

Overview of InfiniBand Architecture



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

Limitations of Traditional Host-based Protocols

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

Capabilities of Current High-Performance Networks

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

Previous High Performance Network Stacks

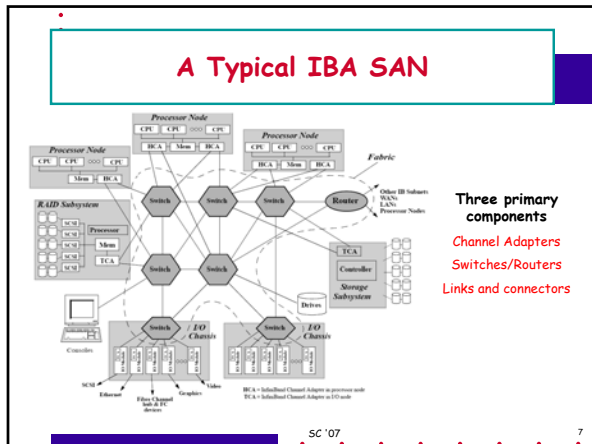
- Virtual Interface Architecture
 - Standardized by Intel, Compaq, Microsoft
- Fast Messages (FM)
 - Developed by UIUC
- Myricom GM
 - Proprietary protocol stack from Myricom
- These network stacks set the trend for high-performance communication requirements
 - Hardware offloaded protocol stack
 - Support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

IBA Trade Organization

- IBA Trade Organization 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 IBA architecture specification
- InfiniBand Architecture (Volume 1, Version 1.0) was released to public on Oct 24, 2000
- www.infinibandta.org

IBA Architecture Overview

- Architecture and Basic Components
- Communication and I/O Operations
- Transport Layer/Services and Reliability
- Routing, QoS, Congestion Control, Multicast, WAN capability
- Management and Services



Components: Channel Adapters

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

SC '07 8

Components: Switches and Routers

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

SC '07 9

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
 - Up to 100m length
 - 550 picoseconds copper-to-optical conversion latency
- Repeaters (Vol. 2 of InfiniBand specification)

Courtesy Intel
SC '07 10

Communication Queuing Model on HCA

- Each request is a Work Queue Element (WQE)
- WQEs placed in Work Queues
- HCA executes WQE
- Completion Queue Entry (CQE) generated for each WQE
- CQEs placed in associated Completion Queue (CQ)

(courtesy IBTA)

SC '07 11

Communication Operations

- Channel Semantic
 - Send and Recv
- Memory Semantic
 - RDMA read
 - RDMA write
 - RDMA atomic operations
 - E.g. Fetch & Add, Compare & Swap

SC '07 12

Basic Communication model

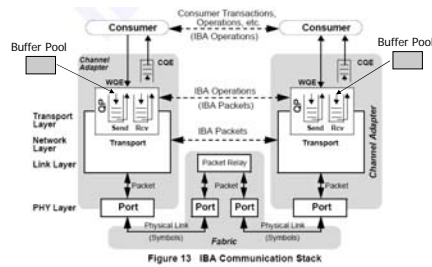
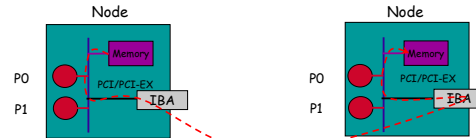


Figure 13 IBA Communication Stack

SC '07

13

Principles behind RDMA Mechanism



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

SC '07

14

Transport Services

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

SC '07

15

IBA Transport Service Types

Attribute	Reliable Connection	Reliable Datagram	Unreliable Datagram	Unreliable Connection	Raw Datagram (both IPv4 & ethernet)
Scalability (M processes on N Processor nodes communicating with all processes on all nodes)	M*N QPs required on each processor node, per CA.	M QPs required on each processor node, per CA.	M QPs required on each processor node, per CA.	M*N QPs required on each processor node, per CA.	1 QP required on each end node, per CA.
Corrupt data detected			Yes		
Data delivery guarantee	Data delivered exactly once			No guarantees	
Data order guaranteed	Yes, per connection	Yes, packets from any one source QP are ordered to multiple destination QPs.	No	Unordered and duplicate packets are detected.	No
Data loss detected	Yes	Yes	No	Yes	No
Error recovery	Reliable: Errors are detected at both the requestor and the responder. The requestor can transparently recover from errors (retransmission, alternate path, etc.) without any involvement of the client application. QP processing is halted only if the destination is inoperable or all fabric paths between the channel adapters have failed.	Unreliable: Packets with errors, including sequence errors, are detected and may be logged by the responder. The requestor is not informed.	Unreliable: Packets with errors, including sequence errors, are detected and may be logged by the responder. The requestor is not informed.	Unreliable: Packets with errors are not delivered. The requestor and responder are not informed of dropped packets.	Unreliable: Packets with errors are not delivered. The requestor and responder are not informed of dropped packets.

SC '07

16

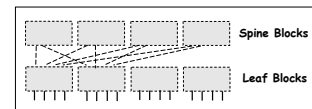
IBA Architecture Overview

- Architecture and Basic Components
- Communication and I/O Operations
- Transport Layer/Services and Reliability
- Routing, QoS, Congestion Control, Multicast, WAN capability
- Management and Services

SC '07

17

Destination Based Routing in IB



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

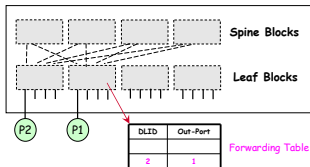
Switching: IBA supports Virtual Cut Through (VCT)
 Routing: Unspecified by IBA SPEC
 Up*/Down*, Shift are popular routing engines supported by OFED

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

SC '07

18

IB Routing: An Example



- Subnet Manager discovers, configures and maintains the cluster
 - Assigns LID(s) to ports and forwarding table to switches
- Different routing algorithms may give different paths
- Multi-Pathing may also be used for avoiding hot-spots

SC '07

19

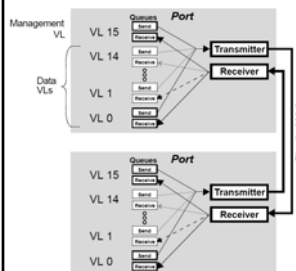
Automatic Path Migration

- Enables migrating connections to a different path
 - Connection recovery in the case of failures
 - Optional Feature
- Available for RC, UC, and RD
- Reliability guarantees for service type maintained during migration

SC '07

20

Quality Management: Virtual Lanes



- Multiple virtual links within same physical link
- Separate buffers and flow control
- VL15: reserved for management
- Each port supports one or more data VL

SC '07

21

QoS Mechanisms

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

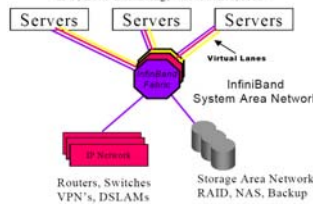
SC '07

22

Benefits of IBA for having Multiple Traffic on the same Network

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

IPC, Load Balancing, Web Caches, ASP



- 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 of Mellanox Technologies, Inc.

SC '07

23

Congestion Control

- Switch detects congestion on a VL
 - Detects whether it is the root or victim of congestion
- InfiniBand follows a three-step protocol
 - Forward Explicit Congestion Notification
 - Used to communicate congested port status
 - Switch sets FECN bit; marks packets leaving the congested state
 - Backward Explicit Congestion Notification
 - Destination sends BECN to sender informing about congestion
 - Injection Rate Control (Throttling)
 - Source throttles its send rate temporarily (timer based)
 - Original injection rate reduces over time
 - Congestion control may be performed per QP or SL
- Pro-active → does not wait for packet drops to occur

SC '07

24

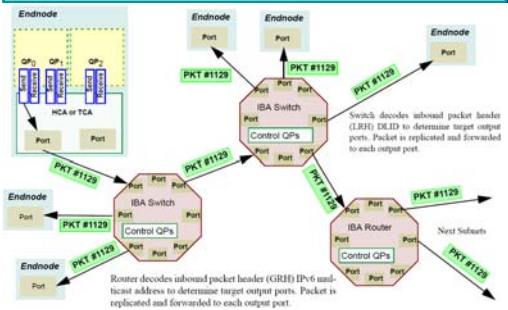
Multicast

- Based on Multicast LIDs and GIDs
- Switches and Routers replicate packets
- Ensures at-most-once delivery and loop-free forwarding
- IBA defines interface for multicast group management protocol
 - Create multicast group
 - Join/Leave multicast group
 - Prune multicast group
 - Delete multicast group

SC '07

25

Multicast Example



SC '07

26

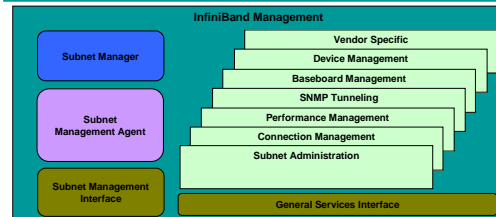
IB WAN Capability

- Getting increased attention for
 - Remote Storage
 - Cluster Aggregation (Cluster-of-clusters)
 - Remote Visualization
- IB-Optical hybrid switch by Obsidian Research Corporation
 - www.obsidianresearch.com
 - Transparently extends IB fabric over dark fiber
 - Low-latency copper-optical-copper conversion
- Link-level buffer credit flow-control
 - Data messages do not have to wait for round-trip hops
 - Important in the wide-area network

SC '07

27

Management Model

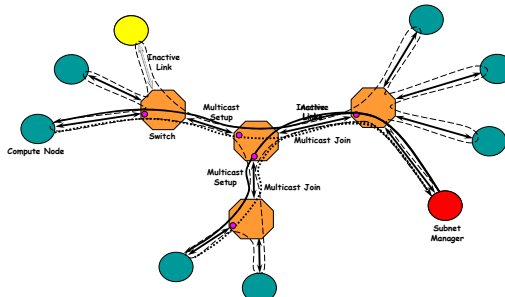


- QP0 and QP1 are special QPs on each port
 - QP0 provides Subnet Management Interface (services)
 - QP1 provides General Services Interface (services)

SC '07

28

Subnet Manager



SC '07

29

IBA Vendors

- Many different vendors
 - Mellanox
 - Voltaire
 - Cisco
 - QLogic
- IBA vendors are being aligned with different Server Vendors (Intel, IBM, SUN, Dell) and Integrators (Linux Networks, Microway, Appro)

SC '07

30

IB Hardware Products

- **Adapters:**
 - Dual port 4X (10.0 Gbps bidir) with PCI-X 64 bit/133 MHz
 - Dual-port 4X (20.0 Gbps bidir) with PCIe x8
- **Mem Free Adapter**
 - No memory on HCA for storing connection information
 - Uses main memory of the system (through PCIe)
 - Good for 'Landed on Motherboard (LOM)' design, blades
- **IBTA has released the specs for**
 - DDR (Double Data Rate) and QDR (Quad Data Rate)
 - Single-port and Dual-port DDR cards available (4X: 20.0 Gbps)
- **Some emerging 12X SDR adapters**
 - 12X: 30.0 Gbps unidirectional and 60.0 Gbps bidirectional

SC '07

31

IB Hardware Products (contd.)

- **Customized adapters to work with IBA switches**
 - Cray XD1 (formerly by Octigabay)
 - Qlogic (tightly integrated with HT; also available with PCIe)
- **Switches:**
 - 4X SDR switch (8-288 ports; value added services/features)
 - Larger switches can be built using these switches
 - 12X ports (30 Gbps) available for inter-switch connectivity
 - 4X DDR switch (mainly available in 8 to 288 port models)
 - 3456-port "Magnum" switch from SUN → used at TACC
 - 12X switches (smaller sizes available)
- **Switch Routers with Gateways**
 - IB-to-FC; IB-to-IP

SC '07

32

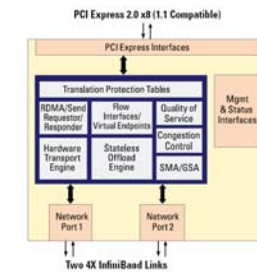
IB Software Products

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

SC '07

33

Mellanox ConnectX Architecture



ConnectX Architecture, courtesy Mellanox

- **Fourth Generation Silicon**
 - DDR (Double Data Rate)
 - PCI-Express Gen1
 - QDR (Quad Data Rate)
 - PCI-Express Gen2
- **Flexibility to configure each individual port to either InfiniBand or 10G**
- **Hardware support for Virtualization**
- **Quality of Service Stateless Offloads**

SC '07

34

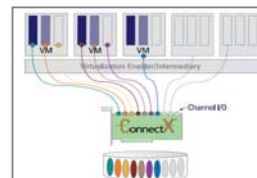
Stateless Offload In ConnectX

- Available for IPoIB protocol (TCP/IP)
 - Hardware segmentation
 - Checksum Operations
 - Large/Giant Send Offload
 - Receive Side Scaling
 - Split Header and Payload Processing
 - Interrupt Moderation
 - Multiple Send and Receive Queues

SC '07

35

Direct Hardware Access in Virtual Machines



ConnectX Virtualization Support, courtesy Mellanox

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

SC '07

36

Quality of Service

- Marking, shaping and Queuing
 - IEEE 802.1 and DiffServ
- End-to-end congestion control
- Granular flow-control

ConnectX QoS Support, courtesy Mellanox

SC '07 37

OpenFabrics

- www.openfabrics.org
- Open source organization (formerly known as OpenIB)
- Incorporates both IB and iWARP in a unified manner
- Focusing on effort for Open Source IBA and iWARP support for Linux and Windows
- Design of complete software stack with 'best of breed' components
 - Gen1
 - Gen2 (current focus)
- Users should be able to download the entire stack and run without any problem
 - Latest release is OFED 1.2.5
 - OFED 1.3 is being worked out

SC '07 38

OpenFabrics Stack with Unified Verbs Interface

SC '07 39

OpenFabrics Software Stack

ISA	Socket Administration
IBAD	Management Datapath
ISBA	Subnet Manager Agent
IBRA	Performance Manager Agent
IBCB	IPsec Datapath
ISDP	Socket Direct Protocol
SRP	SCSI RDMA Protocol
ISDR	SCSI RDMA Protocol
ISDS	RDMA Datagram Service
UDAPL	User Direct Access Programming Library
IBCA	Host Channel Adapter
IBHCC	IBMA HCC

SC '07 40