Presentation Overview

- Introduction

- Why InfiniBand and 10-Gigabit Ethernet?

- Overview of IB, 10GE, their Convergence and Features

- IB and 10GE HW/SW Products and Installations

- Sample Case Studies and Performance Numbers

- Conclusions and Final Q&A
Current and Next Generation Applications and Computing Systems

• Big demand for
  – High Performance Computing (HPC)
  – File-systems, multimedia, database, visualization
  – Enterprise Multi-tier datacenters

• Processor performance continues to grow
  – Chip density doubling every 18 months (multi-cores)

• Commodity networking also continues to grow
  – Increase in speed and features & affordable pricing

• Clusters are increasingly becoming popular to design next generation computing systems
  – Scalability, Modularity and Upgradeability with compute and network technologies
## Trends for Computing Clusters in the Top 500 List

- Top 500 list of Supercomputers ([www.top500.org](http://www.top500.org))

<table>
<thead>
<tr>
<th>Date</th>
<th>Number of Clusters</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jun. 2001:</td>
<td>33/500 (6.6%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2001:</td>
<td>43/500 (8.6%)</td>
<td></td>
</tr>
<tr>
<td>Jun. 2002:</td>
<td>80/500 (16%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2002:</td>
<td>93/500 (18.6%)</td>
<td></td>
</tr>
<tr>
<td>Jun. 2003:</td>
<td>149/500 (29.8%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2003:</td>
<td>208/500 (41.6%)</td>
<td></td>
</tr>
<tr>
<td>Jun. 2004:</td>
<td>291/500 (58.2%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2004:</td>
<td>294/500 (58.8%)</td>
<td></td>
</tr>
<tr>
<td>Jun. 2005:</td>
<td>304/500 (60.8%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2005:</td>
<td>360/500 (72.0%)</td>
<td></td>
</tr>
<tr>
<td>Jun. 2006:</td>
<td>364/500 (72.8%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2006:</td>
<td>361/500 (72.2%)</td>
<td></td>
</tr>
<tr>
<td>Jun. 2007:</td>
<td>373/500 (74.6%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2007:</td>
<td>406/500 (81.2%)</td>
<td></td>
</tr>
<tr>
<td>Jun. 2008:</td>
<td>400/500 (80.0%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2008:</td>
<td>410/500 (82.0%)</td>
<td></td>
</tr>
<tr>
<td>Jun. 2009:</td>
<td>410/500 (82.0%)</td>
<td></td>
</tr>
<tr>
<td>Nov. 2009:</td>
<td>To be announced</td>
<td></td>
</tr>
</tbody>
</table>
Integrated High-End Computing Environments

Enterprise Multi-tier Datacenter for Visualization and Mining

Supercomputing '09
Networking and I/O Requirements

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

- **Hardware Components**
  - Processing Core and Memory sub-system
  - I/O Bus
  - Network Adapter
  - Network Switch

- **Software Components**
  - Communication software

Supercomputing '09
Processing Bottlenecks in Traditional Protocols

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

- Traditionally relied on bus-based technologies
  - E.g., PCI, PCI-X, Shared Ethernet
  - 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
InfiniBand ("Infinite Bandwidth") and 10-Gigabit Ethernet

- Industry Networking Standards

- Processing Bottleneck:
  - Hardware offloaded protocol stacks with user-level communication access

- Network Bottleneck:
  - 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)
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 10GE today come as network adapters that plug into existing I/O technologies
Presentation Overview

- Introduction

- Why InfiniBand and 10-Gigabit Ethernet?

- Overview of IB, 10GE, their Convergence and Features

- IB and 10GE HW/SW Products and Installations

- Sample Case Studies and Performance Numbers

- Conclusions and Final Q&A
Trends in I/O Interfaces with Servers

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

<table>
<thead>
<tr>
<th>I/O Interface</th>
<th>Year(s)</th>
<th>Speed (shared bidirectional)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCI</td>
<td>1990</td>
<td>33MHz/32bit: 1.05Gbps</td>
</tr>
<tr>
<td>PCI-X</td>
<td>1998 (v1.0)</td>
<td>133MHz/64bit: 8.5Gbps</td>
</tr>
<tr>
<td></td>
<td>2003 (v2.0)</td>
<td>266-533MHz/64bit: 17Gbps</td>
</tr>
<tr>
<td>HyperTransport (HT) by AMD</td>
<td>2001 (v1.0), 2004 (v2.0), 2006 (v3.0), 2008 (v3.1)</td>
<td>102.4Gbps (v1.0), 179.2Gbps (v2.0), 332.8Gbps (v3.0), 409.6Gbps (v3.1)</td>
</tr>
<tr>
<td>PCI-Express (PCle) by Intel</td>
<td>2003 (Gen1), 2007 (Gen2), 2009 (Gen3 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>Intel QuickPath</td>
<td>2009</td>
<td>153.6-204.8Gbps per link</td>
</tr>
</tbody>
</table>

Supercomputing '09
Growth in Commodity Network Technology

Representative commodity networks; their entries into the market

<table>
<thead>
<tr>
<th>Technology</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>
<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>InfiniBand (2007 - )</td>
<td>32 Gbit/sec (4X QDR)</td>
</tr>
<tr>
<td>InfiniBand (2011 - )</td>
<td>64 Gbit/sec (4X EDR)</td>
</tr>
</tbody>
</table>

16 times in the last 9 years
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

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

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

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

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

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

- TCP Offload Engines (TOE)
  - Hardware Acceleration for the entire TCP/IP stack
  - Initially patented by Tehuti Networks
  - Actually refers to the IC on the network adapter that implements TCP/IP
  - In practice, 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

• Popularly known as Datacenter Ethernet
• Combines a number of Ethernet (optional) standards into one umbrella; sample enhancements include:
  – Priority-based flow-control: Link-level flow control for each Class of Service (CoS)
  – Enhanced Transmission Selection: Bandwidth assignment to each CoS
  – Datacenter Bridging Exchange Protocols: Congestion notification, Priority classes
  – End-to-end Congestion notification: Per flow congestion control to supplement per link flow control
Presentation Overview

• Introduction

• Why InfiniBand and 10-Gigabit Ethernet?

• Overview of IB, 10GE their Convergence and Features

• IB and 10GE HW/SW Products and Installations

• Sample Case Studies and Performance Numbers

• Conclusions and Final Q&A
IB, 10GE and their Convergence

• InfiniBand
  – Architecture and Basic Hardware Components
  – Novel Features
  – IB Verbs Interface
  – Management and Services
• 10-Gigabit Ethernet Family
  – Architecture and Components
  – Existing Implementations of 10GE/iWARP
• InfiniBand/Ethernet Convergence Technologies
  – Virtual Protocol Interconnect
  – InfiniBand over Ethernet
  – RDMA over Converged Enhanced Ethernet
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
  - Management and Services
    - Subnet Management
    - Hardware support for scalable network management
A Typical IB Network

Three primary components:
- Channel Adapters
- Switches/Routers
- Links and connectors
Components: Channel Adapters

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

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

• Network Links
  – Copper, Optical, Printed Circuit wiring on Back Plane
  – Not directly addressable

• Traditional adapters built for copper cabling
  – Restricted by cable length (signal integrity)

• Intel Connects: Optical cables with Copper-to-optical conversion hubs (acquired by Emcore)
  – Up to 100m length
  – 550 picoseconds copper-to-optical conversion latency

• Available from other vendors (Luxtera)

• Repeaters (Vol. 2 of InfiniBand specification)

(Courtesy Intel)
IB 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
  – Management and Services
    • Subnet Management
    • Hardware support for scalable network management
InfiniBand Communication Model

Basic InfiniBand Communication Semantics

Supercomputing '09
Communication in InfiniBand uses a Queue Pair Model for all data transfer

- Each QP has two queues
  - Send Queue (SQ)
  - Receive Queue (RQ)
- A QP must be linked to a Complete Queue (CQ)
  - Gives notification of operation completion from QPs
Queue Pair Model: WQEs and CQEs

- Entries used for QP communication are data structures called Work Queue Requests (WQEs)
  - Called “Wookies”
- Completed WQEs are placed in the CQ with additional information
  - They are now called CQEs (“Cookies”)
WQEs and CQEs

- Send WQEs contain data about what buffer to send from, how much to send, etc.
- Receive WQEs contain data about what buffer to receive into, how much to receive, etc.
- CQEs contain data about which QP the completed WQE was posted on, how much data actually arrived
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, the initiator must have the \( r\_key \) for the remote virtual address
  - Possibly exchanged with a send/recv
  - \( r\_key \) is not encrypted in IB

\( l\_key \) 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

Receive WQE contains information on the receive buffer; Incoming messages have to be matched to a receive WQE to know where to place the data
Communication in the Memory Semantics (RDMA Model)

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 and the receive buffer
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
  – Management and Services
    • Subnet Management
    • Hardware support for scalable network management
Hardware Protocol Offload

Complete Hardware Implementations Exist

Supercomputing '09
Link Layer Capabilities

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

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

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

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

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

• Service Level (SL):
  – Packets may operate at one of 16 different SLs
  – Meaning not defined by IB

• SL to VL mapping:
  – SL determines which VL on the next link is to be used
  – Each port (switches, routers, end nodes) has a SL to VL mapping table configured by the subnet management

• Partitions:
  – Fabric administration (through Subnet Manager) may assign specific SLs to different partitions to isolate traffic flows
Traffic Segregation Benefits

• Segregation of Server, Network and Storage Traffic – On the same physical network

\[\text{Servers} \rightarrow \text{InfiniBand Network} \rightarrow \text{IP Network} \rightarrow \text{Storage Area Network} \rightarrow \text{Servers}\]

• 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

Courtesy: Mellanox Technologies
Switching (Layer-2 Routing) and Multicast

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

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

Switching: 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 Clusters
- Different over-subscription ratio may be used
IB Switching/Routing: An Example

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

Switch decodes inbound packet header (LRH) DLID to determine target output ports.

Router decodes inbound packet header (GRH) GID multicast address to determine target output ports.
IB WAN Capability

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

Complete Hardware Implementations Exist
IB Network Layer Capabilities

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

• Routing follows the IPv6 packet format
  – Easy interoperability with Wide-area translations
  – Link layer might still need to be translated to the appropriate layer-2 protocol (e.g., Ethernet, SONET)

• Flow Labels allow routers to specify which packets belong to the same connection
  – Switches can optimize communication by sending packets with the same label in order
  – Flow labels can change in the router, but packets belonging to one label will always do so
Hardware Protocol Offload

Complete Hardware Implementations Exist

Consumer Transactions, Operations, etc. (IBA Operations)

IBA Operations (IBA Packets)

Transport Layer
Network Layer
Link Layer
PHY Layer

Channel Adapter

QP

CQE

WQE

Send

Rcv

Packet

Packet Relay

Physical Link (Symbols)

Transport

Consumer

Complete Implementations Exist

Supercomputing '09
## IB Transport Services

<table>
<thead>
<tr>
<th>Service Type</th>
<th>Connection Oriented</th>
<th>Acknowledged</th>
<th>Transport</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reliable Connection</td>
<td>Yes</td>
<td>Yes</td>
<td>IBA</td>
</tr>
<tr>
<td>Unreliable Connection</td>
<td>Yes</td>
<td>No</td>
<td>IBA</td>
</tr>
<tr>
<td>Reliable Datagram</td>
<td>No</td>
<td>Yes</td>
<td>IBA</td>
</tr>
<tr>
<td>Unreliable Datagram</td>
<td>No</td>
<td>No</td>
<td>IBA</td>
</tr>
<tr>
<td>RAW Datagram</td>
<td>No</td>
<td>No</td>
<td>Raw</td>
</tr>
</tbody>
</table>

- Each transport service can have zero or more QPs associated with it
- e.g., you can have 4 QPs based on RC and one based on UD
## Trade-offs in Different Transport Types

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

- SRQ is a hardware mechanism for a process to share receive resources (memory) across multiple connections
  - Introduced in specification v1.2
- $0 < p \ll m(n-1)$
eXtended Reliable Connection (XRC)

- Each QP takes at least one page of memory
  - Connections between all processes is very costly for RC
- **New** IB Transport added: eXtended Reliable Connection
  - Allows connections between nodes instead of processes
IB 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
  - Management and Services
    - Subnet Management
    - Hardware support for scalable network management
Concepts in IB Management

• Agents
  – Processes or hardware units running on each adapter, switch, router (everything on the network)
  – Provide capability to query and set parameters

• Managers
  – Make high-level decisions and implement it on the network fabric using the agents

• Messaging schemes
  – Used for interactions between the manager and agents (or between agents)

• Messages
Subnet Manager
10GE Overview

- 10-Gigabit Ethernet Family
  - Architecture and Components
    - Stack Layout
    - Out-of-Order Data Placement
    - Dynamic and Fine-grained Data Rate Control
  - Existing Implementations of 10GE/iWARP
## IB and 10GE: Commonalities and Differences

<table>
<thead>
<tr>
<th>Feature</th>
<th>IB</th>
<th>iWARP/10GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Acceleration</td>
<td>Supported</td>
<td>Supported (for TOE and iWARP)</td>
</tr>
<tr>
<td>RDMA</td>
<td>Supported</td>
<td>Supported (for iWARP)</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>Data Placement</td>
<td>Ordered</td>
<td>Out-of-order (for iWARP)</td>
</tr>
<tr>
<td>Data Rate-control</td>
<td>Static and Coarse-grained</td>
<td>Dynamic and Fine-grained (for TOE and iWARP)</td>
</tr>
<tr>
<td>QoS</td>
<td>Prioritization</td>
<td>Prioritization and Fixed Bandwidth QoS</td>
</tr>
</tbody>
</table>

Supercomputing '09
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)

(Courtesy iWARP Specification)
Decoupled Data Placement and Data Delivery

• Place data as it arrives, whether in or out-of-order
• If data is out-of-order, place it at the appropriate offset
• Issues from the application’s perspective:
  – Second half of the message has been placed does not mean that the first half of the message has arrived as well
  – If one message has been placed, it does not mean that the previous messages have been placed
Protocol Stack Issues with Out-of-Order Data Placement

• The receiver network stack has to understand each frame of data
  – If the frame is unchanged during transmission, this is easy!

• Issues to consider:
  – *Can we guarantee that the frame will be unchanged?*
  – *What happens when intermediate switches segment data?*
Intermediate Ethernet switches (e.g., those which support splicing) can segment a frame to multiple segments or coalesce multiple segments to a single segment.
Marker PDU Aligned (MPA) Protocol

- Deterministic Approach to find segment boundaries
- Approach:
  - Places strips of data at regular intervals (based on data sequence number)
  - Interval is set to be 512 bytes (small enough to ensure that each Ethernet frame has at least one)
    - Minimum IP packet size is 536 bytes (RFC 879)
  - Each strip points to the RDDP header
- Each segment independently has enough information about where it needs to be placed
MPA Frame Format

RDDP Header

Data Payload (if any)

Segment Length

Pad

CRC

Data Payload (if any)

Supercomputing '09
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 vs. Fixed Bandwidth QoS

• Can allow for simple prioritization:
  – E.g., connection 1 performs better than connection 2
  – 8 classes provided (a connection can be in any class)
    • Similar to SLs in InfiniBand
  – Two priority classes for high-priority traffic
    • E.g., management traffic or your favorite application

• Or can allow for specific bandwidth requests:
  – E.g., can request for 3.62 Gbps bandwidth
  – Packet pacing and stalls used to achieve this

• Query functionality to find out “remaining bandwidth”
10GE Overview

- 10-Gigabit Ethernet Family
  - Architecture and Components
    - Stack Layout
    - Out-of-Order Data Placement
    - Dynamic and Fine-grained Data Rate Control
  - Existing Implementations of 10GE/iWARP
Current Usage of Ethernet

Regular Ethernet

TOE

iWARP

System Area Network or Cluster Environment

Wide Area Network

Regular Ethernet Cluster

Distributed Cluster Environment

iWARP Cluster

Supercomputing '09
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

Application

User-level iWARP

Sockets

TCP (Modified with MPA)

IP

Device Driver

Network Adapter

Regular Ethernet Adapters

OSU, ANL

Application

High Performance Sockets

Sockets

TCP

IP

Device Driver

Network Adapter

TCP Offload Engines

Chelsio, NetEffect, Ammasso

Application

High Performance Sockets

Sockets

TCP

IP

Device Driver

Network Adapter

Offloaded TCP

Offloaded IP

Offloaded iWARP

Network Adapter

iWARP compliant Adapters

Supercomputing '09
IB and 10GE Convergence

- InfiniBand/Ethernet Convergence Technologies
  - Virtual Protocol Interconnect
  - InfiniBand over Ethernet
  - RDMA over Converged Enhanced Ethernet
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)

- 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
- Has all the benefits of IB verbs

Cons:
- Network bandwidth limited to Ethernet switches (currently 10Gbps), even though IB can provide 32Gbps
- Some IB native link-layer features are optional in (regular) Ethernet
**InfiniBand** RDMA over Converged Enhanced Ethernet (RoCEE)

- Very similar to IB over Ethernet
  - Often used interchangeably with IBoE
  - Can be used to explicitly specify link layer is Converged Enhanced Ethernet (CEE)

- **Pros:**
  - Works natively in Ethernet environments
  - Has all the benefits of IB verbs
  - CEE is very similar to the link layer of native IB, so there are no missing features

- **Cons:**
  - Network bandwidth limited to Ethernet switches (currently 10Gbps), even though IB can provide 32Gbps

---

Supercomputing '09 81
<table>
<thead>
<tr>
<th>Feature</th>
<th>IB</th>
<th>iWARP/10GE</th>
<th>IBoE</th>
<th>RoCEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Acceleration</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>
</tr>
<tr>
<td>Atomic Operations</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Multicast</td>
<td>Optional</td>
<td>No</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>
</tr>
<tr>
<td>Prioritization</td>
<td>Optional</td>
<td>Optional</td>
<td>Optional</td>
<td>Yes</td>
</tr>
<tr>
<td>Fixed BW QoS (ETS)</td>
<td>No</td>
<td>Optional</td>
<td>Optional</td>
<td>Yes</td>
</tr>
<tr>
<td>Ethernet Compatibility</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>TCP/IP Compatibility</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

Supercomputing ’09
Presentation Overview

- Introduction
- Why InfiniBand and 10-Gigabit Ethernet?
- Overview of IB, 10GE, their Convergence and Features
- IB and 10GE 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)
Tyan Thunder S2935 Board

(Courtesy Tyan)

Similar boards from Supermicro with LOM features are also available
• 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 early 2008
    • Up to 648-port QDR switch by SUN
  – New IB switch silicon from Qlogic introduced at SC ’08
    • Up to 846-port QDR switch by Qlogic

• Switch Routers with Gateways
  – IB-to-FC; IB-to-IP
IB Software Products

• Low-level software stacks
  – VAPI (Verbs-Level API) from Mellanox
  – Modified and customized VAPI from other vendors
  – New initiative: Open Fabrics (formerly OpenIB)
    • http://www.openfabrics.org
    • Open-source code available with Linux distributions
    • Initially IB; later extended to incorporate iWARP

• High-level software stacks
  – MPI, SDP, IPoIB, SRP, iSER, DAPL, NFS, PVFS on various stacks (primarily VAPI and OpenFabrics)
10G, 40G and 100G Ethernet Products

- 10GE adapters
  - Intel, Myricom, Mellanox (ConnectX)
- 10GE/iWARP adapters
  - Chelsio, NetEffect (now owned by Intel)
- 10GE switches
  - Fulcrum Microsystems
    - Low latency switch based on 24-port silicon
    - FM4000 switch with IP routing, and TCP/UDP support
  - Fujitsu, Woven Systems (144 ports), Myricom (512 ports), Quadrics (96 ports), Force10, Cisco, Arista (formerly Arastra)
- 40GE and 100GE switches
  - Nortel Networks
    - 10GE downlinks with 40GE and 100GE uplinks
Mellanox ConnectX Architecture

- Early adapter supporting IB/10GE convergence
  - Support for VPI and IBoE
- Includes other features as well
  - Hardware support for Virtualization
  - Quality of Service
  - Stateless Offloads

(Courtesy Mellanox)
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 1.4.2
  - OFED 1.5 is underway
OpenFabrics Stack with Unified Verbs Interface
OpenFabrics on Convergent IB/10GE

- For IBoE and RoCEE, 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.

Supercomputing '09
InfiniBand in the Top500

Systems

Performance

Percentage share of InfiniBand is steadily increasing
### Large-scale InfiniBand Installations

- 151 IB clusters (30.2%) in the June ’09 TOP500 list ([www.top500.org](http://www.top500.org))
- Installations in the Top 30 (15 of them):

<table>
<thead>
<tr>
<th>129,600 cores (RoadRunner) at LANL (1&lt;sup&gt;st&lt;/sup&gt;)</th>
<th>12,288 cores at GENCI-CINES, France (20&lt;sup&gt;th&lt;/sup&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>51,200 cores (Pleiades) at NASA Ames (4&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>8,320 cores in UK (25&lt;sup&gt;th&lt;/sup&gt;)</td>
</tr>
<tr>
<td>62,976 cores (Ranger) at TACC (8&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>8,320 cores in UK (26&lt;sup&gt;th&lt;/sup&gt;)</td>
</tr>
<tr>
<td>26,304 cores (Juropa) at TACC (10&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>8,064 cores (DKRZ) in Germany (27&lt;sup&gt;th&lt;/sup&gt;)</td>
</tr>
<tr>
<td>30,720 cores (Dawning) at Shanghai (15&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>12,032 cores at JAXA, Japan (28&lt;sup&gt;th&lt;/sup&gt;)</td>
</tr>
<tr>
<td>14,336 cores at New Mexico (17&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>10,240 cores at TEP, France (29&lt;sup&gt;th&lt;/sup&gt;)</td>
</tr>
<tr>
<td>14,384 cores at Tata CRL, India (18&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>13,728 cores in Sweden (30&lt;sup&gt;th&lt;/sup&gt;)</td>
</tr>
<tr>
<td>18,224 cores at LLNL (19&lt;sup&gt;th&lt;/sup&gt;)</td>
<td><strong>More are getting installed!</strong></td>
</tr>
</tbody>
</table>

Supercomputing ’09
10GE Installations

• Several Enterprise Computing Domains
  – Enterprise Datacenters (HP, Intel) and Financial Markets
  – Animation firms (e.g., Universal Studios created “The Hulk” and many new movies using 10GE)

• Scientific Computing Installations
  – 5,600-core installation in Purdue with Chelsio/iWARP
  – 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 3, 7, 9, 14, 24 in the Top 25)

• ESnet to install 62M $ 100GE infrastructure for US DOE
Dual IB/10GE Systems

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

(Courtesy Taisuke Boku, University of Tsukuba)

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

• Introduction
• Why InfiniBand and 10-Gigabit Ethernet?
• Overview of IB, 10GE, their Convergence and Features
• IB and 10GE 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)

• File Systems
Low-level Latency Measurements

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

Supercomputing '09
Low-level Uni- and Bi-directional Bandwidth Measurements

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

Supercomputing '09
Case Studies

• Low-level Performance

• Message Passing Interface (MPI)

• File Systems
Message Passing Interface (MPI)

- De-facto message passing standard
  - Point-to-point communication
  - Collective communication (broadcast, multicast, reduction, barrier)
  - MPI-1 and MPI-2 available; MPI-3 under discussion
- Has been implemented for various past commodity networks (Myrinet, Quadrics)
- How can it be designed and efficiently implemented for InfiniBand and iWARP?
MVAPICH/MVAPICH2 Software

- High Performance MPI Library for IB and 10GE
  - MVAPICH (MPI-1) and MVAPICH2 (MPI-2)
  - Used by more than 975 organizations in 51 countries
  - More than 34,000 downloads from OSU site directly
  - Empowering many TOP500 clusters
    - 8th ranked 62,976-core cluster (Ranger) at TACC
  - Available with software stacks of many IB, 10GE and server vendors including Open Fabrics Enterprise Distribution (OFED)
  - Also supports uDAPL device to work with any network supporting uDAPL
    - http://mvapich.cse.ohio-state.edu/
MPICH2 Software Stack

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

Supercomputing '09
One-way Latency: MPI over IB

**Small Message Latency**

- InfiniHost III and ConnectX-DDR: 2.33 GHz Quad-core (Clovertown) Intel with IB switch
- ConnectX-QDR-PCIe2: 2.83 GHz Quad-core (Harpertown) Intel with back-to-back

**Large Message Latency**

- MVAPICH-InfiniHost III-DDR
- MVAPICH-Qlogic-SDR
- MVAPICH-ConnectX-DDR
- MVAPICH-ConnectX-QDR-PCIe2
- MVAPICH-Qlogic-DDR-PCIe2

Supercomputing '09
Bandwidth: MPI over IB

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

ConnectX-QDR-PCIe2: 2.4 GHz Quad-core (Nehalem) Intel with IB switch

Supercomputing '09
One-way Latency: MPI over iWARP

One-way Latency

- Chelsio (TCP/IP)
- Chelsio (iWARP)

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

Supercomputing '09
Bandwidth: MPI over iWARP

Unidirectional Bandwidth

- Blue diamonds: Chelsio (TCP/IP)
- Red squares: Chelsio (iWARP)

Message Size (bytes)

- MillionBytes/sec

Message Size (bytes)

- MillionBytes/sec

Bidirectional Bandwidth

- Blue diamonds: Chelsio (TCP/IP)
- Red squares: Chelsio (iWARP)

Message Size (bytes)

- MillionBytes/sec

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

Supercomputing ’09
Convergent Technologies: MPI Latency

**Small Messages**
- Native IB
- VPI-IB
- VPI-Eth
- IBoE

**Large Messages**

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

Supercomputing '09
Convergent Technologies: MPI Uni- and Bi-directional Bandwidth

Uni-directional Bandwidth

- Native IB
- VPI-IB
- VPI-Eth
- IBoE

Bi-directional Bandwidth

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

- Low-level Performance
- Message Passing Interface (MPI)
- File Systems
Sample file systems:
- Lustre, Panasas, GPFS, Sistina/Redhat GFS
- PVFS, Google File systems, Oracle Cluster File system (OCFS2)
• Lustre over Native IB
  – Write: 1.38X faster than IPoIB; Read: 2.16X faster than IPoIB
• Memory copies in IPoIB and Native IB
  – Reduced throughput and high overhead; I/O servers are saturated
CPU Utilization

- 4 OSS nodes, IOzone record size 1MB
- Offers potential for greater scalability
NFS/RDMA Performance

- IOzone Read Bandwidth up to 913 MB/s (Sun x2200’s with x8 PCIe)
- Read-Write design by OSU, available with the latest OpenSolaris
- NFS/RDMA is being added into OFED 1.4


Supercomputing ’09
Summary of Design Performance Results

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

- Introduction
- Why InfiniBand and 10-Gigabit Ethernet?
- Overview of IB, 10GE, their Convergence and Features
- IB and 10GE HW/SW Products and Installations
- Sample Case Studies and Performance Numbers
- Conclusions and Final Q&A
Concluding Remarks

• Presented network architectures & trends in Clusters
• Presented background and details of IB and 10GE
  – Highlighted the main features of IB and 10GE and their convergence
  – Gave an overview of IB and 10GE hw/sw products
  – Discussed sample performance numbers in designing various high-end systems with IB and 10GE
• IB and 10GE are emerging as new architectures leading to a new generation of networked computing systems, opening many research issues needing novel solutions

Supercomputing '09
Funding Acknowledgments

Our research is supported by the following organizations

- **Funding support by**
  - [Office of Science](#)
  - [NSF](#)
  - [Mellanox Technologies](#)
  - [Cisco Systems](#)
  - [QLogic](#)
  - [Intel](#)
  - [Linux Networx](#)
  - [NetApp](#)
  - [Sun](#)

- **Equipment support by**
  - [Mellanox Technologies](#)
  - [Intel](#)
  - [AMD](#)
  - [Sun](#)
  - [APPR](#)
  - [advanced clustering technologies, inc.](#)
  - [Microway](#)
  - [QLogic](#)

Supercomputing ’09
Personnel Acknowledgments

Current Students

– M. Kalaiya (M. S.)
– K. Kandalla (M.S.)
– P. Lai (Ph.D.)
– M. Luo (Ph.D.)
– G. Marsh (Ph. D.)
– X. Ouyang (Ph.D.)
– S. Potluri (M. S.)
– 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.)

Current Post-Doc

– E. Mancini

Current Programmer

– J. Perkins

– A. Mamidala (Ph.D.)
– 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.)

Supercomputing '09
Web Pointers

http://www.cse.ohio-state.edu/~panda
http://www.mcs.anl.gov/~balaji
http://www.cse.ohio-state.edu/~koop
http://nowlab.cse.ohio-state.edu

MVAPICH Web Page
http://mvapich.cse.ohio-state.edu

panda@cse.ohio-state.edu
balaji@mcs.anl.gov
matthew.koop@nasa.gov