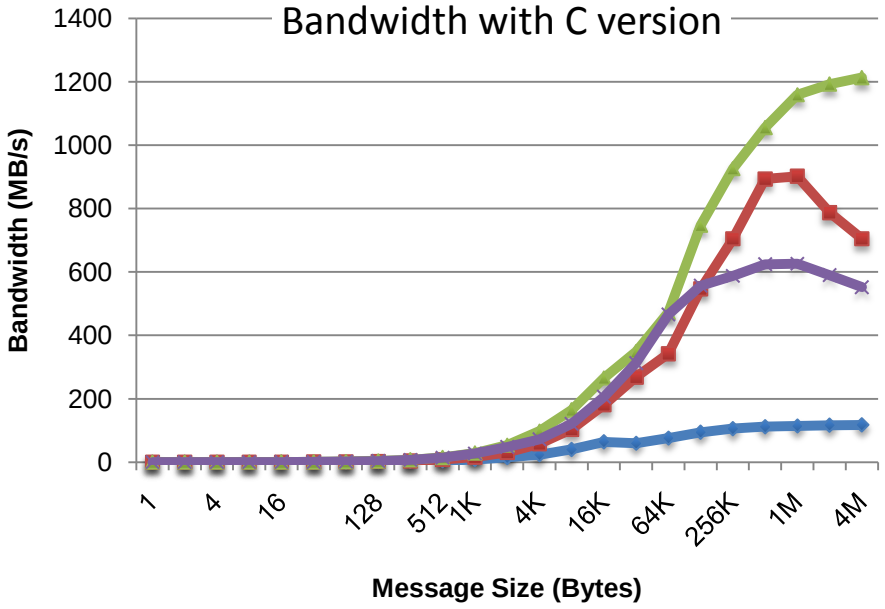
Intel Clovertown Cluster (IB: DDR)

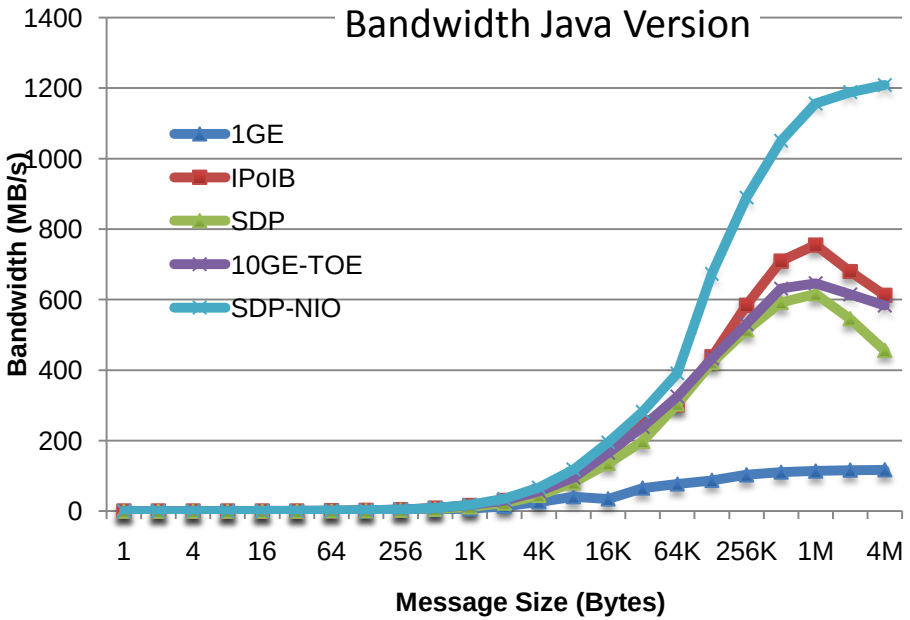


Intel Westmere Cluster (IB: QDR)

- Memcached Get transactions per second for 4 bytes
  - On IB DDR about 600K/s for 16 clients
  - On IB QDR 1.9M/s for 16 clients
- Almost factor of *six* improvement over 10GE (TOE)
- We are in the process of evaluating iWARP on 10GE

Hadoop: Java Communication Benchmark

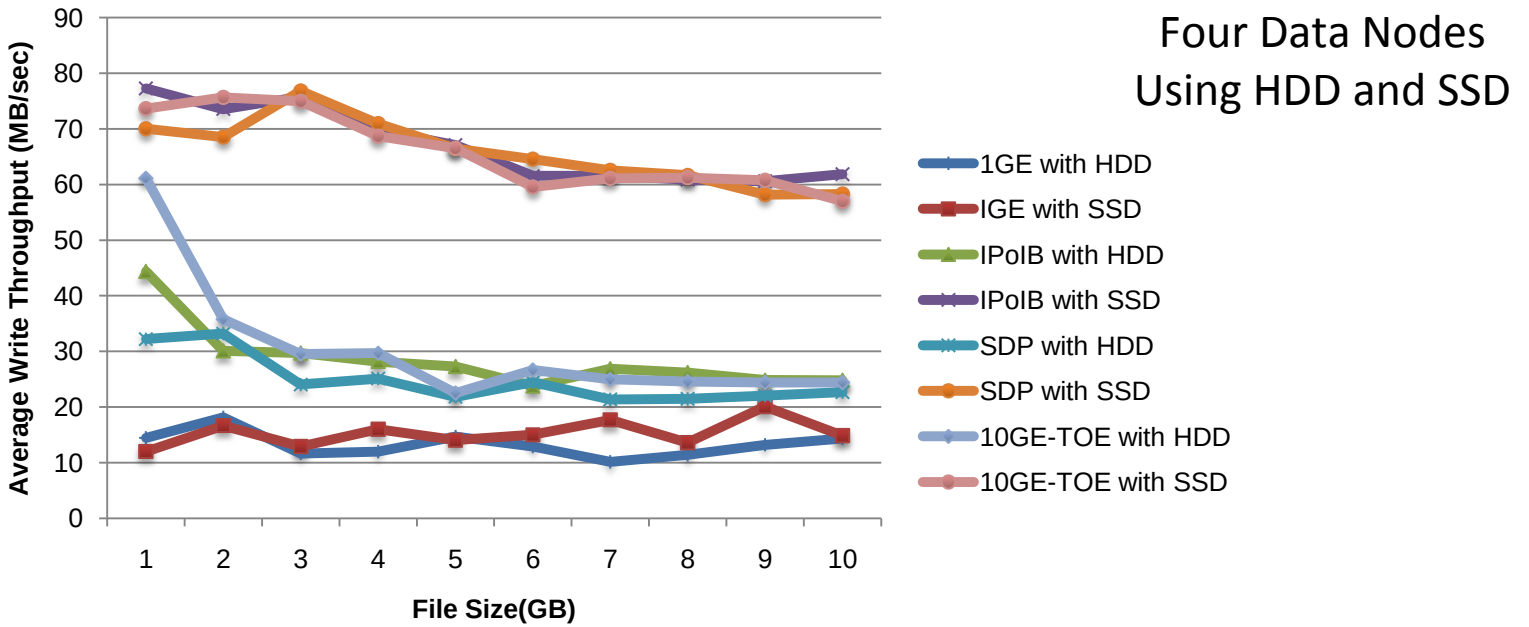




- Sockets level ping-pong bandwidth test
- Java performance depends on usage of NIO (allocateDirect)
- C and Java versions of the benchmark have similar performance
- HDFS does not use direct allocated blocks or NIO on DataNode

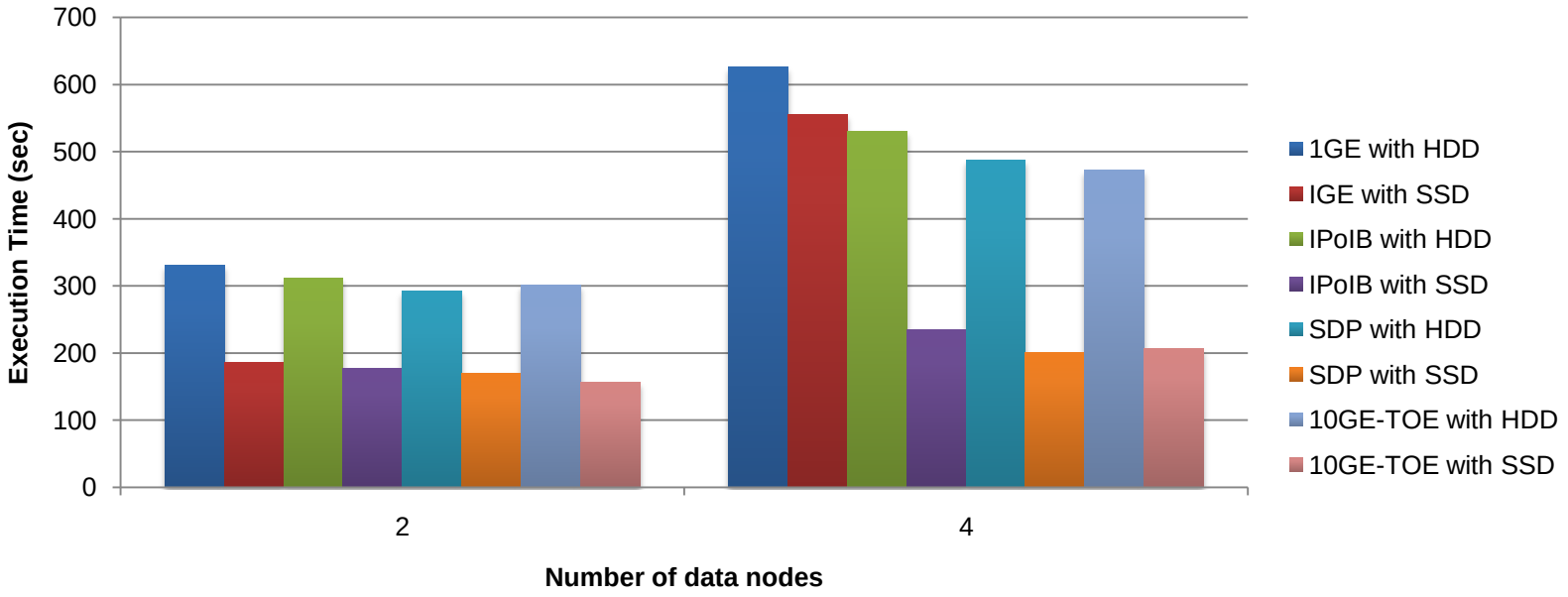
S. Sur, H. Wang, J. Huang, X. Ouyang and D. K. Panda *"Can High-Performance Interconnects Benefit Hadoop Distributed File System?"*, MASVDC '10 in conjunction with MICRO 2010, Atlanta, GA.

Hadoop: DFS IO Write Performance



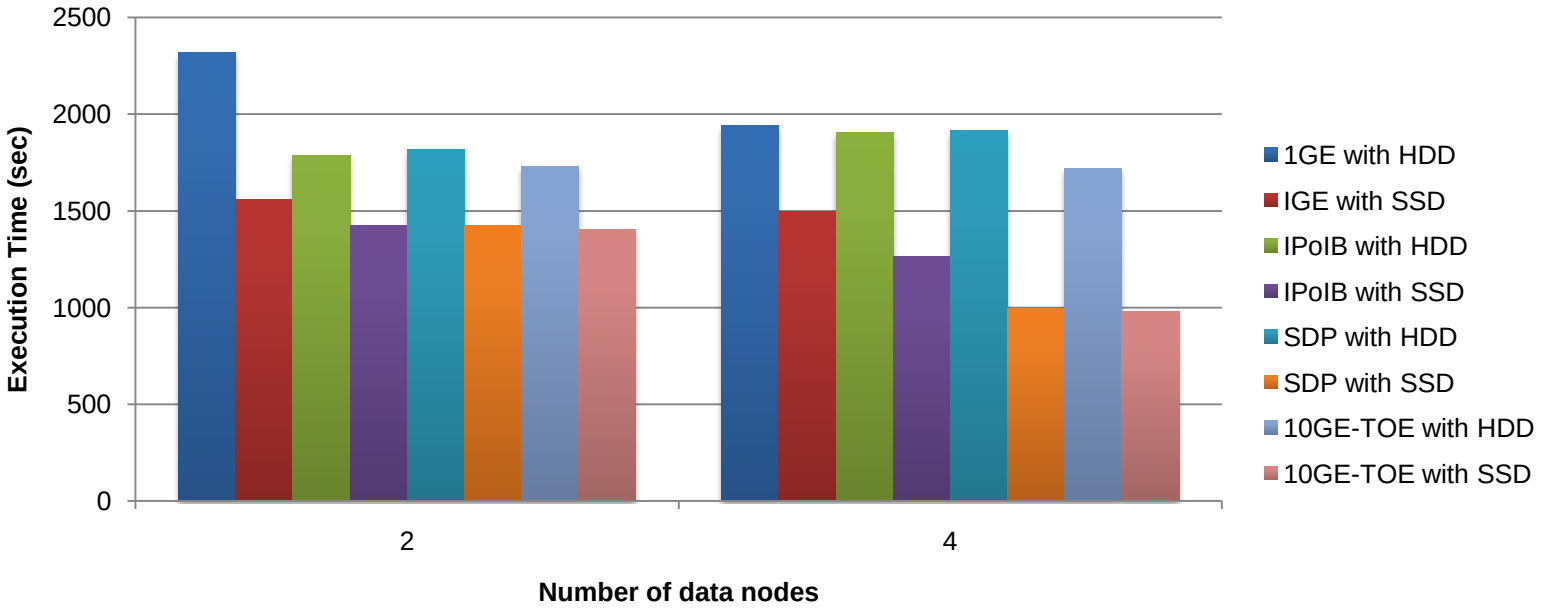
- DFS IO included in Hadoop, measures sequential access throughput
- We have two map tasks each writing to a file of increasing size (1-10GB)
- Significant improvement with IPoIB, SDP and 10GigE
- With SSD, performance improvement is almost seven or eight fold!
- SSD benefits not seen without using high-performance interconnect!
  - In-line with comment on Google keynote about I/O performance

Hadoop: RandomWriter Performance



- Each map generates 1GB of random binary data and writes to HDFS
- SSD improves execution time by 50% with 1GigE for two DataNodes
- For four DataNodes, benefits are observed only with HPC interconnect
- IPoIB, SDP and 10GigE can improve performance by 59% on four DataNodes

Hadoop Sort Benchmark



- Sort: baseline benchmark for Hadoop
- Sort phase: I/O bound; Reduce phase: communication bound
- SSD improves performance by 28% using 1GigE with two DataNodes
- Benefit of 50% on four DataNodes using SDP, IPoIB or 10GigE

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 for Clusters, Grid, Multi-tier Datacenters and Cloud Computing Systems
- 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. Dandapanthula (M.S.)
- R. Darbha (M.S.)
- V. Dhanraj (M.S.)
- J. Huang (Ph.D.)
- J. Jose (Ph.D.)
- K. Kandalla (Ph.D.)
- M. Luo (Ph.D.)
- V. Meshram (M.S.)
- X. Ouyang (Ph.D.)
- S. Pai (M.S.)
- S. Potluri (Ph.D.)
- R. Rajachandrasekhar (Ph.D.)
- A. Singh (Ph.D.)
- H. Subramoni (Ph.D.)

Past Students

- P. Balaji (Ph.D.)
- D. Buntinas (Ph.D.)
- S. Bhagvat (M.S.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (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 (Ph. D.)
- J. Liu (Ph.D.)
- A. Mamidala (Ph.D.)
- G. Marsh (M.S.)
- S. Naravula (Ph.D.)
- R. Noronha (Ph.D.)
- G. Santhanaraman (Ph.D.)
- J. Sridhar (M.S.)
- S. Sur (Ph.D.)
- K. Vaidyanathan (Ph.D.)
- A. Vishnu (Ph.D.)
- J. Wu (Ph.D.)
- W. Yu (Ph.D.)

Current Research Scientist

- S. Sur

Current Post-Docs

- H. Wang
- J. Vienne
- X. Besseron

Past Post-Docs

- E. Mancini
- S. Marcarelli
- H.-W. Jin

Current Programmers

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

Web Pointers

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

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

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

MVAPICH Web Page

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

panda@cse.ohio-state.edu

surs@cse.ohio-state.edu