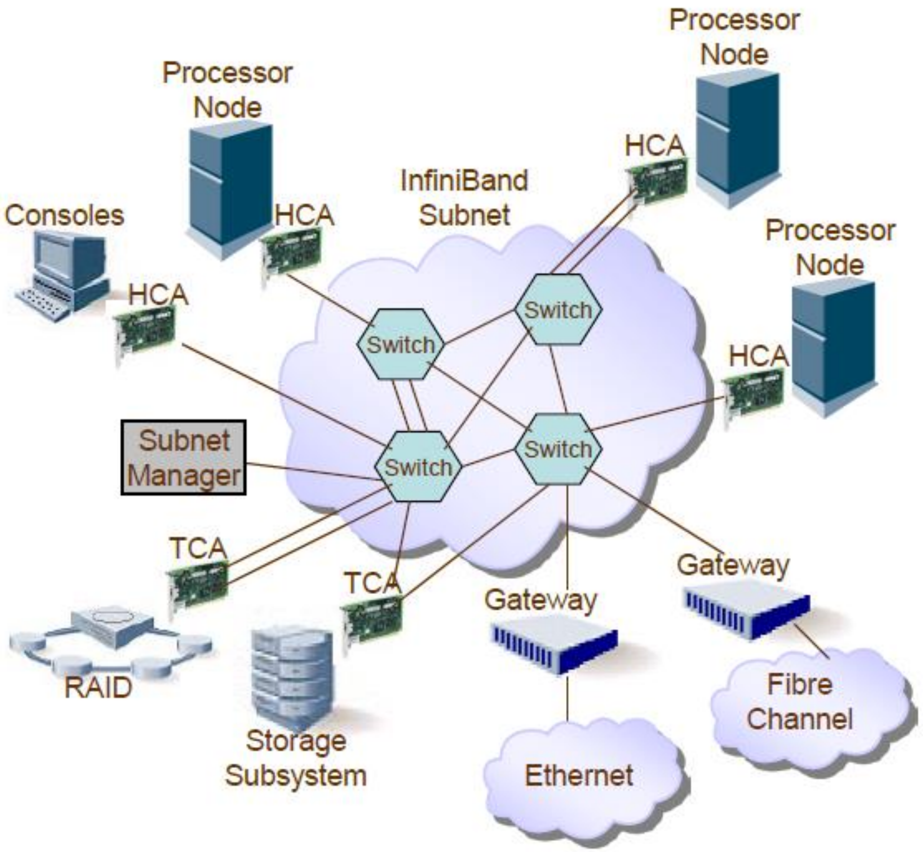
InfiniBand Technology

What is InfiniBand?

- Industry standard defined by the InfiniBand Trade Association (IBTA)
  - Originated in 1999
- Input/output architecture used to interconnect servers, communications infrastructure equipment, storage and embedded systems
- Pervasive, low-latency, high-bandwidth interconnect which requires low processing overhead and is ideal to carry multiple traffic types (clustering, communications, storage, management) over a single connection.
- As a mature and field-proven technology, InfiniBand is used in thousands of data centers, high-performance compute clusters and embedded applications that scale from small scale to large scale

The InfiniBand Architecture

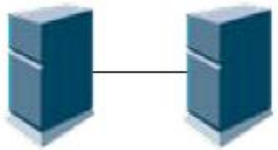
- Defines System Area Network architecture
- Architecture supports
  - Host Channel Adapters (HCA)
  - Target Channel Adapters (TCA)
  - Switches
  - Routers
- Facilitated HW design for
  - Low latency / high bandwidth
  - Transport offload



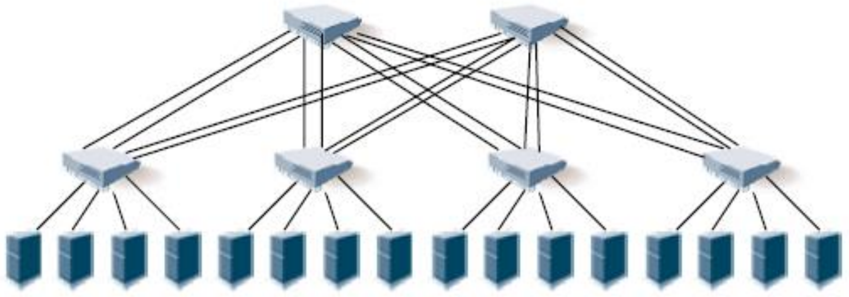
InfiniBand Feature Highlights

- Serial high-bandwidth, ultra-low-latency links
- Reliable, lossless, self-managing fabric
- Full CPU offload
- Quality Of Service
- Cluster scalability, flexibility and simplified management

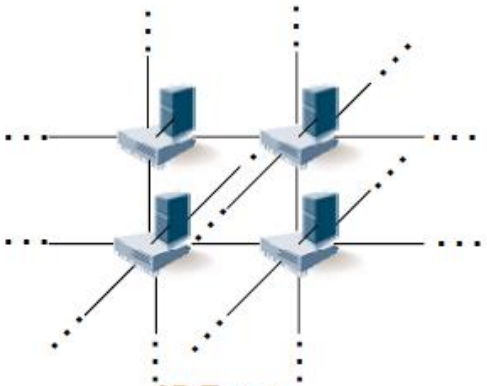
InfiniBand Topologies



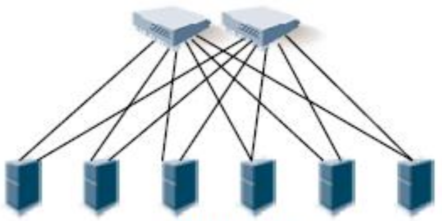
Back to Back



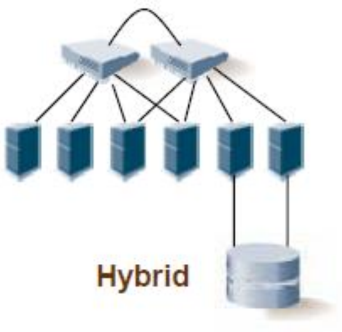
2 Level Fat Tree



3D Torus



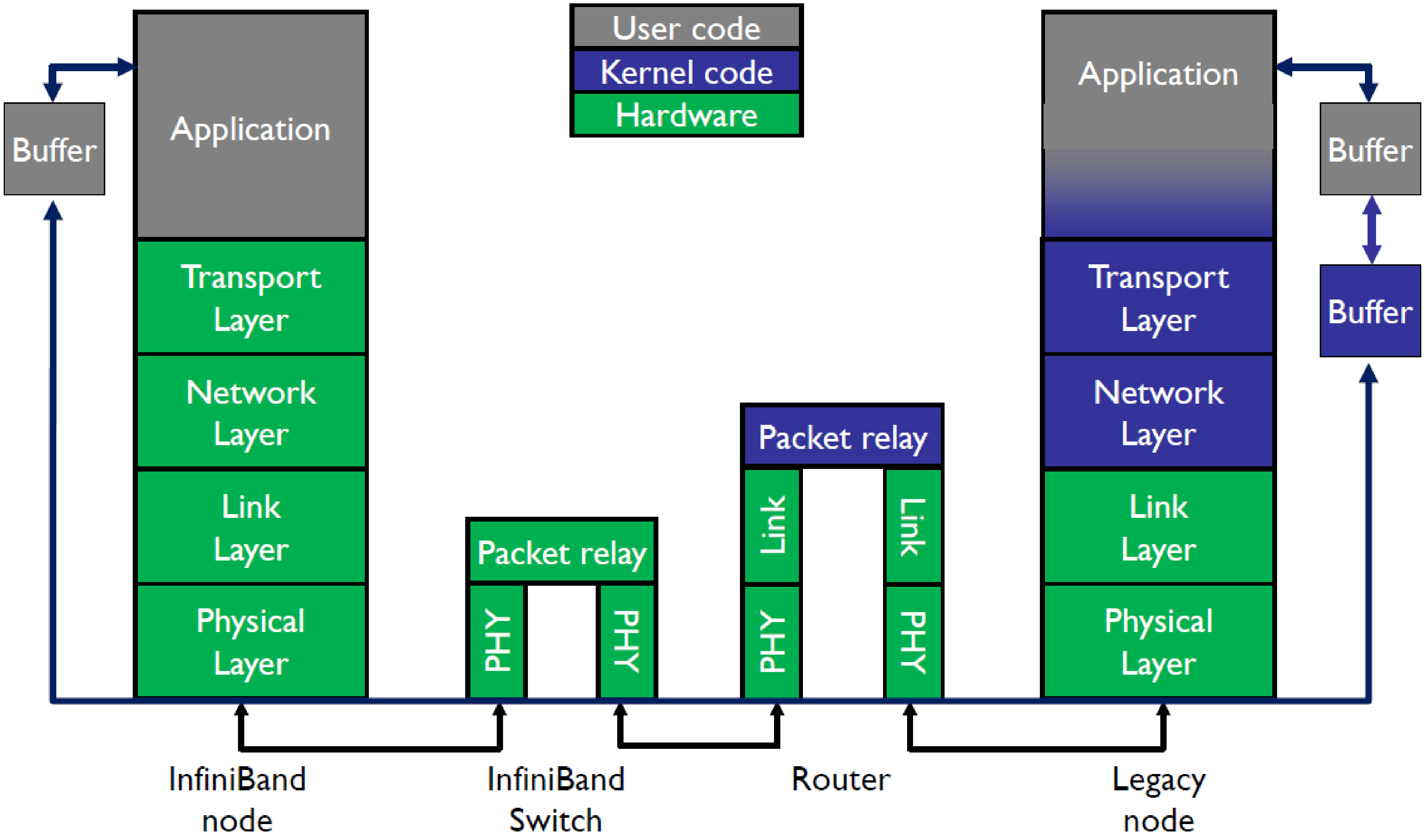
Dual Star



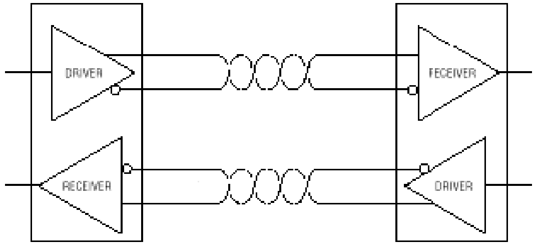
Hybrid

- Example topologies commonly used
- Architecture does not limit topology
- Modular switches are based on fat tree architecture

InfiniBand Network Stack



Physical Layer



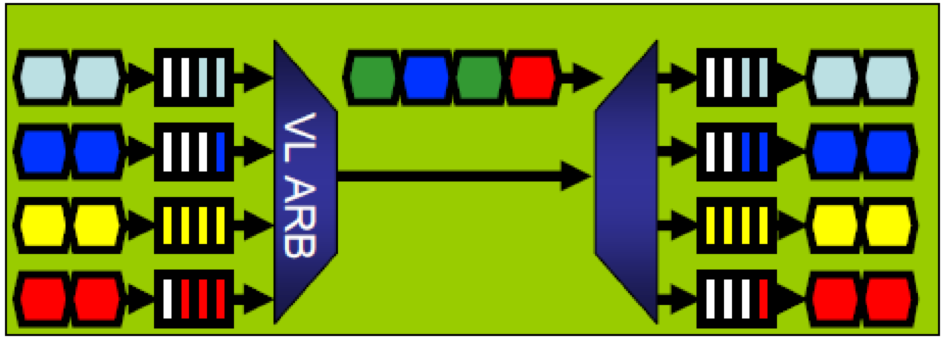
- ❐ Data transfer over serial bit streams
- ❐ Auto-negotiation of link speed and width
- ❐ Power management
- ❐ Bit encoding
- ❐ Control symbols

Link Speed (10^9 bit/sec)

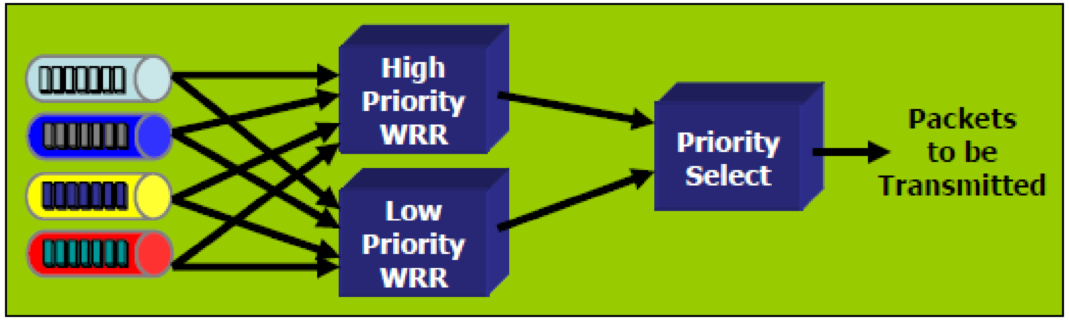
Lane Speed → Link Width ↓	SDR (2.5GHz)	DDR (5GHz)	QDR (10GHz)	FDR (14GHz)	EDR (25GHz)
1X	2.5	5	10	14	25
4X	10	20	40	56	100
8X	20	40	80	102	200
12X	30	60	120	168	300

Link Layer

- Addressing and Switching
  - Local Identifier (LID) addressing
  - Unicast LID - 48K addresses
  - Multicast LID – up to 16K addresses
  - Efficient linear lookup
  - Cut through switching supported
  - Multi-pathing support through LMC
- Independent Virtual Lanes
  - Flow control (lossless fabric)
  - Service level
  - VL arbitration for QoS
- Data Integrity
  - Invariant CRC
  - Variant CRC



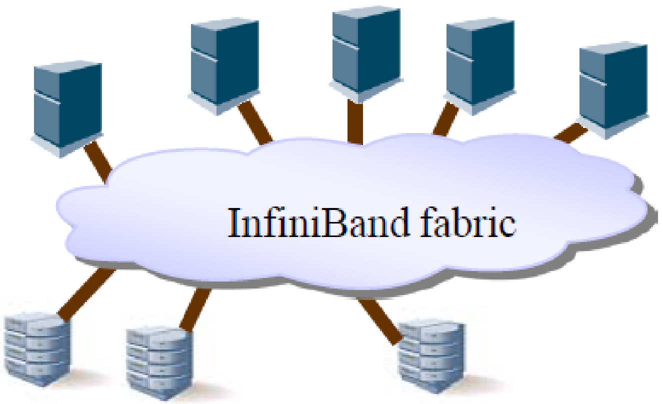
Independent Virtual Lanes (VLs)



H/L Weighted Round Robin (WRR) VL Arbitration

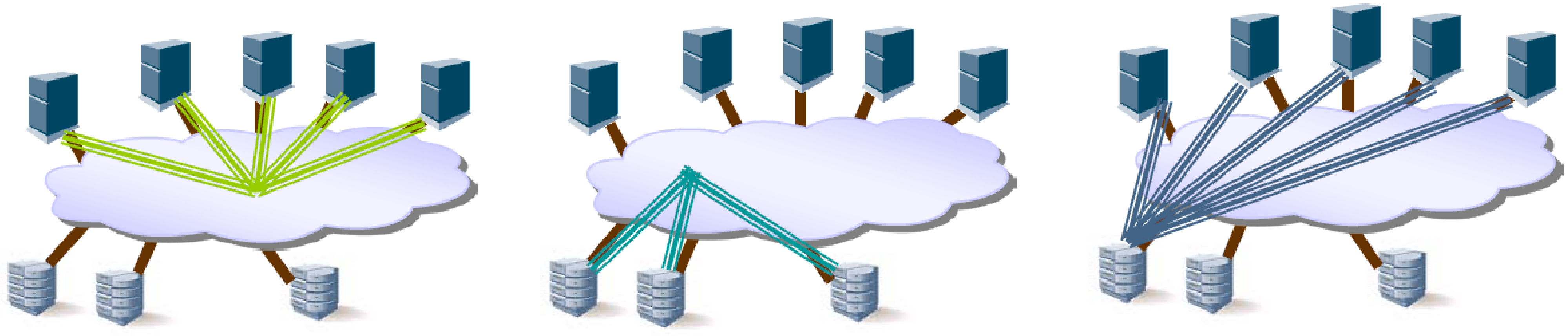
Virtual Lanes and Scheduling

Physical:



- Dynamically configure and adjust VLs and scheduling to match application performance needs

Logical:



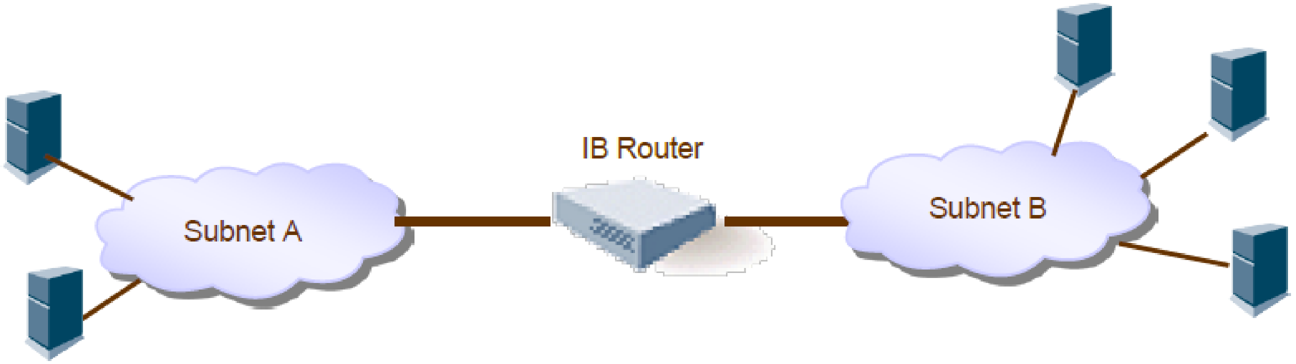
Low-latency VL for clustering

Backup VL
Day: ≥ 20% BW
Night: ≥ 60% BW

Mainstream storage VL
Day: ≥ 40% BW
Night: ≥ 20% BW

Network Layer

- Global Identifier (GID) addressing
  - Based on IPv6 addressing scheme
  - GID = {64 bit GID prefix, 64 bit GUID}
    - GUID = Global Unique Identifier (64 bit EUI-64)
    - GUID 0 – assigned by the manufacturer
    - GUID 1..(N-1) – assigned by the subnet manager
- Used for multicast distribution within end nodes



Transport Layer

- Queue Pair (QP) – transport endpoint
  - Asynchronous interface
    - Send Queue, Receive Queue, Completion Queue
  - Full transport offload
    - Segmentation, reassembly, timers, retransmission, etc…
- Kernel bypass
  - Enables low latency and CPU offload
  - Exposure of application buffers to the network
- Polling and interrupt models supported

InfiniBand Packet Format

8B	40B	12B	var	0..4096B	4B	2B
LRH	GRH	BTH	Ext HDRs	Payload	ICRC	VCRC

InfiniBand Data Packet

VL	LVer	SL	rsvd	LNH	DLID
rsvd	Len				SLID

LRH L2-Local Route Header

Opcode	S	M	Pad	TVer	Partition Key
rsvd	Destination QP				
A	rsvd	PSN			

BTH L4-Base Transport Header

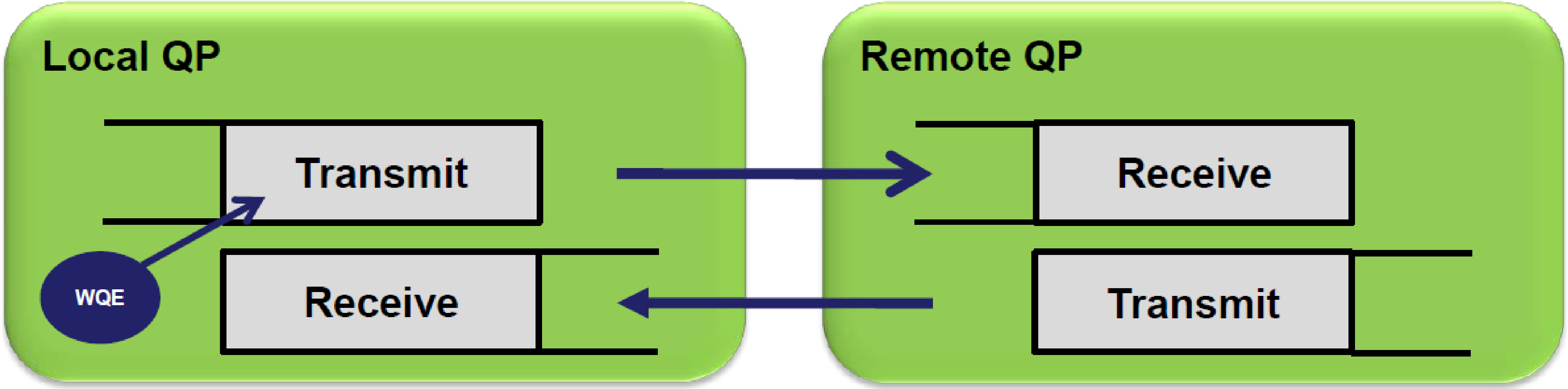
IPVer	TClass	Flow Label
Payload Len	Next Header	Hop Lim
SGID[127:96]		
SGID[95:64]		
SGID[63:32]		
SGID[31:0]		
DGID[127:96]		
DGID[95:64]		
DGID[63:32]		
DGID[31:0]		

GRH (Optional) L3-Global Route Header

Extended headers:

- Reliable Datagram ETH (4B)
- Datagram ETH (8B)
- RDMA ETH (16B)
- Atomic ETH (28B)
- ACK ETH (4B)
- Atomic ACK ETH (8B)
- Immediate Data ETH (4B)
- Invalidate ETH (4B)

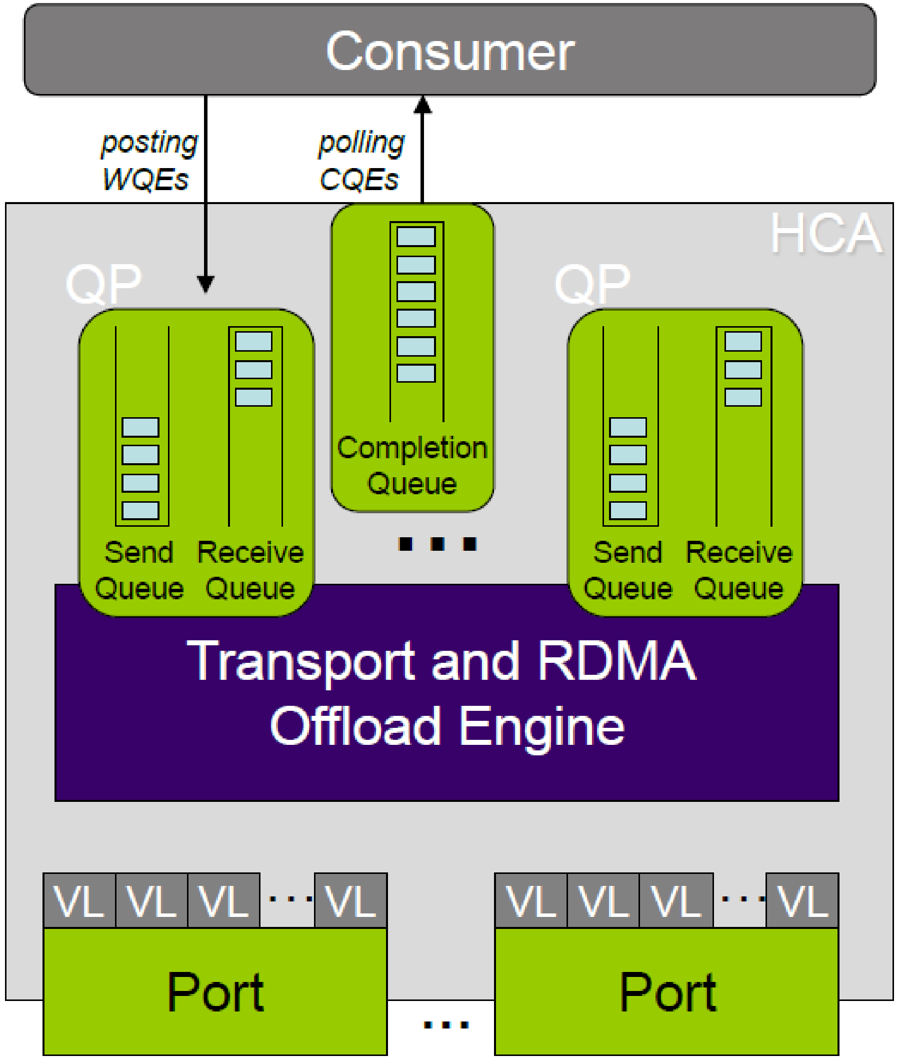
Transport Layer – Queue Pairs



- QPs are in pairs (Send/Receive)
- Work Queue is the consumer/producer interface to the fabric
- The consumer/producer initiates a Work Queue Element (WQE)
- The channel adapter executes the work request
- The channel adapter notifies on completion or errors by writing a Completion Queue Element (CQE) to a Completion Queue (CQ)

Transport – HCA Model

- Asynchronous interface
  - Consumer posts work requests
  - HCA processes
  - Consumer polls completions
- Transport executed by HCA
- I/O channel exposed to the application



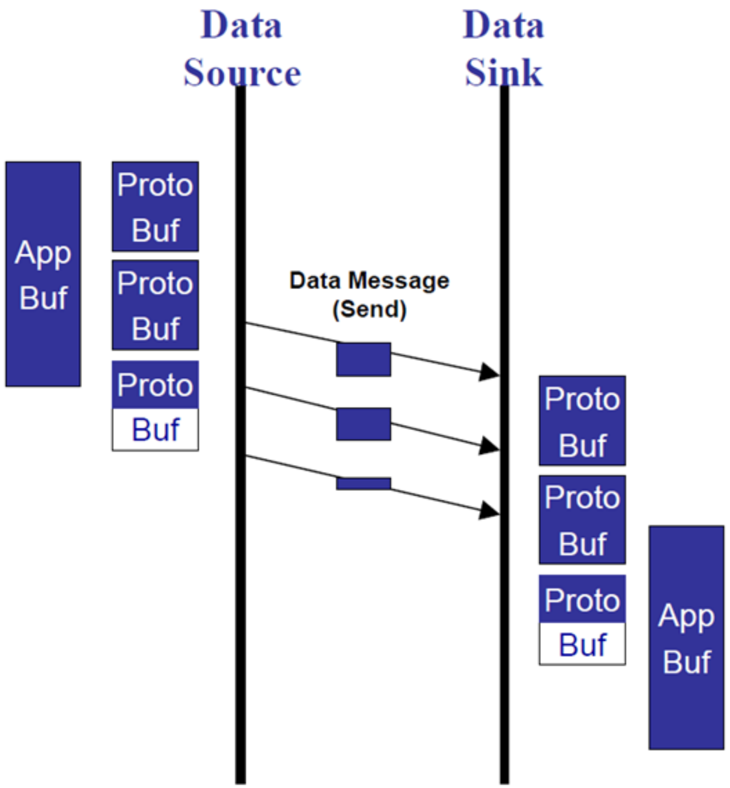
Transport Layer – Types Transfer Operations

- SEND
  - Read message from HCA local system memory
  - Transfers data to responder HCA Receive Queue logic
  - Does not specify where the data will be written in remote memory
  - Immediate Data option available

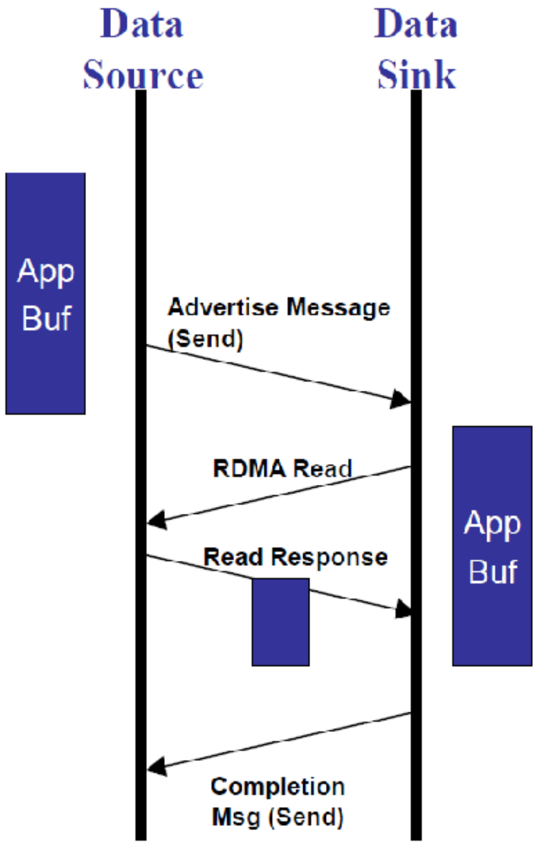
- RDMA Read
  - Responder HCA reads its local memory and returns it to the requesting HCA
  - Requires remote memory access rights, memory start address and message length

- RDMA Write
  - Requester HCA sends data to be written into the responder HCA system memory
  - Requires remote memory access rights, memory start address and message length

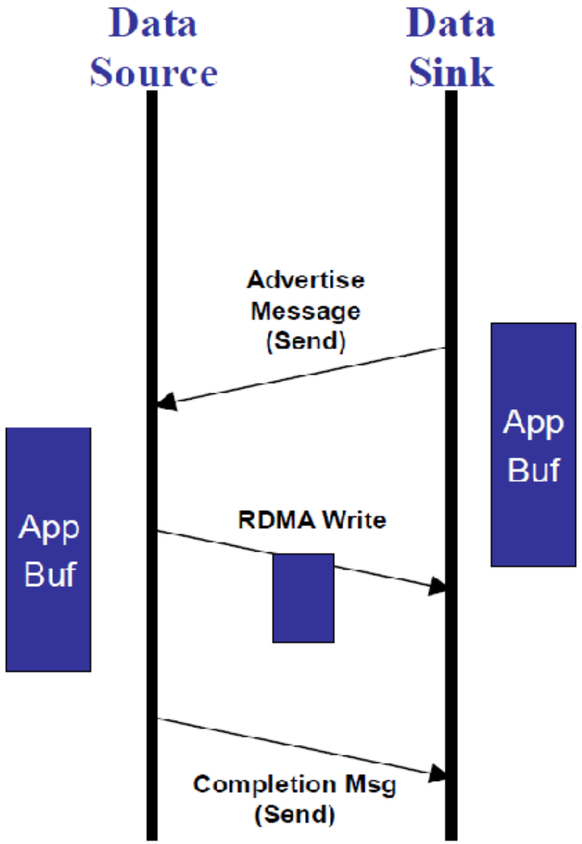
Typical buffer copy flow



Typical Zero-Copy Read flow

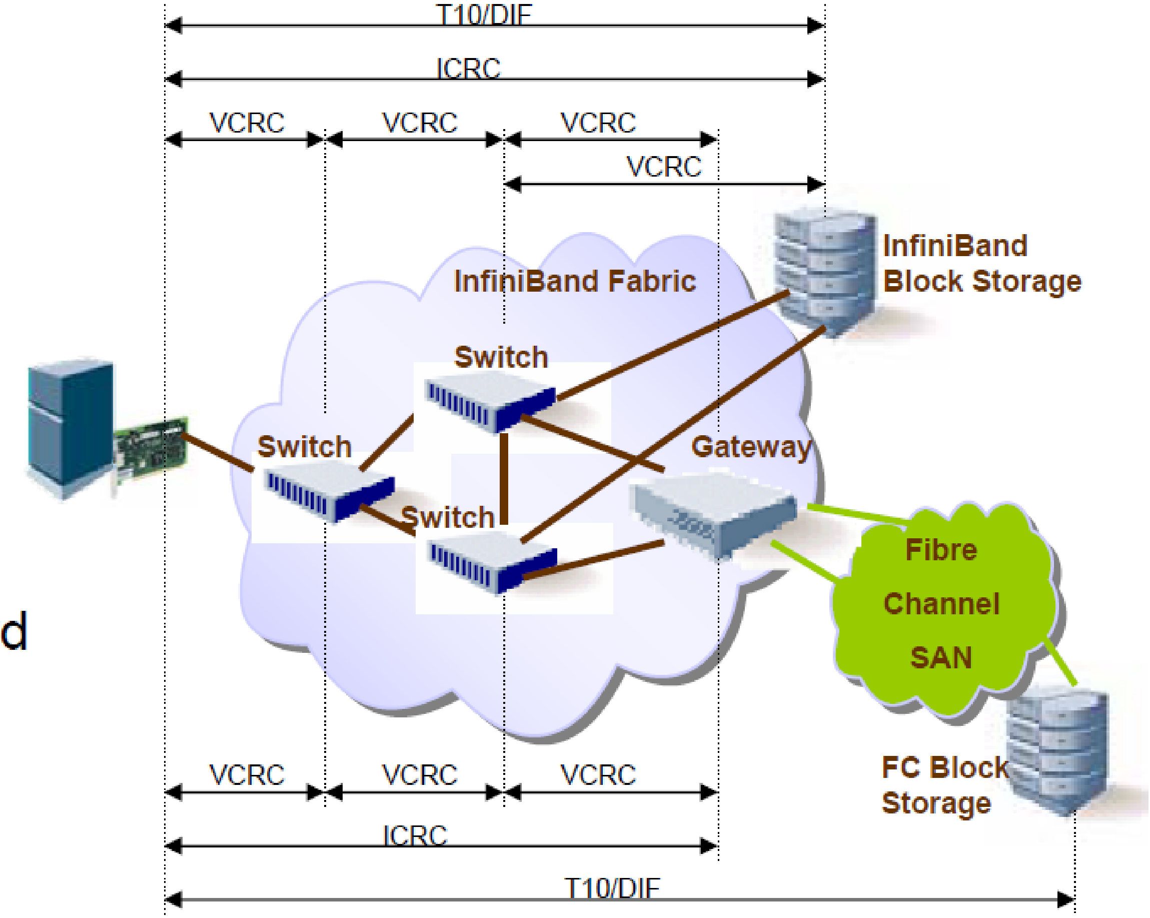


Typical Zero-Copy Write flow



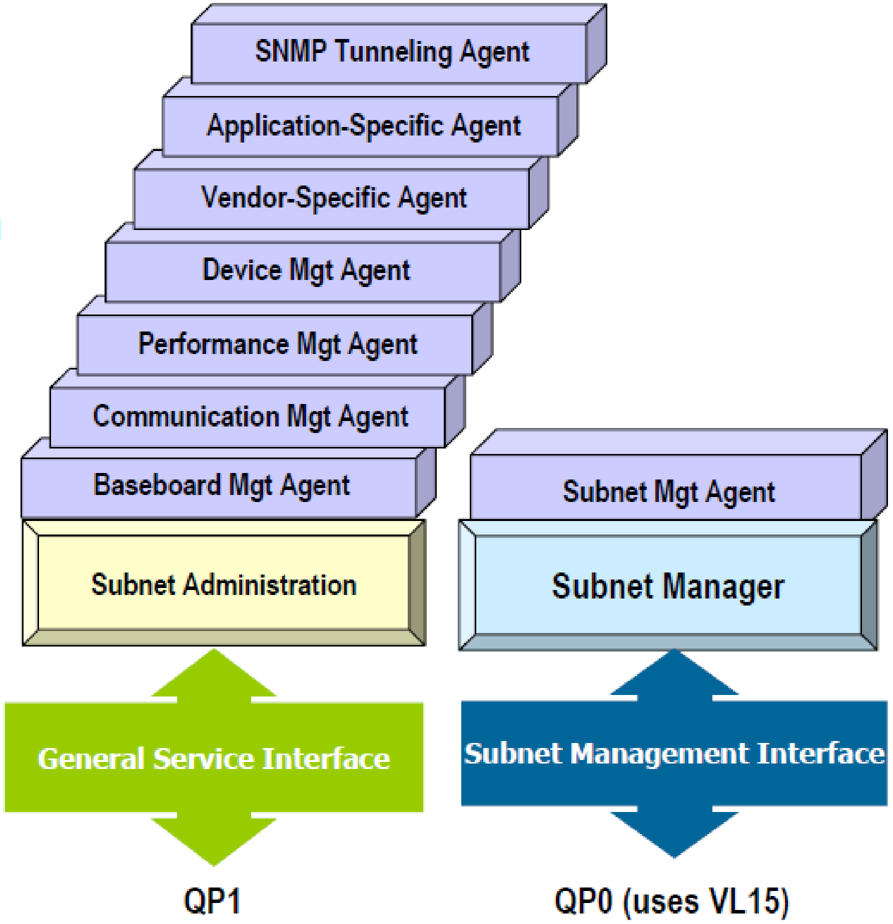
InfiniBand Data Integrity

- Hop by hop
  - VCRC – 16 bit CRC
  - CRC16 0x100B
- End to end
  - ICRC – 32 bit CRC
  - CRC32 0x04C11DB7
  - Same CRC as Ethernet
- Application level
  - T10/DIF Logical Block Guard
    - Per block CRC
  - 16 bit CRC 0x8BB7



Management Model

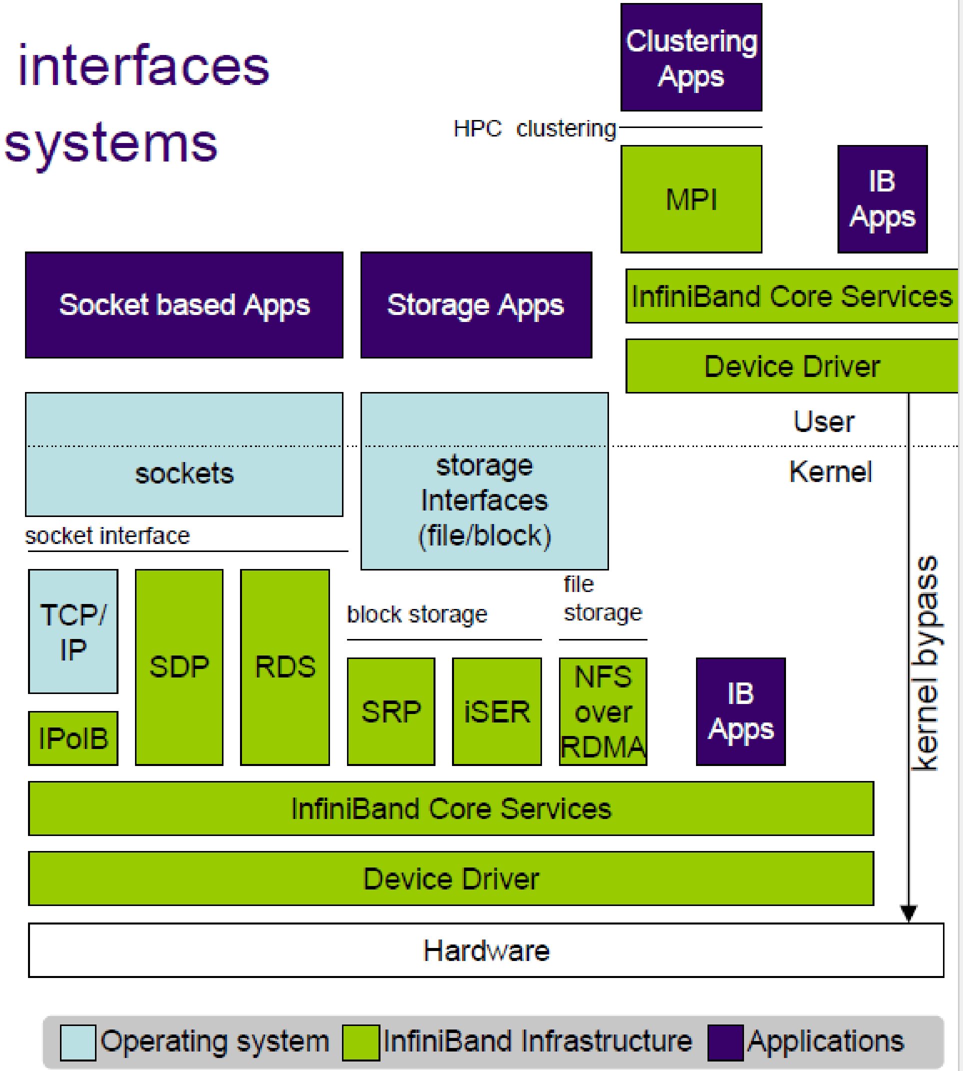
- Subnet Manager (SM)
  - Configures/Administers fabric topology
  - Implemented at an end-node or a switch
  - Active/Passive model when more than one SM is present
  - Talks with SM Agents in nodes/switches
- Subnet Administration
  - Provides path records
  - QoS management
- Communication Management
  - Connection establishment processing



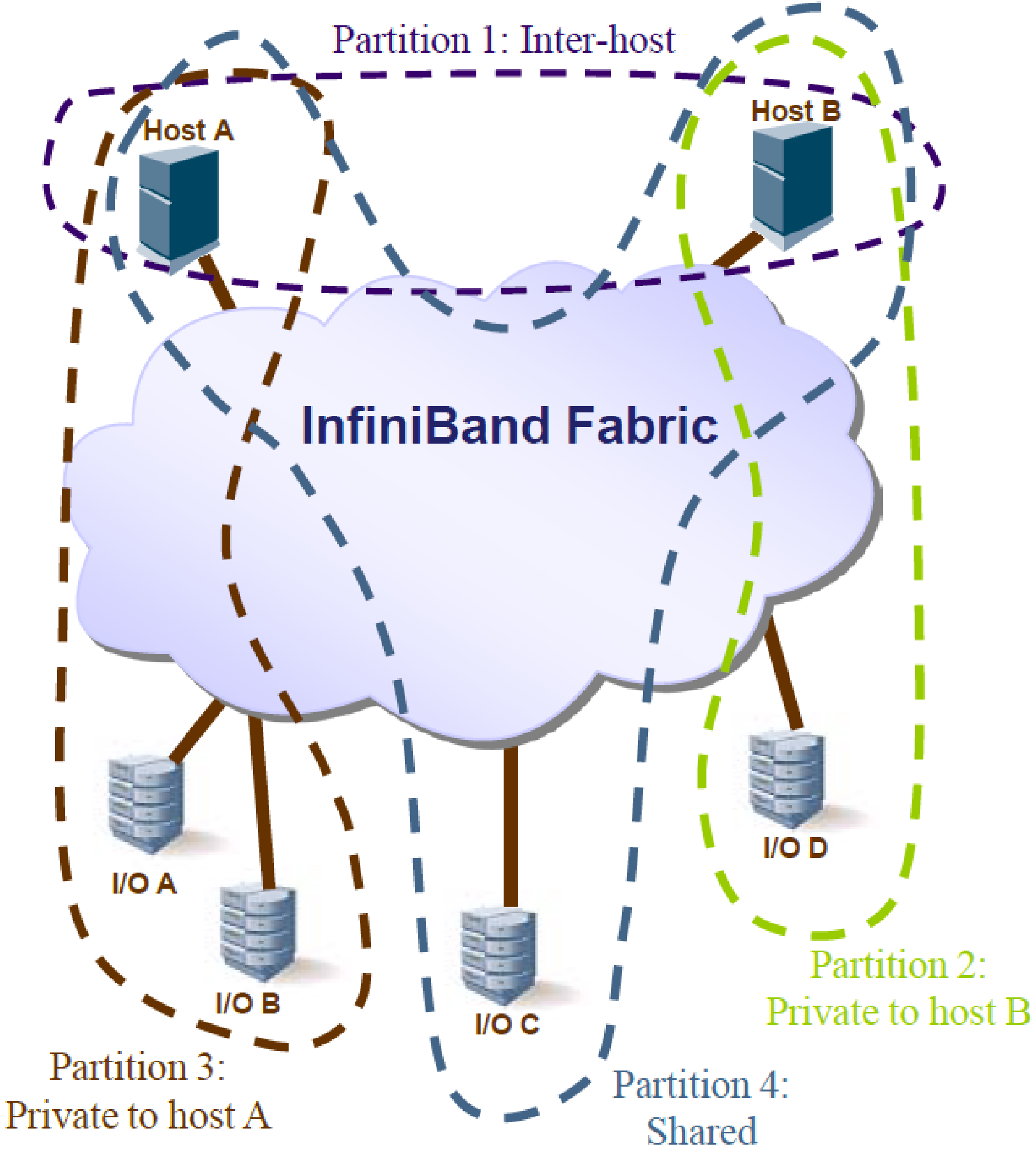
Upper Layer Protocols

- ULPs connect InfiniBand to common interfaces
- Supported on mainstream operating systems

- Clustering
  - MPI (Message Passing Interface)
  - RDS (Reliable Datagram Socket)
- Network
  - IPoIB (IP over InfiniBand)
  - SDP (Socket Direct Protocol)
- Storage
  - SRP (SCSI RDMA Protocol)
  - iSER (iSCSI Extensions for RDMA)
  - NFSoRDMA (NFS over RDMA)



Partitions



- Logically divide fabric into isolated domains
- Partial and full membership per partition
- Partition filtering at switches
- Similar to
  - FC Zoning
  - 802.1Q VLANs

High Availability and Redundancy

- Multi-port HCAs
- Redundant fabric topologies
- Link layer multi-pathing (LMC)
- Automatic Path Migration (APM)
- ULP High Availability
  - Application-level multi-pathing (SRP/iSER)
  - Teaming/bonding (IPoIB)

Glossary

- APM - Automatic Path Migration
- BECN - Backward Explicit Congestion Notification
- BTH - Base Transport Header
- CFM - Configuration Manager
- CQ - Completion Queue
- CQE - Completion Queue Element
- CRC - Cyclic Redundancy Check
- DDR - Double Data Rate
- DIF - Data Integrity Field
- FC - Fibre Channel
- FECN - Forward Explicit Congestion Notification
- GbE - Gigabit Ethernet
- GID - Global IDentifier
- GRH - Global Routing Header
- GUID - Globally Unique IDentifier
- HCA - Host Channel Adapter
- IB - InfiniBand
- IBTA - InfiniBand Trade Association
- ICRC - Invariant CRC
- IPoIB - Internet Protocol Over InfiniBand
- IPv6 - Internet Protocol Version 6
- iSER - iSCSI Extensions for RDMA
- LID - Local IDentifier
- LMC - Link Mask Control
- LRH - Local Routing Header
- LUN - Logical Unit Number
- MPI - Message Passing Interface
- MR - Memory Region
- NFSoRDMA - NFS over RDMA
- OSD - Object based Storage Device
- OS - Operating System
- PCIe - PCI Express
- PD - Protection Domain
- QDR - Quadruple Data Rate
- QoS - Quality of Service
- QP - Queue Pair
- RDMA - Remote DMA
- RDS - Reliable Datagram Socket
- RPC - Remote Procedure Call
- SAN - Storage Area Network
- SDP - Sockets Direct Protocol
- SDR - Single Data Rate
- SL - Service Level
- SM - Subnet Manager
- SRP - SCSI RDMA Protocol
- TCA - Target Channel Adapter
- ULP - Upper Layer Protocol
- VCRC - Variant CRC
- VL - Virtual Lane
- WQE - Work Queue Element
- WRR - Weighted Round Robin