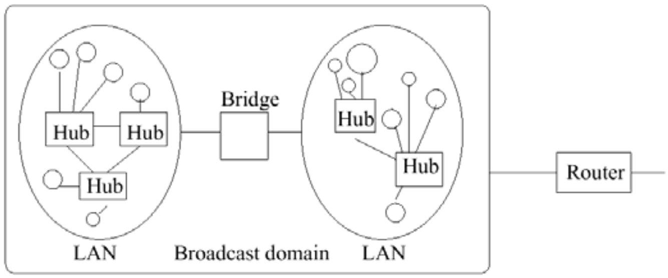
SAN for Clusters

- More detailed discussion of
  - Fast Ethernet
  - Gigabit Ethernet
  - Myrinet
  - Infiniband

Fast Ethernet (I)

- 100 Mbps over UTP or fiber-optic cable
- MAC protocol: CSMA/CD (Carrier Sense Multiple Access with Collision Detection)



Fast Ethernet (II)

- Interconnection devices
  - Repeater
    - restore data and collision signal
    - amplify and regenerate signals
  - Hub
    - central point
    - repeat and copy: All can see it
  - Bridge
    - link adjacent LANs: datalink layer
    - filtering
    - forward to other segment
  - Router
    - link adjacent LANs: network layer
    - shortest path

Gigabit Ethernet Fundamentals

- Gigabit Ethernet (802.3z) 1Gbps
  - Modified Fiber Channel physical layer
  - Actual Bit Rate 1,250,000,000 bits/second!
- Media: Fiber (multimode,singlemode), UTP Cat 5
- Uses the 802.3 Ethernet frame format.
- Full-duplex (point to point) mode and/or half-duplex (switched) mode.
- Half-duplex mode
  - Uses *enhanced* CSMA/CD access method
  - requires carrier extension to 512 byte time slot to preserve 200m collision domain
  - thus wastes bandwidth

Gigabit Ethernet Fundamentals Cont.

- Support Fiber and Copper media.
  - 25 meters short link copper.
  - 100 meters horizontal copper.
  - 500 meters multimode fiber.
  - 3000 meters single mode fiber.
- Paul A. Farrell, Hong Ong, Communication Performance over a Gigabit Network "http://discov.cs.kent.edu/publications/2000/ipccc2000.pdf"
- Ethernet – Next Generation.
  - 10 Gigabit Ethernet (802.3ae).
  - http://www.10gea.org/.

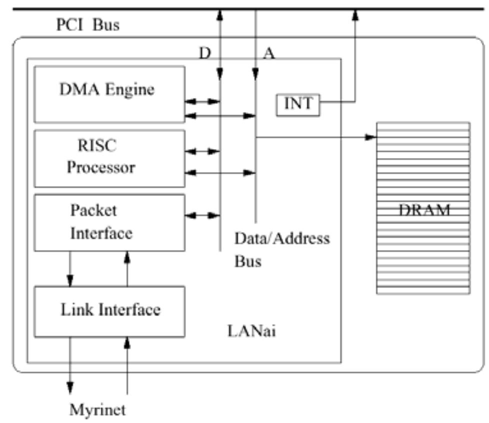
Myrinet (I)

- A SAN evolved from supercomputer technology
- A main product of Myricom (founded in 1994)
- Quite popular in the research community
  - all HW & SW specifications are open & public
- Based on 2 research projects
  - Mosaic by Caltech
    - a fine grain supercomputer, need a truly scalable interconnection network with lots of bandwidth
  - Atomic LAN by USC
    - based on Mosaic technology, a research prototype of Myrinet
- Speed: 1.28 Gbps
- Good price/performance ratio

Myrinet (II)

- Host interface
  - LANai chip
    - a custom VLSI chip, a programmable microcontroller
    - control the data transfer between the host & the network
  - SRAM memory
    - Message data must first be written to the NI SRAM, before it can be injected into the network
  - (+) the great flexibility of the HW due to a programmable microcontroller,
  - (-) but can also be a bottleneck with respect to performance since the LANai runs only at moderate frequencies

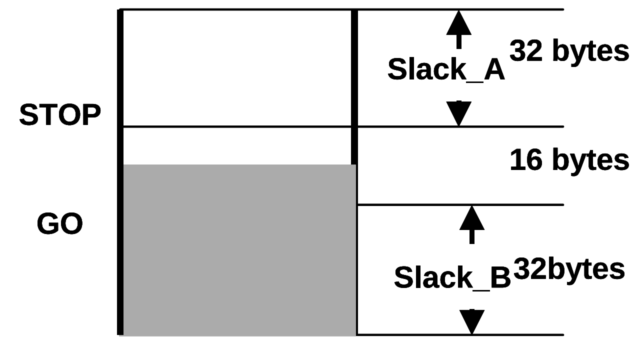
Myrinet Host Interface



Myrinet (III)

- Link and Packet Layer
  - similar to ServerNet
  - full duplex 9 bit parallel channel in one direction running at 80MHz
  - network offer 160Mbyte/s physical bandwidth over one channel
  - two different cable type (SAN, LAN)
    - 3m SAN link, 10m LAN link
  - variable length data format
  - route with wormhole switching
  - source path routing
    - consist of routing header
  - special control symbols (STOP, GO)

Myrinet Flow Control (Slack Buffer Operation)



Myrinet (IV)

- Switches
  - 4, 8 and 16 ports, mixable SAN and LAN
  - any network topology
  - autodetect the absence of a link
  - starting up, host interface detects network topology automatically
- Error Handling
  - MTBF: million hours are reported
  - cable fault and node failure
    - alternative routing by LANai
  - prevent deadlock: time out generates a forward reset (FRES) signal

Performance of Message Layers over Myrinet

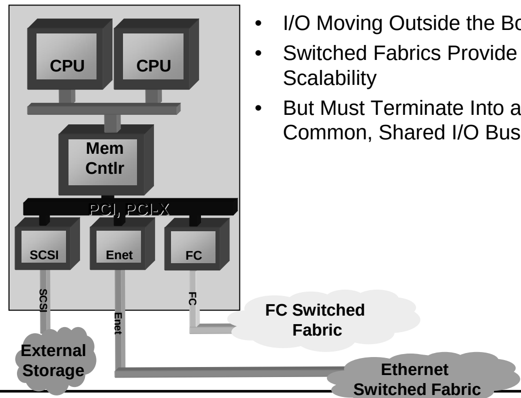
BIP : Basic Interface for Parallelism
http://www.cs.utk.edu/~dongarra/BLACKBERRY/prylli.pdf

PM: SCore Protocol Module - low-level high performance communication library

AM: Active Messages

Machine	API	Latency (μs)	Bandwidth (Mbit/s)	Ref.
200 MHz PPro	BIP	4.8	1009	LHPC
166 MHz Pentium	PM	7.2	941	RWCP
Ultra-1	AM	10	280	GAM
200 MHz PPro	TCP (Linux/BIP)		293	LHPC
200 MHz PPro	UDP (Linux/BIP)		324	LHPC
DEC Alpha 500/266	TCP (Digital Unix)		271	Duke
DEC Alpha 500/266	UDP (Digital Unix)		404	Duke

I/O Architecture Trends



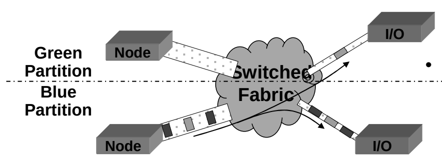
- I/O Moving Outside the Box
- Switched Fabrics Provide Scalability
- But Must Terminate Into a Common, Shared I/O Bus

Infiniband Architectural Goals

- Bandwidth tracking Moore's Law or better
- Improved latency and fewer interrupts compared to current architectures
- Competitive cost for every tier of server market
- Unified interconnect for IPC, network I/O, storage I/O
- Scalable to hundreds-thousands of endpoints
- Message-based model
- Support emerging distributed applications
- Very high reliability and availability
- Complement to Internet computing strategies
- Practical to implement

InfiniBand™ Protocol Features

- Flow & rate control
  - Static rate control to control sources and destinations of different speeds
  - Credit based link-level flow control for efficient congestion avoidance

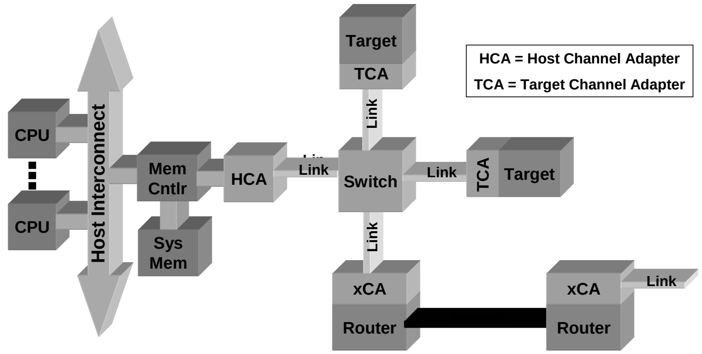


- Partitioning
  - For performance and functional isolation
  - Transparent I/O Sharing
- Multicasts
  - A single message is distributed by the fabric to multiple destinations
- Network topology
  - Subnet support
  - IPv6 GUIDs for unique end-point id
  - IPv6 addressing headers for global routing
  - Speed matched to the backbone
  - IP compatible fabric management

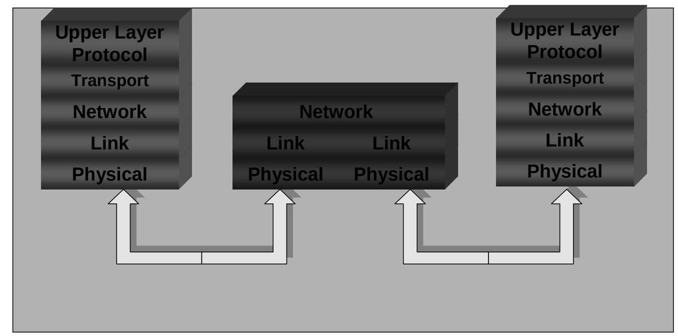
InfiniBand Components

- Link
- Switch
- Router
- Target Channel Adapter (TCA) – I/O devices
- Host Channel Adapter (HCA) - machines

InfiniBand™ Architecture (IBA)

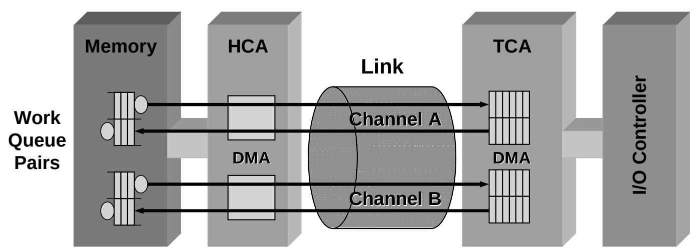


Layered Architecture



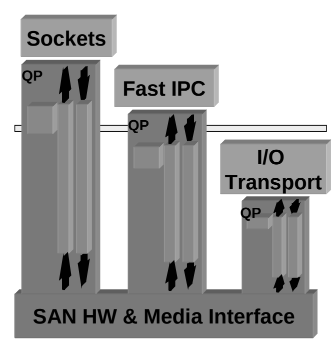
- Provides forward and backward compatibility / inter-operability
- Allows layers to evolve at the rate technology evolves.

Infiniband Channel-Based I/O



- Logical, connection-based path between two address spaces
  - Protected DMA engine at each end
  - Driven by pairs of Work Queues

InfiniBand™ IPC

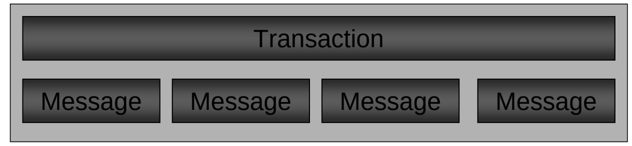


- Sockets
  - Match IPC Hardware to Legacy Software Needs
- Fast IPC
  - OS Vendor Provided
  - Native HW Performance
- I/O Transport
  - Message Based
  - General Purpose Transport

Infiniband Message & Connection Concepts

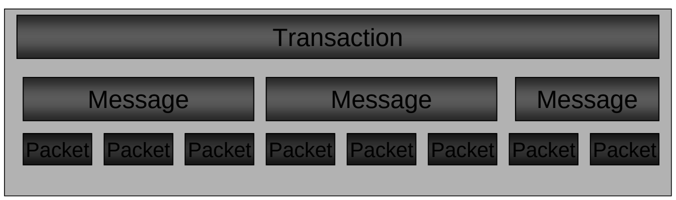
- Message Types
  - Send / Receive
  - Remote Direct Memory Access (RDMA)
    - Reads
    - Writes
  - Optional
    - Atomic (Compare&Swap, Fetch&Add)
    - Multicast
- Connection Types
  - Reliable
    - Acknowledged
  - Unreliable
    - Unacknowledged
  - Connected
    - Specific queue pair relationship
  - Datagram
    - Indirect destination

Message



- Applications, drivers, devices, adapters, etc. execute transactions via logical units of work termed messages.
- Messages are moved via the InfiniBand™ technology transport layer.
- H/W-based memory / resource protection to prevent unauthorized access
- Message semantics supported are:
  - Memory - RDMA Read / Write
  - Channel - Send / Receive
  - Atomics (optional functionality)
  - Multicast (optional functionality)

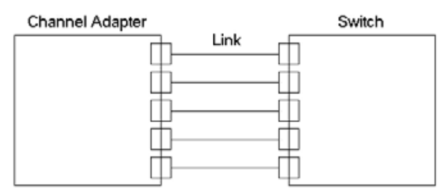
Packet



- InfiniBand™ technology H/W provides automatic message segmentation and re-assembly via packets
- End-to-end fabric unit of transfer
- Routable unit of transfer
- End-to-end unit of acknowledgment for reliable transport services
  - H/W-based timeouts and per QP packet sequence numbers
    - Sequence number space is large to support, e.g. long distance
  - End-to-end unit of recovery for hard failures, e.g. cable disconnect

InfiniBand Link

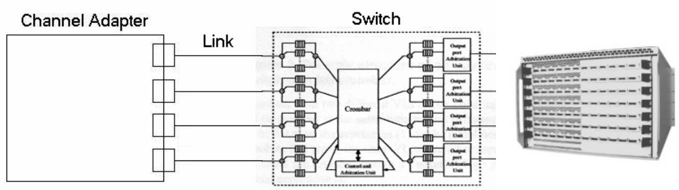
- Link
  - Full duplex
  - Point-to-point
  - 1-bit (1X) – 2.5Gbps
  - 4-bit (4x) – 10Gbps
  - 12-bit (12x) – 30Gbps
- Copper Cable, Optical Fiber, and printed circuit.
- Vendor Specific.
- May be used in Parallel.



InfiniBand Switch

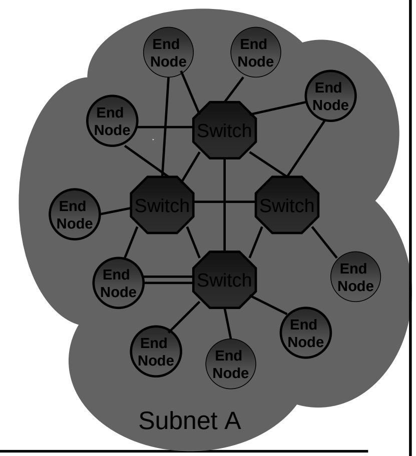
- Switch
  - Local network routing
  - Implements up to 255 ports
  - Local Route Header
    - 16-bit Source Local ID
    - 16-bit Destination Local ID

Port 0
Forwarding table
Service Level to Virtual Lane mapping table
Non-multiplexed
Full-multiplexed



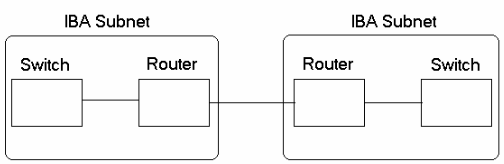
Infiniband Switch

- Switches route packets only within a single subnet to reduce complexity and solution cost.
- Optional switch-based multicast for unreliable, datagram service



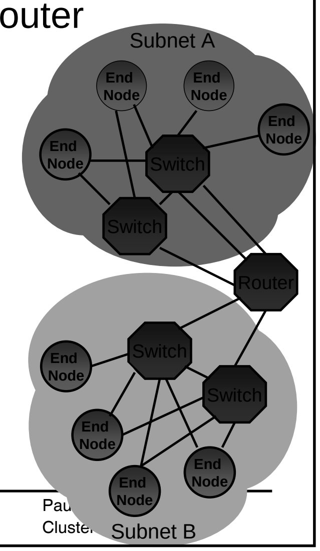
InfiniBand Router

- Router
  - Outside local network routing
  - Implements up to 255 ports
  - IBA, WAN, and MAN
  - Global Route Header
    - 128-bit Source Global ID
    - 128-bit Destination Global ID



Infiniband Router

- Superset of InfiniBand™ technology Switch functionality.
- Provides routing between subnets.
  - Subnetting improves scalability, management, etc.
  - Subnets leverage IP subnet architecture / concepts
- Independent Hardware Vendors may provide a variety of value-add solutions across a wide range of price / performance points.
- Routers may join InfiniBand technology fabric instances to same or disparate fabrics (disparate support is optional)
  - Disparate fabric support allows InfiniBand™ technology subnets to be joined using alternative, intermediate fabrics.
  - Optional multiple protocol support via raw packet over a consolidated data center fabric to remote end nodes.



InfiniBand Target Channel Adapter

- Target Channel Adapter (TCA)
  - Network Interface Card
  - Connects I/O devices
  - Subnet manager
    - Assigns Local IDs

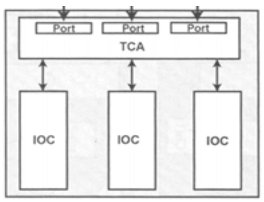
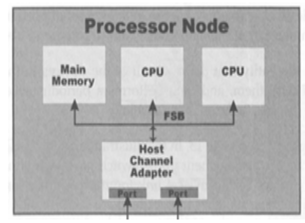
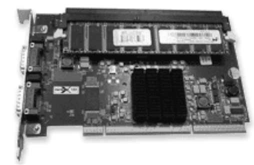


Diagram from [16]

InfiniBand Host Channel Adapter

- Host Channel Adapter
  - Network Interface Card
  - Connects Processors
  - May have memory (0-512MB)
  - Subnet manager agent
    - Assigns Local IDs
    - Location of subnet manager





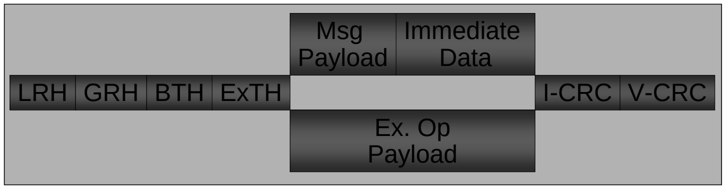
Infiniband Message-level Flow Control

- Prevents a transport-level “Receive Buffer Empty” NAK.
- Message Level Flow control is invoked for:
  - Sends to a Queue Pair with no Receive Buffer Posted
  - Sends or RDMA accesses to memory that is paged out.
- Reliable Connection:
  - ReceiverNotReady & ReceiverReady frames pace the flow of data.
  - Flow control is invisible to the client.
- Reliable Datagram:
  - ReceiverNotReady NAK indicates message wasn’t received.
  - Flow control is visible to the client.
- Unreliable Datagram & Raw Packet:
  - No flow control.

Infiniband Link Characteristics: Addressing

- GUID (Globally Unique Identifier) - Each TCA, HCA, Switch, and Router has a single unique GUID.
- Local ID (LID) - Subnet-unique, 16-bit ID used to identify and route a packet to an endnode port within a single subnet.
- IP Address - Global-unique 128-bit IPv6 ID used to identify an endnode by applications and to route packets between subnets.
  - Each switch and router has one or more LIDs and IPv6 addresses that are used when it is the destination endnode for management messages.

Packet Header Ordering



Red fields are only present based on operation type and destination subnet

- LRH - Local Route Header
- GRH - Global Route Header
- BTH - Base Transport Header
- ExTH - Extended Transport Header
- Msg Payload - Message Payload
- Immediate Data Header
- Ex Op Payload - Extended Op Payload
- I-CRC - Invariant CRC (32-bit)
- V-CRC - Variant CRC (16-bit)

Data Integrity

- **I-CRC - Invariant CRC (32-bit)**
  - covers only the fields that do not change from hop to hop
- **V-CRC - Variant CRC (16-bit)**
  - provides link-level data integrity between two hops
- **InfiniBand includes the I-CRC so that if a bit error is introduced within a device, the error will always be detected.**

Local Route Header

VL(4)	Vers (4)	NH(2)	Rsv(6)	Destination LID (16)
Rsv (5)	Pkt Len (11)			Source LID (16)

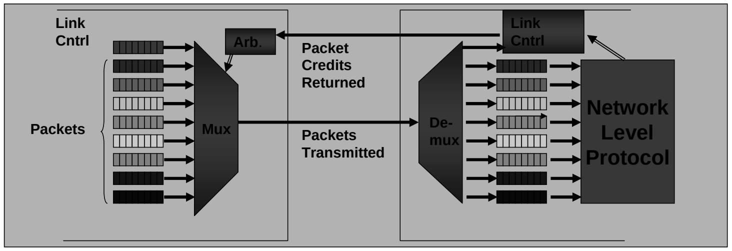
- **LRH used to route packets within subnets.**
- **VL - Actual Virtual Lane Used**
- **Vers - LRH Version**
- **NH - Next Header - indicates next header**
  - InfiniBand™ **technology transport, GRH, IPv6 (raw), Ethertype (raw)**
- **Destination LID - Subnet Local Destination Address Identifier**
- **Rsv - Reserve Field**
- **Pkt Len - Packet Length (multiple of 4 bytes - up to 8KB)**
- **Source LID - Subnet Local Source Address Identifier**

Global Route Header

Vers (4)	TC (8)	Flow ID (20)	Pkt Len (16)	NHdr (8)	Hop (8)
Source IPv6 Address (128)					
Destination IPv6 Address (128)					

- GRH used to route packets between subnets.
- GRH is defined per RFC 2460. This is the base IPv6 header.
- Present in all packets if the LRH Next Header = GRH.
- Each end node shall be assigned an unique IPv6 address
- Applications target an end node by its IPv6 address
  - An end node may be assigned multiple IPv6 addresses

Infiniband Link Level Flow Control



- Credit-based link-level flow control
- Link Receivers grant packet receive buffer space credits per VL
- Multiple Virtual Lanes (VLs) on each Physical link provide:
  - Priority arbitration, with VLs assigned priority scheme.
  - Alleviation of head-of-line blocking.

Infiniband Link Characteristics

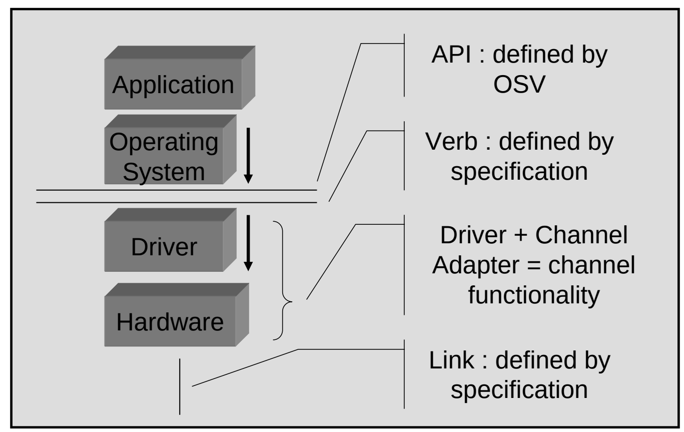
- 2.5 Gbaud signaling rate
  - Auto-negotiation for future higher speed signaling
- All links full-duplex
- 1, 4, and 12 wide link widths
- Common backplane connector(s)
  - With auto-negotiation to mutually acceptable width

Link Width	Signaling Rate	Bandwidth	
		Unidirectional	Bi-directional
1	2.5 Gbaud	250 MBytes/s	500 MBytes/s
4	2.5 Gbaud	1 GByte/s	2 GByte/s
12	2.5 Gbaud	3 GBytes/s	6 GBytes/s

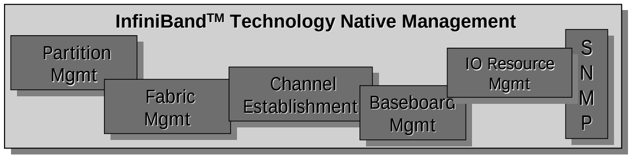
Infiniband Software Characteristics Overview

- Software interface to transport layer
  - Verbs specification
- Major specification topics
  - Partitioning
  - Work Request management
  - Memory management

Infiniband Verb layer

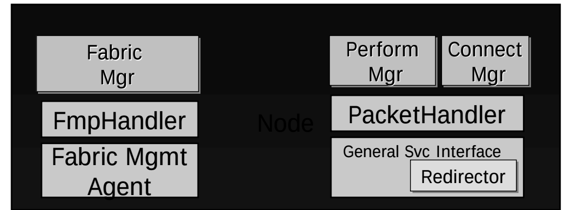


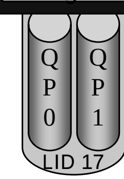
Management Characteristics Infrastructure



- Fabric Management - Fabric Topology & Fabric health
- Partition Management - Communication policy enforcement
- Channel Establishment - Connection primitives
- IO Resource Management - Device enumeration, device mgmt
- Baseboard Mgmt - Enabling instrumentation in-band
- SNMP - Native support of SNMP

Communications are "well known"





- QP 0 is Fabric Mgmt
  - VL 15, QP 0 is reserved for fabric management
- QP 1 is General Service Interface
  - GSI can forward packets to interested applications via packet handler interface
  - Can also send redirect response to originator of MAD to refer to preferred QP

Subnet Manager

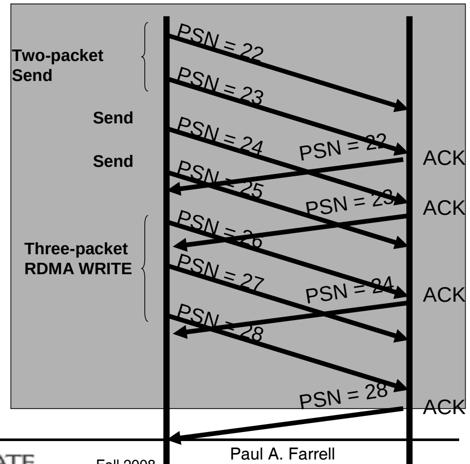
- SM can reside on any port of endnode, switch or router and can be implemented in hardware or in software.
- Discovers the subnet topology, assigns LIDs, GIDs, establishes pathways, and ensures its continued operation.
- All channel adapters and switches must implement an Subnet Management Agent (SMA) for handling communication with the subnet manager.
- There must be at least one subnet manager present to manage initialization and to reconfigure the subnet when a link goes up or down.
- An arbitration scheme selects a master subnet manager and other subnet managers operate in standby mode.

Infiniband Transport Characteristics Services

- When a QP (Queue Pair) is created it is set to provide one of the following services:
  - Reliable Connection
  - Unreliable Connection
  - Reliable Datagram
  - Unreliable Datagram
  - Multicast (optional)
  - Raw Packet (optional)
- Definition: "Reliable"
  - HW generates acknowledgments for every packet.
  - HW generates / checks packet sequence numbers
  - HW rejects duplicates, detects missing packets
  - Client transparent recovery from most fabric level errors.

Infiniband Sample Packet Exchange - Reliable Transports

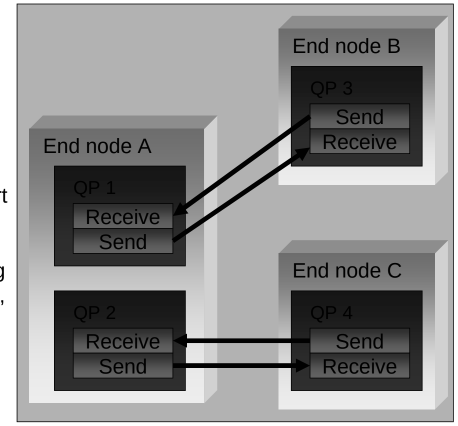
- H/W generated ACKs
- Seq. # within ACK
- Cumulative ACKs



Infiniband Reliable Connection

One-to-one QP relationship between source end node and destination end node.

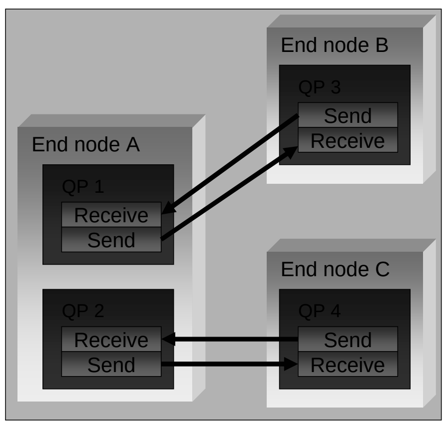
Provides:
- RDMA Support
- Send / Receive Support
- Atomic Support
- H/W generated ACKs
- Strong Packet Ordering
- Arbitrary message size, i.e. multi-packet messages



Infiniband Unreliable Connection

One-to-one QP relationship between source end node and destination end node.

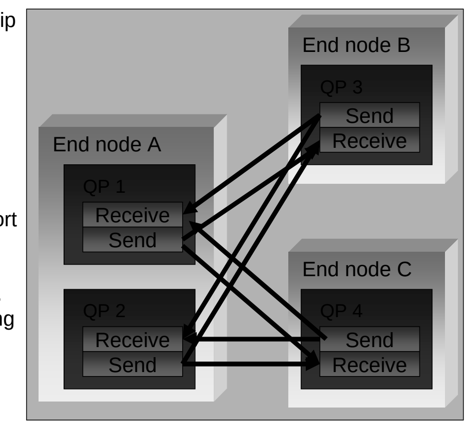
Provides:
- No RDMA Read Support
- Send / Receive Support
- No Atomic Support
- No H/W generated ACKs
- Source provides incrementing sequence numbers - no packet ordering guaranteed at the destination.
- Arbitrary message size, i.e. multi-packet messages



Infiniband Reliable Datagram

One-to-many QP relationship between source end node and destination end node. Optional H/W implementation.

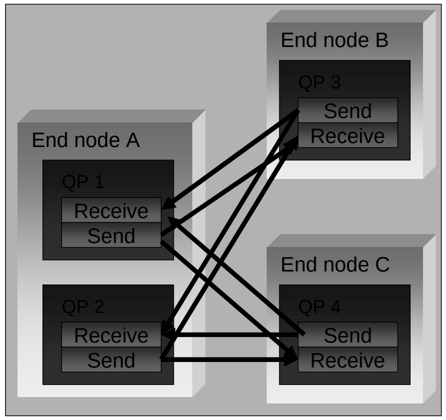
Provides:
- RDMA R / W Support
- Send / Receive Support
- Optional Atomic Support
- H/W generated ACKs.
- Strong Packet Ordering between any two end nodes
- QP Scalability
- Limited message size



Infiniband Unreliable Datagram

One-to-many QP relationship between source end node and destination end node.

Provides:
- No RDMA Support
- Send / Receive Support
- No Atomic Support
- No H/W generated ACKs
- No Packet Ordering
- Good QP Scalability
- Limited message size



Infiniband Unreliable Multicast

One-to-many QP relationship between source and destinations end nodes. Optional functionality.

Provides:
- Automatic packet replication within switches and routers - reduces number of packets injected into the subnet
- Send / Receive Support
- No RDMA Support
- No Atomic Support
- No H/W generated ACKs
- No Packet Ordering between end nodes
- Limited message size