Latest version of the slides can be obtained from
http://www.cse.ohio-state.edu/~panda/it4i-ib-hse.pdf

InfiniBand, Omni-Path, and High-speed Ethernet for Dummies

Tutorial at IT4 Innovations ’18
by

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

- Introduction

- Why InfiniBand and High-speed Ethernet?

- Overview of IB, HSE, their Convergence and Features

- Overview of Omni-Path Architecture

- IB, Omni-Path, 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)

Timeline

Number of Clusters

Percentage of Clusters

87%
Integrated High-End Computing Environments

Compute cluster
- LAN
- Frontend

Storage cluster
- LAN/WAN
- Meta-Data Manager
- Compute Node
- Compute Node
- Compute Node
- I/O Server Node
- I/O Server Node
- I/O Server Node

Enterprise Multi-tier Datacenter for Visualization and Mining
- Tier 1
  - Routers/Servers
  - Switch
  - Application Server

- Tier 2
  - Switch
  - Application Server
  - Database Server

- Tier 3
  - Switch
  - Database Server

Network Based Computing Laboratory
Cloud Computing Environments

LAN / WAN

Virtual Machine
Virtual Machine
Virtual Machine
Virtual Machine

Physical Machine

Virtual Network File System

Physical Meta-Data Manager

Physical I/O Server Node
Physical I/O Server Node
Physical I/O Server Node

Data
Data
Data
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

<table>
<thead>
<tr>
<th>I/O Interface</th>
<th>Year</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCI</td>
<td>1990</td>
<td>33MHz/32bit: 1.05Gbps (shared bidirectional)</td>
</tr>
<tr>
<td>PCI-X</td>
<td>1998 (v1.0)</td>
<td>133MHz/64bit: 8.5Gbps (shared bidirectional)</td>
</tr>
<tr>
<td></td>
<td>2003 (v2.0)</td>
<td>266-533MHz/64bit: 17Gbps (shared bidirectional)</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Network Type</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet (1979 - )</td>
<td>10 Mbit/sec</td>
</tr>
<tr>
<td>Fast Ethernet (1993 -)</td>
<td>100 Mbit/sec</td>
</tr>
<tr>
<td>Gigabit Ethernet (1995 -)</td>
<td>1000 Mbit /sec</td>
</tr>
<tr>
<td>ATM (1995 -)</td>
<td>155/622/1024 Mbit/sec</td>
</tr>
<tr>
<td>Myrinet (1993 -)</td>
<td>1 Gbit/sec</td>
</tr>
<tr>
<td>Fibre Channel (1994 -)</td>
<td>1 Gbit/sec</td>
</tr>
</tbody>
</table>
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

• Overview of Omni-Path Architecture

• IB, Omni-Path, 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
  - Several annexes released after that (RDMA_CM - Sep’06, iSER – Sep’06, XRC – Mar’09, RoCE – Apr’10, RoCEv2 – Sep’14, Virtualization – Nov’16)
  - Latest version 1.3.1 released November 2016
- [http://www.infinibandta.org](http://www.infinibandta.org)
High-speed Ethernet Consortium (10GE/25GE/40GE/50GE/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](http://www.ethernetalliance.org)
- 40-Gbps (Servers) and 100-Gbps Ethernet (Backbones, Switches, Routers): IEEE 802.3 WG
- 25-Gbps Ethernet Consortium targeting 25/50Gbps (July 2014)
  - [http://25gethernet.org](http://25gethernet.org)
- Energy-efficient and power-conscious protocols
  - On-the-fly link speed reduction for under-utilized links
- Ethernet Alliance Technology Forum looking forward to 2026
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

- 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

<table>
<thead>
<tr>
<th>Network Type</th>
<th>Speed (Mbit/sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet (1979 - )</td>
<td>10 Mbit/sec</td>
</tr>
<tr>
<td>Fast Ethernet (1993 - )</td>
<td>100 Mbit/sec</td>
</tr>
<tr>
<td>Gigabit Ethernet (1995 - )</td>
<td>1000 Mbit/sec</td>
</tr>
<tr>
<td>ATM (1995 - )</td>
<td>155/622/1024 Mbit/sec</td>
</tr>
<tr>
<td>Myrinet (1993 - )</td>
<td>1 Gbit/sec</td>
</tr>
<tr>
<td>Fibre Channel (1994 - )</td>
<td>1 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2001 - )</td>
<td>2 Gbit/sec (1X SDR)</td>
</tr>
<tr>
<td>10-Gigabit Ethernet (2001 - )</td>
<td>10 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2003 - )</td>
<td>8 Gbit/sec (4X SDR)</td>
</tr>
<tr>
<td>InfiniBand (2005 - )</td>
<td>16 Gbit/sec (4X DDR)</td>
</tr>
<tr>
<td>InfiniBand (2007 - )</td>
<td>24 Gbit/sec (12X SDR)</td>
</tr>
<tr>
<td>40-Gigabit Ethernet (2010 - )</td>
<td>32 Gbit/sec (4X QDR)</td>
</tr>
<tr>
<td>InfiniBand (2011 - )</td>
<td>40 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2012 - )</td>
<td>54.6 Gbit/sec (4X FDR)</td>
</tr>
<tr>
<td>InfiniBand (2014 - )</td>
<td>2 x 54.6 Gbit/sec (4X Dual-FDR)</td>
</tr>
<tr>
<td>25-/50-Gigabit Ethernet (2014 -)</td>
<td>25/50 Gbit/sec</td>
</tr>
<tr>
<td>100-Gigabit Ethernet (2015 -)</td>
<td>100 Gbit/sec</td>
</tr>
<tr>
<td>Omni-Path (2015 - )</td>
<td>100 Gbit/sec</td>
</tr>
<tr>
<td>InfiniBand (2015 - )</td>
<td>100 Gbit/sec (4X EDR)</td>
</tr>
<tr>
<td>InfiniBand (2016 - )</td>
<td>200 Gbit/sec (4X HDR)</td>
</tr>
</tbody>
</table>

---

100 times in the last 15 years
InfiniBand Link Speed Standardization Roadmap

- XDR = eXtreme Data Rate
- NDR = Next Data Rate
- HDR = High Data Rate
- EDR = Enhanced Data Rate
- FDR = Fourteen Data Rate
- QDR = Quad Data Rate
- DDR = Double Data Rate *(not shown)*
- 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 \(\rightarrow\) 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 \(\rightarrow\) 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
  – Support acceleration features (like IB) for Ethernet

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)

<table>
<thead>
<tr>
<th></th>
<th>Year(s)</th>
<th>Bandwidth (Speeds)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PCI</strong></td>
<td>1990</td>
<td>33MHz/32bit: 1.05Gbps (shared bidirectional)</td>
</tr>
<tr>
<td><strong>PCI-X</strong></td>
<td>1998 (v1.0), 2003 (v2.0)</td>
<td>133MHz/64bit: 8.5Gbps (shared bidirectional)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>266-533MHz/64bit: 17Gbps (shared bidirectional)</td>
</tr>
<tr>
<td><strong>AMD HyperTransport (HT)</strong></td>
<td>2001 (v1.0), 2004 (v2.0), 2006 (v3.0), 2008 (v3.1)</td>
<td>102.4Gbps (v1.0), 179.2Gbps (v2.0)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>332.8Gbps (v3.0), 409.6Gbps (v3.1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(32 lanes)</td>
</tr>
<tr>
<td><strong>PCI-Express (PCIe)</strong> by Intel</td>
<td>2003 (Gen1), 2007 (Gen2), 2009 (Gen3 standard), 2017 (Gen4 standard)</td>
<td>Gen1: 4X (8Gbps), 8X (16Gbps), 16X (32Gbps)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gen2: 4X (16Gbps), 8X (32Gbps), 16X (64Gbps)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gen3: 4X (~32Gbps), 8X (~64Gbps), 16X (~128Gbps)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gen4: 4X (~64Gbps), 8X (~128Gbps), 16X (~256Gbps)</td>
</tr>
<tr>
<td><strong>Intel QuickPath Interconnect (QPI)</strong></td>
<td>2009</td>
<td>153.6-204.8Gbps (20 lanes)</td>
</tr>
</tbody>
</table>
Upcoming I/O Interface Architectures

- Cache Coherence Interconnect for Accelerators (CCIX)
  - https://www.ccixconsortium.com/
- NVLink
- CAPI/OpenCAPI
  - http://opencapi.org/
- GenZ
  - http://genzconsortium.org/
Presentation Overview

- Introduction
- Why InfiniBand and High-speed Ethernet?
  - Overview of IB, HSE, their Convergence and Features
  - Overview of Omni-Path Architecture
  - IB, Omni-Path, 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

**Application Layer**
- MPI, PGAS, File Systems
- OpenFabrics Verbs

**Transport Layer**
- OpenSM (management tool)
- RC (reliable), UD (unreliable)

**Network Layer**
- DNS management tools
- Routing

**Link Layer**
- Flow-control, Error Detection
- Copper, Optical or Wireless

**Physical Layer**
- InfiniBand
- Copper or Optical

**Traditional Ethernet**
- HTTP, FTP, MPI, File Systems
- Sockets Interface
- TCP, UDP
- Routing
- Flow-control and Error Detection

**InfiniBand**
- Physical Layer
TCP/IP Stack and IPoIB

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Ethernet Driver

IPoIB

InfiniBand Adapter

Ethernet Adapter

InfiniBand Switch

IPoIB

1/10/25/40/50/100 GigE

Switch

Sockets

Application / Middleware
TCP/IP, IPoIB and Native IB Verbs

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

Sockets

Verbs

RDMA

User Space

Application / Middleware

Ethernet Adapter

InfiniBand Adapter

InfiniBand Switch

1/10/25/40/50/100 GigE

IB Native

RDMA

InfiniBand Adapter

InfiniBand Switch

IPoIB

Ethernet Switch

IPoIB

Ethernet Adapter

IPoIB

IPoIB

IPoIB

TCP/IP

Ethernet Adapter

IPoIB

Ethernet Switch

IPoIB

TCP/IP

Ethernet Adapter

IPoIB

1/10/25/40/50/100 GigE
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
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)
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
IB Communication Model

Basic InfiniBand Communication Semantics
Two-sided Communication Model

HCA  P 1

HCA  P 2

HCA  P 3

Recv from P1

Post Recv Buffer

Poll HCA

No Data

Recv from P3

Post Recv Buffer

Poll HCA

Send to P2

HCA Send

Data to P2

HCA Send

Data to P2

Recv Data from P3

Poll HCA

Post Send Buffer

Recv Data from P1

Recv Data from P1
One-sided Communication Model

Global Region Creation
(Buffer Info Exchanged)

- HCA Write Data to P2
- HCA Write Data to P3
- Write to P3
- Write Data from P1
- Write to P2
- Post to HCA
- Post to HCA
- Buffer at P1
- Buffer at P2
- Buffer at P3
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

`l_key` is needed for RDMA operations

`r_key` is needed for RDMA operations
Communication in the Channel Semantics (Send/Receive Model)

Processor is involved only to:

1. Post receive WQE
2. Post send WQE
3. Pull out completed CQEs from the CQ

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

Hardware ACK
Initiator processor is involved only to:
1. Post send WQE
2. Pull out completed CQE from the send CQ
No involvement from the target processor

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

Initiator processor is involved only to:
1. Post send WQE
2. Pull out completed CQE from the send CQ

No involvement from the target processor
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
Hardware Protocol Offload

Complete Hardware Implementations Exist
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, Service Levels, and QoS

• Virtual Lanes (VL)
  – Multiple (between 2 and 16) virtual links within same physical link
    • 0 – default data VL; 15 – VL for management traffic
  – Separate buffers and flow control
  – Avoids Head-of-Line Blocking

• Service Level (SL):
  – Packets may operate at one of 16, user defined SLs

• 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
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*
• Someone has to setup the forwarding tables and give every port an LID
  – “Subnet Manager” does this work
• Different routing algorithms give different paths

Switching: IB supports Virtual Cut Through (VCT)
Routing: Unspecified by IB SPEC
  Up*/Down*, Shift are popular routing engines supported by OFED

• 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
    • 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

• Efforts underway to create InfiniBand connectivity around the world by Astar Computational Resource Centre and partner organizations[1]

Hardware Protocol Offload

Complete Hardware Implementations Exist
### IB Transport Types and Associated Trade-offs

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Reliable Connection</th>
<th>Reliable Datagram</th>
<th>Dynamic Connected</th>
<th>eXtended Reliable Connection</th>
<th>Unreliable Connection</th>
<th>Unreliable Datagram</th>
<th>Raw Datagram</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scalability (M processes, N nodes)</td>
<td>M²N QPs per HCA</td>
<td>M QPs per HCA</td>
<td>M QPs per HCA</td>
<td>MN QPs per HCA</td>
<td>M²N QPs per HCA</td>
<td>M QPs per HCA</td>
<td>1 QP per HCA</td>
</tr>
</tbody>
</table>

**Reliability**

<table>
<thead>
<tr>
<th></th>
<th>Corrupt data detected</th>
<th>Data Delivery Guarantee</th>
<th>Data Order Guarantees</th>
<th>Data Loss Detected</th>
<th>Error Recovery</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Data delivered exactly once</td>
<td>Per connection</td>
<td>Yes</td>
<td>Errors (retransmissions, alternate path, etc.) handled by transport layer. Client only involved in handling fatal errors (links broken, protection violation, etc.)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Unordered, duplicate data detected</td>
<td>One source to multiple destinations</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No guarantees</td>
<td>No connection</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>No</td>
<td>None</td>
</tr>
</tbody>
</table>

### Notes

- **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

- **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 & Transaction Ordering

• 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

• 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 & Rate 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
- IB allows link rates to be statically changed to fixed values
  - On a 4X link, we can set data to be sent at 1X
    - Cannot set rate requirement to 3.16 Gbps, for example
  - 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
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
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

Inactive Links

Compute Node

Inactive Link

Switch

Inactive Links

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

Traditional Model

- Sockets App
- Sockets API

Possible SDP Model

- Sockets Application
- Sockets API

Kernel

TCP/IP Sockets Provider

TCP/IP Transport Driver

IPoIB Driver

InfiniBand CA

SDP

OS Modules

InfiniBand Hardware

TCP/IP Sockets Provider

TCP/IP Transport Driver

IPoIB Driver

InfiniBand CA

Sockets Direct Protocol

Kernel

Bypass

RDMA

Semantics

(Source: InfiniBand Trade Association)
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
  – Alternate Vendor-specific Stacks
    • MX over Ethernet (for Myricom 10GE adapters)
    • Datagram Bypass Layer (for Myricom 10GE adapters)
    • Solarflare OpenOnload (for Solarflare 10/40GE adapters)
    • Emulex FastStack DBL (for OneConnect OCe12000-D 10GE adapters)
# IB and 10/40GE RDMA Models: Commonalities and Differences

<table>
<thead>
<tr>
<th>Features</th>
<th>IB</th>
<th>iWARP/HSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Acceleration</td>
<td>Supported</td>
<td>Supported</td>
</tr>
<tr>
<td>RDMA</td>
<td>Supported</td>
<td>Supported</td>
</tr>
<tr>
<td>Atomic Operations</td>
<td>Supported</td>
<td>Not supported</td>
</tr>
<tr>
<td>Multicast</td>
<td>Supported</td>
<td>Supported</td>
</tr>
<tr>
<td>Congestion Control</td>
<td>Supported</td>
<td>Supported</td>
</tr>
<tr>
<td>Data Placement</td>
<td>Ordered</td>
<td>Out-of-order</td>
</tr>
<tr>
<td>Data Rate-control</td>
<td>Static and Coarse-grained</td>
<td>Dynamic and Fine-grained</td>
</tr>
<tr>
<td>QoS</td>
<td>Prioritization</td>
<td>Prioritization and Fixed Bandwidth QoS</td>
</tr>
<tr>
<td>Multipathing</td>
<td>Using DLIDs</td>
<td>Using VLANs</td>
</tr>
</tbody>
</table>
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)
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
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
    - Alternate Vendor-specific Stacks
      - MX over Ethernet (for Myricom 10GE adapters)
      - Datagram Bypass Layer (for Myricom 10GE adapters)
      - Solarflare OpenOnload (for Solarflare 10/40GE adapters)
      - Emulex FastStack DBL (for OneConnect OCe12000-D 10GE adapters)
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”
iWARP and TOE

Application / Middleware Interface

Protocol

Kernel Space

TCP/IP

IPoIB

TCP/IP

Hardware Offload

RSockets

SDP

TCP/IP

RDMA

RDMA

User Space

User Space

Application / Middleware

Sockets

Verbs

Adapter

Ethernet Adapter

InfiniBand Adapter

Ethernet Adapter

InfiniBand Adapter

InfiniBand Adapter

InfiniBand Adapter

InfiniBand Adapter

Ethernet Switch

IPoIB

Ethernet Switch

30/40 GigE-TOE

InfiniBand Switch

RDMA

InfiniBand Switch

IB Native

Network Based Computing Laboratory

IT4 Innovations’18

83
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
  - Alternate Vendor-specific Stack
    - Datagram Bypass Layer (for Myricom 10GE adapters)
    - Solarflare OpenOnload (for Solarflare 10/40GE adapters)
    - Emulex FastStack DBL (for OneConnect OCe12000-D 10GE adapters)
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

[Diagram: Typical Commodity Networking Stack vs Typical HPC Networking Stack]

Courtesy Solarflare communications (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)

Network Stack Comparison

- Takes advantage of IB and Ethernet
  - Software written with IB-Verbs
  - Link layer is Converged (Enhanced) Ethernet (CE)
  - 100Gb/s support from latest EDR and ConnectX-3 Pro adapters

- Pros: IB Vs RoCE
  - Works natively in Ethernet environments
    - Entire Ethernet management ecosystem is available
  - Has all the benefits of IB verbs
  - Link layer is very similar to the link layer of native IB, so there are no missing features

- RoCE v2: Additional Benefits over RoCE
  - Traditional Network Management Tools Apply
  - ACLs (Metering, Accounting, Firewalling)
  - GMP Snooping for Optimized Multicast
  - Network Monitoring Tools

Packet Header Comparison

- ETH L2 Hdr
- IP Hdr
- UDP Hdr
- IB BTH+ L4 Hdr

APPLICATION

IB Verbs

IB Transport

IB Network

InfiniBand Link Layer

RoCE

Application

IB Verbs

IB Transport

IB Network

Ethernet Link Layer

RoCE v2

Application

IB Verbs

IB Transport

UDP / IP

Ethernet Link Layer

ETH
L2 Hdr

IB GRH
L3 Hdr

IB BTH+
L4 Hdr

ETH
L2 Hdr

IP Hdr

UDP Hdr

IB BTH+
L4 Hdr

Courtesy: OFED, Mellanox
Presentation Overview

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

- Pathscale (2003 – 2006) came up with initial version of IB-based product
- QLogic enhanced the product with the PSM software interface
- The IB product of QLogic was acquired by Intel
- Intel enhanced the QLogic IB product to create the Omni-Path product
Omni-Path Fabric Overview

• **Layer 1.5: Link Transfer Protocol**
  – Features
    • Traffic Flow Optimization
    • Packet Integrity Protection
    • Dynamic Lane Switching
  – Error detection/replay occurs in Link Transfer Packet units
  – 1 Flit = 65b; LTP = 1056b = 16 flits + 14b CRC + 2b Credit
  – LTPs implicitly acknowledged
  – Retransmit request via NULL LTP; carries replay command flit

• **Layer 2: Link Layer**
  – Supports 24 bit fabric addresses
  – Allows 10KB of L4 payload; 10,368 byte max packet size
  – Congestion Management
    • Adaptive / Dispersive Routing
    • Explicit Congestion Notification
  – QoS support
    • Traffic Class, Service Level, Service Channel and Virtual Lane

• **Layer 3: Data Link Layer**
  – Fabric addressing, switching, resource allocation and partitioning support
### IB, Omni-Path, and HSE: Feature Comparison

<table>
<thead>
<tr>
<th>Features</th>
<th>IB</th>
<th>iWARP/HSE</th>
<th>RoCE</th>
<th>RoCE v2</th>
<th>Omni-Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Acceleration</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RDMA</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Congestion Control</td>
<td>Yes</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multipathning</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Atomic Operations</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multicast</td>
<td>Optional</td>
<td>No</td>
<td>Optional</td>
<td>Optional</td>
<td>Optional</td>
</tr>
<tr>
<td>Data Placement</td>
<td>Ordered</td>
<td>Out-of-order</td>
<td>Ordered</td>
<td>Ordered</td>
<td>Ordered</td>
</tr>
<tr>
<td>Prioritization</td>
<td>Optional</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Fixed BW QoS (ETS)</td>
<td>No</td>
<td>Optional</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ethernet Compatibility</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TCP/IP Compatibility</td>
<td>Yes (using IPoIB)</td>
<td>Yes (using IPoIB)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Presentation Overview

- Introduction
- Why InfiniBand and High-speed Ethernet?
- Overview of IB, HSE, their Convergence and Features
- Overview of Omni-Path Architecture
  - IB, Omni-Path, 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 (acquired by Mellanox) and QLogic (acquired by Intel)
  – Aligned with many server vendors: Intel, IBM, Oracle, Dell
  – And many integrators: Appro, Advanced Clustering, Microway
• New vendors like Oracle are entering the market with IB products
• Broadly two kinds of adapters
  – Offloading (Mellanox) and Onloading (Intel TrueScale / 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), EDR (100 Gbps), HDR (200 Gbps)
• ConnectX-2, ConnectX-3, ConnectIB, Connectx-3 Pro, ConnectX-4, ConnectX-5, ConnectX-6 adapters from Mellanox supports offload for collectives (Barrier, Broadcast, etc.) and offload for tag matching
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 (Intel)
    - Up to 846-port QDR switch by QLogic
  - FDR (54.6 Gbps) switch silicon (Bridge-X) and associated switches (18-648 ports)
  - EDR (100Gbps) switch from Oracle and Mellanox
  - Switch-X-2 silicon from Mellanox with VPI and SDN (Software Defined Networking) support announced in Oct ’12
  - SwitchIB-2 from Mellanox  EDR 100Gb/s, offloads MPI communications, announced Nov'15
  - Quantum from Mellanox  HDR 200Gb/s, offloads MPI communications, announced Nov'16

- **Switch Routers with Gateways**
  - IB-to-FC; IB-to-IP
10G, 25G, 40G, 50G, 56G and 100G Ethernet Products

• 10GE adapters
  – Intel, Intilop, Myricom, Emulex, Mellanox (ConnectX, ConnectX-4 Lx EN), Solarflare (Flareon)
• 10GE/iWARP adapters
  – Chelsio, NetEffect (now owned by Intel)
• 25GE adapters
  – Mellanox ConnectX-4 Lx EN
• 40GE adapters
  – Mellanox ConnectX3-EN 40G, Mellanox ConnectX-4 Lx EN
  – Chelsio (T5 2x40 GigE), Solarflare (Flareon)
• 50GE adapters
  – Mellanox ConnectX-4 Lx EN
• 100GE adapters
  – FPGA-based 100GE adapter from inveaTECH
  – FPGA based Dual-port 100GE adapter from Accolade Technology (ANIC-200K)
  – ConnectX-4 EN single/dual-port 100GE adapter from Mellanox

• 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
• 25GE, 40GE, 50GE, 56GE and 100GE switches
  – Mellanox SN2410, SN2100, and SN2700 supports 10/25/40/50/56/100 GE
  – Gnodal, Arista, Brocade, Cisco, Juniper, Huawei and Mellanox 40GE (SX series)
  – Arista 7504R, 7508R, 7512R supports 10/25/40/100 GE
  – Broadcom has switch architectures for 10/40/100GE
    – Trident, Trident2, Tomahawk and, Tomahawk2
  – Nortel Networks - 10GE downlinks with 40GE and 100GE uplinks
  – Mellanox – Spectrum 25/100 Gigabit Open Ethernet-based Switch
  – Atrica A-8800 provides 100 GE optical Ethernet

Price for different adapters and switches are available from: http://colfaxdirect.com
Omni-Path Products

- Intel Omni-Path Edge Switches 100 Series

- Intel Omni-Path Director Class Switches 100 Series

- Intel Omni-Path Host Fabric Interface
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), 54.6 Gbps (FDR) and 100 Gbps (EDR) rates available for IB
  – 10GE, 40GE, 56GE (only with Mellanox Switches) and 100GE rates available for RoCE
Software Convergence with OpenFabrics

- Open source organization (formerly OpenIB)
  - www.openfabrics.org

- Incorporates both IB, RoCE, and iWARP in a unified manner
  - Support for Linux and Windows

- Users can download the entire stack and run
  - Latest stable release is OFED 4.8.1
    - New naming convention to get aligned with Linux Kernel Development
    - OFED 4.8.2 is under development
OpenFabrics Stack with Unified Verbs Interface

Verbs Interface (libibverbs)

User Level
- Mellanox (libmthca, libmlx*)
- Intel/QLogic (libipathverbs)
- IBM (libehca)
- Chelsio (libcxgb*)
- Emulex (libocrdma)
- Intel/NetEffect (libnes)

Kernel Level
- Mellanox (ib_mthca, ib_ml*)
- Intel/QLogic (ib_ipath)
- IBM (ib_ehca)
- Chelsio (ib_cxgb*)
- Emulex
- Intel/NetEffect

Mellanox Adapters
Intel/QLogic Adapters
IBM Adapters
Chelsio Adapters
Emulex Adapters
Intel/NetEffect Adapters
Libfabrics Software Stack

Open Fabrics Interface (OFI)
- Control Services
  - Discovery
- Communication Services
  - Connection Management
  - Address Vectors
- Completion Services
  - Event Queues
  - Counters
- Data Transfer Services
  - Message Queues
  - RMA
  - Tag Matching
  - Atomics
- Triggered Operations

OFI Provider
- Discovery
- Connection Management
- Address Vectors
- Event Queues
- Counters
- Message Queues
- RMA
- Tag Matching
- Atomics
- Triggered Operations

NIC
- TX Command Queues
- RX Command Queues

Trends of Networking Technologies in TOP500 Systems

Interconnect Family – Systems Share
InfiniBand in the Top500 (November 2017)
Large-scale InfiniBand Installations

- 163 IB Clusters (32.6%) in the Nov’17 Top500 list
  - (http://www.top500.org)
- Installations in the Top 50 (17 systems):

<table>
<thead>
<tr>
<th>Large-scale InfiniBand Installations</th>
<th>Large-scale InfiniBand Installations</th>
</tr>
</thead>
<tbody>
<tr>
<td>19,860,000 core (Gyoukou) in Japan (4th)</td>
<td>60,512 core (DGX SATURN V) at NVIDIA/USA (36th)</td>
</tr>
<tr>
<td>241,108 core (Pleiades) at NASA/Ames (17th)</td>
<td>72,000 core (HPC2) in Italy (37th)</td>
</tr>
<tr>
<td>220,800 core (Pangea) in France (21st)</td>
<td>152,692 core (Thunder) at AFRL/USA (40th)</td>
</tr>
<tr>
<td>144,900 core (Cheyenne) at NCAR/USA (24th)</td>
<td>99,072 core (Mistral) at DKRZ/Germany (42nd)</td>
</tr>
<tr>
<td>155,150 core (Jureca) in Germany (29th)</td>
<td>147,456 core (SuperMUC) in Germany (44th)</td>
</tr>
<tr>
<td>72,800 core Cray CS-Storm in US (30th)</td>
<td>86,016 core (SuperMUC Phase 2) in Germany (45th)</td>
</tr>
<tr>
<td>72,800 core Cray CS-Storm in US (31st)</td>
<td>74,520 core (Tsubame 2.5) at Japan/GSIC (48th)</td>
</tr>
<tr>
<td>78,336 core (Electra) at NASA/USA (33rd)</td>
<td>66,000 core (HPC3) in Italy (51st)</td>
</tr>
<tr>
<td>124,200 core (Topaz) SGI ICE at ERDC DSRC in US (34th)</td>
<td>194,616 core (Cascade) at PNNL (53rd)</td>
</tr>
<tr>
<td>60,512 core (NVIDIA DGX-1/Relion) at Facebook in USA (35th)</td>
<td>and many more!</td>
</tr>
</tbody>
</table>
### Large-scale Omni-Path Installations

- 35 Omni-Path Clusters (7%) in the Nov’17 Top500 list
  - [http://www.top500.org](http://www.top500.org)

<table>
<thead>
<tr>
<th>Core Count</th>
<th>Location/Name</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>556,104</td>
<td>Oakforest-PACS at JCAHPC in Japan</td>
<td>9th</td>
</tr>
<tr>
<td>368,928</td>
<td>Stampede2 at TACC in USA</td>
<td>12th</td>
</tr>
<tr>
<td>135,828</td>
<td>Tsubame 3.0 at TiTech in Japan</td>
<td>13th</td>
</tr>
<tr>
<td>314,384</td>
<td>Marconi XeonPhi at CINECA in Italy</td>
<td>14th</td>
</tr>
<tr>
<td>153,216</td>
<td>MareNostrum at BSC in Spain</td>
<td>16th</td>
</tr>
<tr>
<td>95,472</td>
<td>Quartz at LLNL in USA</td>
<td>49th</td>
</tr>
<tr>
<td>95,472</td>
<td>Jade at LLNL in USA</td>
<td>50th</td>
</tr>
<tr>
<td>49,432</td>
<td>Mogon II at Universitaet Mainz in Germany</td>
<td>65th</td>
</tr>
<tr>
<td>38,552</td>
<td>Molecular Simulator in Japan</td>
<td>70th</td>
</tr>
<tr>
<td>35,280</td>
<td>Quriosity at BASF in Germany</td>
<td>71st</td>
</tr>
<tr>
<td>554,432</td>
<td>Marconi Xeon at CINECA in Italy</td>
<td>72nd</td>
</tr>
<tr>
<td>46,464</td>
<td>Peta4 at University of Cambridge in UK</td>
<td>75th</td>
</tr>
<tr>
<td>53,352</td>
<td>Girzzly at LANL in USA</td>
<td>85th</td>
</tr>
<tr>
<td>45,680</td>
<td>Endeavor at Intel in USA</td>
<td>86th</td>
</tr>
<tr>
<td>59,776</td>
<td>Cedar at SFU in Canada</td>
<td>94th</td>
</tr>
<tr>
<td>27,200</td>
<td>Peta HPC in Taiwan</td>
<td>95th</td>
</tr>
<tr>
<td>40,392</td>
<td>Serrano at SNL in USA</td>
<td>112th</td>
</tr>
<tr>
<td>39,774</td>
<td>Nel at LLNL in USA</td>
<td>101st</td>
</tr>
<tr>
<td>40,392</td>
<td>Cayenne at SNL in USA</td>
<td>113th</td>
</tr>
</tbody>
</table>

**and many more!**
HSE Scientific Computing Installations

- 204 HSE compute systems with ranking in the Nov’17 Top500 list
  - 39,680-core installation in China (#73)
  - 66,560-core installation in China (#101) – new
  - 66,280-core installation in China (#103) – new
  - 64,000-core installation in China (#104) – new
  - 64,000-core installation in China (#105) – new
  - 72,000-core installation in China (#108) – new
  - 78,000-core installation in China (#125)
  - 59,520-core installation in China (#128) – new
  - 59,520-core installation in China (#129) – new
  - 64,800-core installation in China (#130) – new
  - 67,200-core installation in China (#134) – new
  - 57,600-core installation in China (#135) – new
  - 57,600-core installation in China (#136) – new
  - 64,000-core installation in China (#138) – new
  - 84,000-core installation in China (#139)
  - 84,000-core installation in China (#140)
  - 51,840-core installation in China (#151) – new
  - 51,200-core installation in China (#156) – new
  - and many more!
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 has installed a 100GE infrastructure for US DOE and have recently expanded it across the Atlantic to Europe

- They also have a 100G SDN overlay network from LBL to StarLight

https://www.es.net/network-r-and-d/experimental-network-testbeds/100ge-sdn-testbed/
Presentation Overview

- Introduction
- Why InfiniBand and High-speed Ethernet?
- Overview of IB, HSE, their Convergence and Features
- Overview of Omni-Path Architecture
- IB, Omni-Path, 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

Small Messages

ConnectX-4 EDR (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back
ConnectX-4 EN (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back

Large Messages

IB-EDR (100Gbps)
RoCE (100Gbps)
Low-level Uni-directional Bandwidth Measurements

ConnectX-4 EDR (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back
ConnectX-4 EN (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back
Low-level Latency Measurements

Small Messages

Message Size (bytes)

Latency (us)

ConnectX-4 EDR (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back

Large Messages

Message Size (bytes)

Latency (us)
Low-level Uni-directional Bandwidth Measurements

ConnectX-4 EDR (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back
Case Studies

- Low-level Performance
- Message Passing Interface (MPI)
Overview of the MVAPICH2 Project

- High Performance open-source MPI Library for InfiniBand, Omni-Path, Ethernet/iWARP, and RDMA over Converged Ethernet (RoCE)
  - MVAPICH (MPI-1), MVAPICH2 (MPI-2.2 and MPI-3.0), Started in 2001, First version available in 2002
  - MVAPICH2-X (MPI + PGAS), Available since 2011
  - Support for GPGPUs (MVAPICH2-GDR) and MIC (MVAPICH2-MIC), Available since 2014
  - Support for Virtualization (MVAPICH2-Virt), Available since 2015
  - Support for Energy-Awareness (MVAPICH2-EA), Available since 2015
  - Support for InfiniBand Network Analysis and Monitoring (OSU INAM) since 2015

- Used by more than 2,850 organizations in 85 countries
- More than 440,000 (> 0.44 million) downloads from the OSU site directly
- Empowering many TOP500 clusters (Nov ’17 ranking)
  - 1st, 10,649,600-core (Sunway TaihuLight) at National Supercomputing Center in Wuxi, China
  - 12th, 368,928-core (Stampede2) at TACC
  - 17th, 241,108-core (Pleiades) at NASA
  - 48th, 76,032-core (Tsubame 2.5) at Tokyo Institute of Technology
- Available with software stacks of many vendors and Linux Distros (RedHat and SuSE)
- http://mvapich.cse.ohio-state.edu

- Empowering Top500 systems for over a decade
  - System-X from Virginia Tech (3rd in Nov 2003, 2,200 processors, 12.25 TFlops) ->
  - Sunway TaihuLight (1st in Jun’17, 10M cores, 100 PFllops)
One-way Latency: MPI over IB with MVAPICH2

**Small Message Latency**

<table>
<thead>
<tr>
<th>Message Size (bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>1.19</td>
</tr>
<tr>
<td></td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>1.04</td>
</tr>
</tbody>
</table>

**Large Message Latency**

<table>
<thead>
<tr>
<th>Message Size (bytes)</th>
<th>Latency (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2K</td>
<td>1.01</td>
</tr>
<tr>
<td>4K</td>
<td>1.04</td>
</tr>
<tr>
<td>8K</td>
<td>1.04</td>
</tr>
<tr>
<td>16K</td>
<td>1.04</td>
</tr>
<tr>
<td>32K</td>
<td>1.04</td>
</tr>
<tr>
<td>64K</td>
<td>1.04</td>
</tr>
<tr>
<td>128K</td>
<td>1.04</td>
</tr>
<tr>
<td>256K</td>
<td>1.04</td>
</tr>
</tbody>
</table>

*TrueScale-QDR* - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
*ConnectX-3-FDR* - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
*ConnectIB-DualFDR* - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
*ConnectX-4-EDR* - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB Switch
*Omni-Path* - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch
Bandwidth: MPI over IB with MVAPICH2

Unidirectional Bandwidth

![Graph showing unidirectional bandwidth with message size and bandwidth values for different network types.]

Bidirectional Bandwidth

![Graph showing bidirectional bandwidth with message size and bandwidth values for different network types.]

- TrueScale-QDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
- ConnectX-3-FDR - 2.8 GHz Deca-core (IvyBridge) Intel PCI Gen3 with IB switch
- ConnectIB-Dual FDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with IB switch
- ConnectX-4-EDR - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 IB switch
- Omni-Path - 3.1 GHz Deca-core (Haswell) Intel PCI Gen3 with Omni-Path switch
One-way Latency: MPI over iWARP

- Chelsio T4 (TCP/IP)
- Chelsio T4 (iWARP)
- Intel-NetEffect NE20 (TCP/IP)
- Intel-NetEffect NE20 (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

Unidirectional Bandwidth

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-4 EDR (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back
ConnectX-4 EN (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back
Convergent Technologies: 
MPI Uni- and Bi-directional Bandwidth

ConnectX-4 EDR (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back
ConnectX-4 EN (100 Gbps): 3.1 GHz Deca-core (Haswell) Intel Back-to-back
Presentation Overview

- Introduction
- Why InfiniBand and High-speed Ethernet?
- Overview of IB, HSE, their Convergence and Features
- Overview of Omni-Path Architecture
- IB, Omni-Path, 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, Omni-Path and HSE
  - Highlighted the main features of IB and HSE and their convergence
  - Gave an overview of IB, Omni-Path, and HSE hardware/software products
  - Discussed sample performance numbers in designing various high-end systems with IB, Omni-Path, and HSE
- IB, Omni-Path, 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**

Office of Science
NSF
Mellanox Technologies
Microsoft
QLogic
CRAY
intel
LINUX NETWORK
NVIDIA
NetApp
Sun Microsystems
Cisco Systems

**Equipment Support by**

Mellanox Technologies
intel
AMD
Sun Microsystems
AFPR
advanced clustering technologies, inc.
Microway
NVIDIA
QLogic
Personnel Acknowledgments

Current Students (Graduate)
- A. Awan (Ph.D.)
- R. Biswas (M.S.)
- M. Bayatpour (Ph.D.)
- S. Chakraborty (Ph.D.)
- C.-H. Chu (Ph.D.)
- S. Guganani (Ph.D.)

Current Students (Undergraduate)
- J. Hashmi (Ph.D.)

Current Research Scientists
- X. Lu

Past Students
- A. Augustine (M.S.)
- P. Balaji (Ph.D.)
- S. Bhagvat (M.S.)
- A. Bhat (M.S.)
- D. Buntinas (Ph.D.)
- L. Chai (Ph.D.)
- B. Chandrasekharan (M.S.)
- N. Dandapanthula (M.S.)
- V. Dhanraj (M.S.)
- T. Gangadharappa (M.S.)
- K. Gopalakrishnan (M.S.)

Past Research Scientist
- K. Hamidouche
- S. Sur

Past Post-Docs
- D. Banerjee
- X. Besseron
- H.-W. Jin

Past Post-Docs
- J. Lin
- M. Luo
- E. Mancini

Past Programmers
- D. Bureddy
- J. Perkins

Current Students (Undergraduate)
- N. Sarkauskas (B.S.)

Current Research Scientists
- H. Subramoni

Current Post-doc
- A. Ruhela

Current Research Specialist
- J. Smith
- M. Arnold

Current Research Specialist
- J. Smith
- M. Arnold

Past Programmers
- D. Bureddy
- J. Perkins

Current Research Scientist
- J. Lin
- M. Luo
- E. Mancini

Past Post-Docs
- J. Vienne
- H. Wang
Thank You!

panda@cse.ohio-state.edu; subramon@cse.ohio-state.edu

Network-Based Computing Laboratory
http://nowlab.cse.ohio-state.edu/

The High-Performance MPI/PGAS Project
http://mvapich.cse.ohio-state.edu/

The High-Performance Deep Learning Project
http://hidl.cse.ohio-state.edu/