Introduction to High-Speed InfiniBand Interconnect

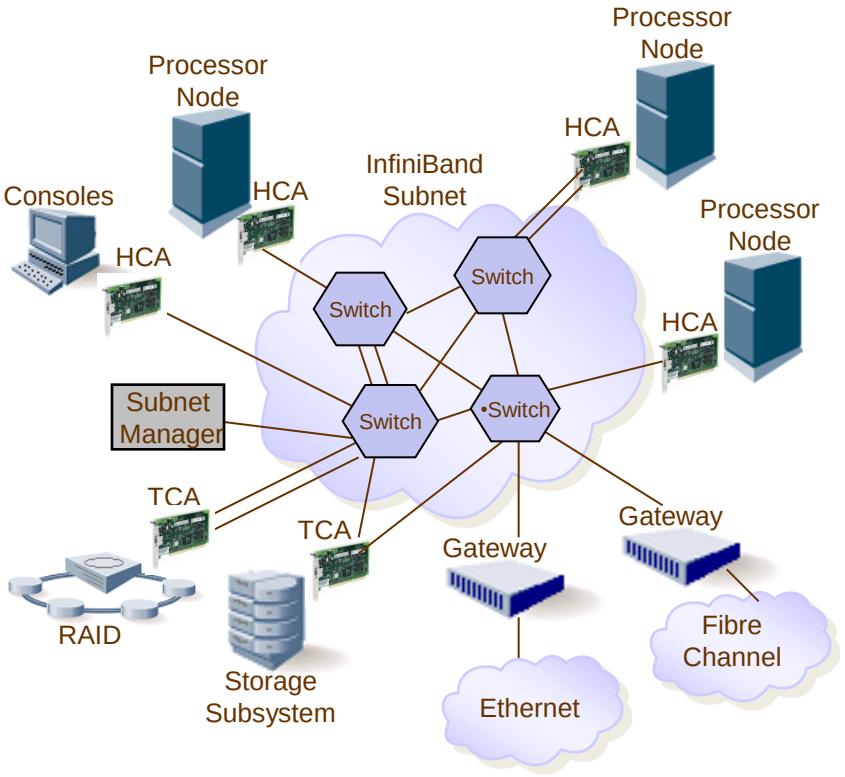
What is InfiniBand?

- **Industry standard defined by the InfiniBand Trade Association**
  - Originated in 1999
- **InfiniBand™ specification defines an input/output architecture used to interconnect servers, communications infrastructure equipment, storage and embedded systems**
- **InfiniBand is a 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**

Source: InfiniBand® Trade Association (IBTA) www.infinibandta.org

The InfiniBand Architecture

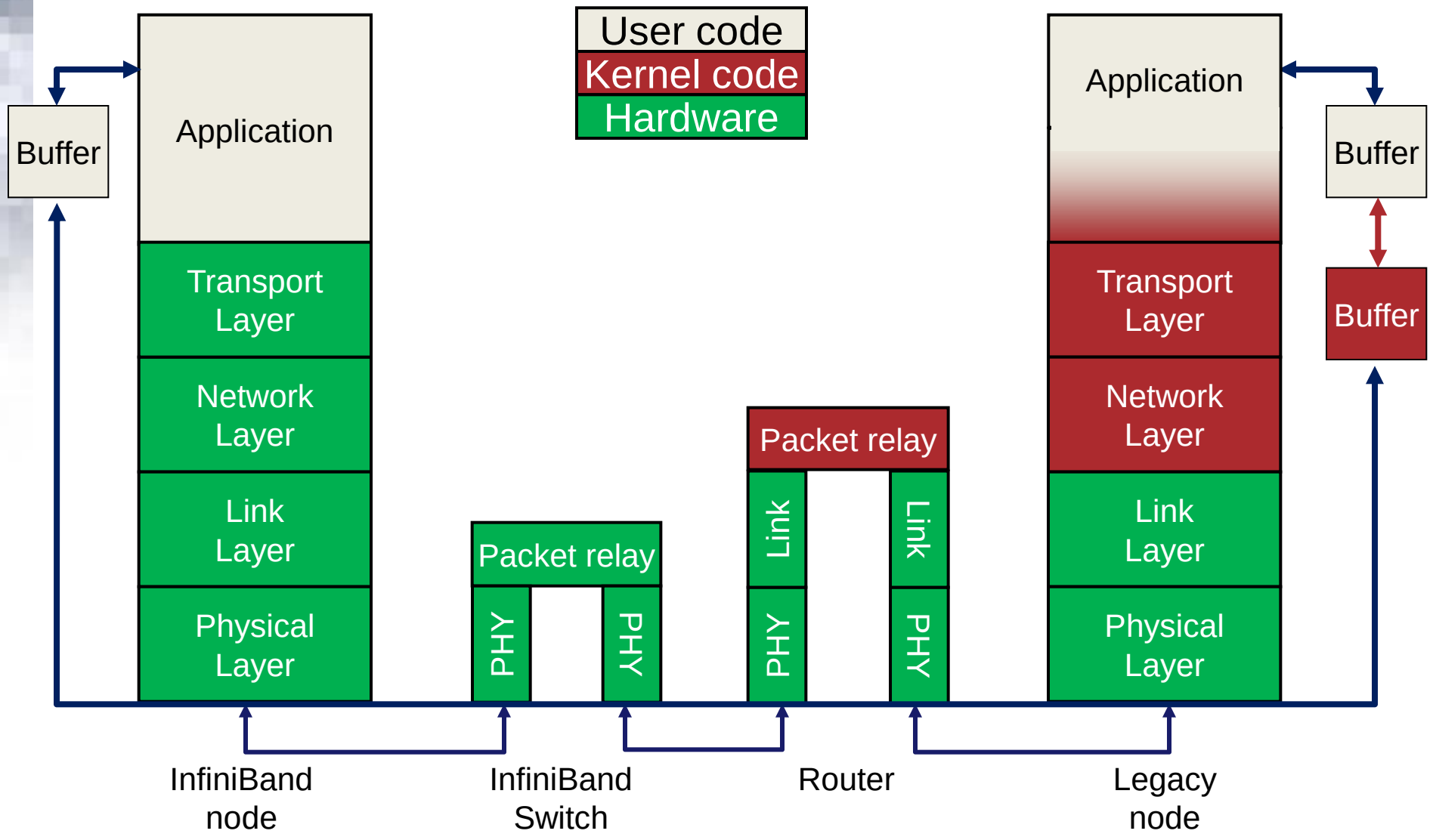
- **Industry standard defined by the InfiniBand Trade Association**
- **Defines System Area Network architecture**
  - Comprehensive specification: from physical to applications
- **Architecture supports**
  - Host Channel Adapters (HCA)
  - Target Channel Adapters (TCA)
  - Switches
  - Routers
- **Facilitated HW design for**
  - Low latency / high bandwidth
  - Transport offload



InfiniBand Highlights

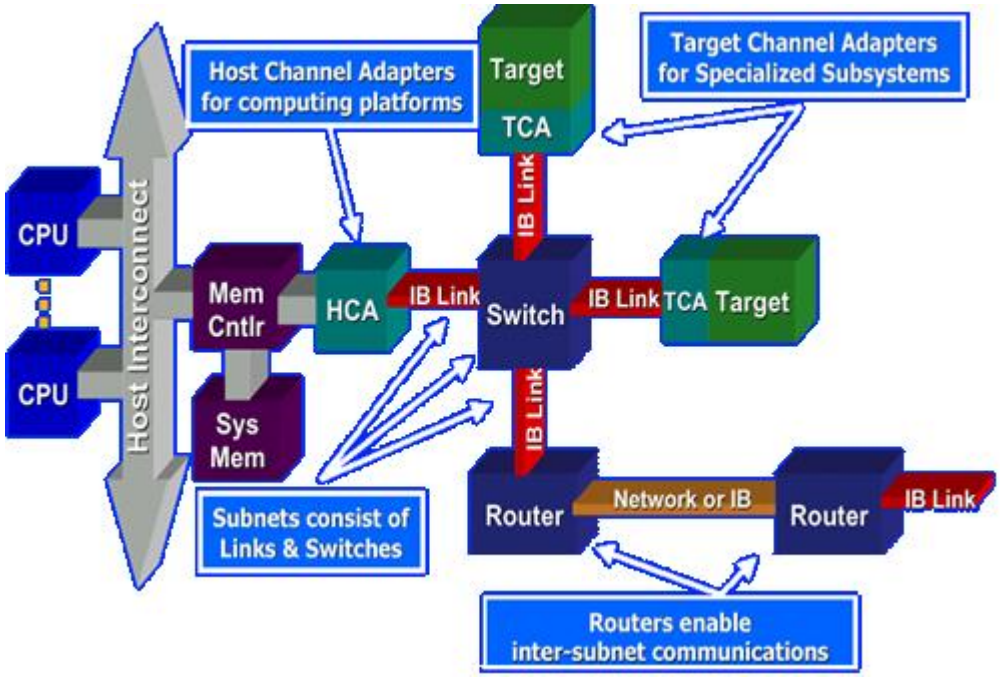
- **Serial High Bandwidth Links**
  - SDR: 10Gb/s HCA links
  - DDR: 20Gb/s HCA links
  - QDR: 40Gb/s HCA links
  - FDR: 56Gb/s HCA links
  - EDR: 100Gb/s HCA links
- **Ultra low latency**
  - Under 1 us application to application
- **Reliable, lossless, self-managing fabric**
  - Link level flow control
  - Congestion control to prevent HOL blocking
- **Full CPU Offload**
  - Hardware Based Reliable Transport Protocol
  - Kernel Bypass (User level applications get direct access to hardware)
- **Memory exposed to remote node access**
  - RDMA-read and RDMA-write
  - Atomic operations
- **Quality Of Service**
  - Independent I/O channels at the adapter level
  - Virtual Lanes at the link level
- **Cluster Scalability/flexibility**
  - Up to 48K nodes in subnet, up to 2^{128} in network
  - Parallel routes between end nodes
  - Multiple cluster topologies possible
- **Simplified Cluster Management**
  - Centralized route manager
  - In-band diagnostics and upgrades

InfiniBand Network Stack



InfiniBand Components Overview

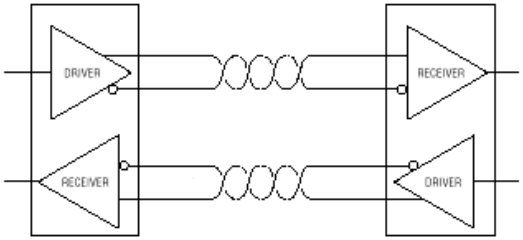
- **Host Channel Adapter (HCA)**
  - Device that terminates an IB link and executes transport-level functions and support the verbs interface
- **Switch**
  - A device that routes packets from one link to another of the same IB Subnet
- **Router**
  - A device that transports packets between IBA subnets
- **Bridge**
  - InfiniBand to Ethernet



Physical Layer – Link Rate

- **InfiniBand uses serial stream of bits for data transfer**

- **Link width**
  - 1x – One differential pair per Tx/Rx
  - 4x – Four differential pairs per Tx/Rx
  - 12x - Twelve differential pairs per Tx and per Rx



- **Link Speed**
  - Single Data Rate (SDR) - 2.5Gb/s per lane (10Gb/s for 4x)
  - Double Data Rate (DDR) - 5Gb/s per lane (20Gb/s for 4x)
  - Quad Data Rate (QDR) - 10Gb/s per lane (40Gb/s for 4x)
  - Fourteen Data Rate (FDR) - 14Gb/s per lane (56Gb/s for 4x)
  - Enhanced Data rate (EDR) - 25Gb/s per lane (100Gb/s for 4x)

- **Link rate**
  - Multiplication of the link width and link speed
  - Most common shipping today is 4x ports

Physical Layer – Cables

- **Media types**
  - PCB: several inches
  - Passive copper: 20m SDR, 10m DDR, 7m QDR
  - Fiber: 300m SDR, 150m DDR, 100/300m QDR

4X QSFP Copper

- **Link encoding**
  - SDR, DDR, QDR: 8 to 10 bit encoding
  - FDR, EDR: 64 to 66 bit encoding

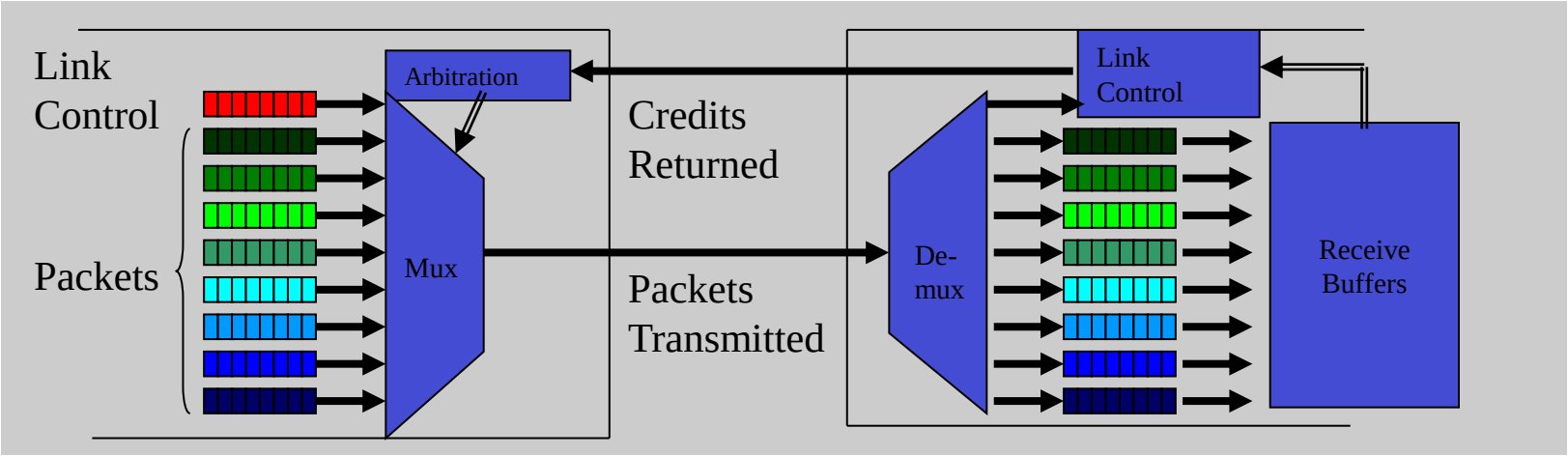
4x QSFP Fiber

- **Industry standard components**
  - Copper cables / Connectors
  - Optical cables
  - Backplane connectors

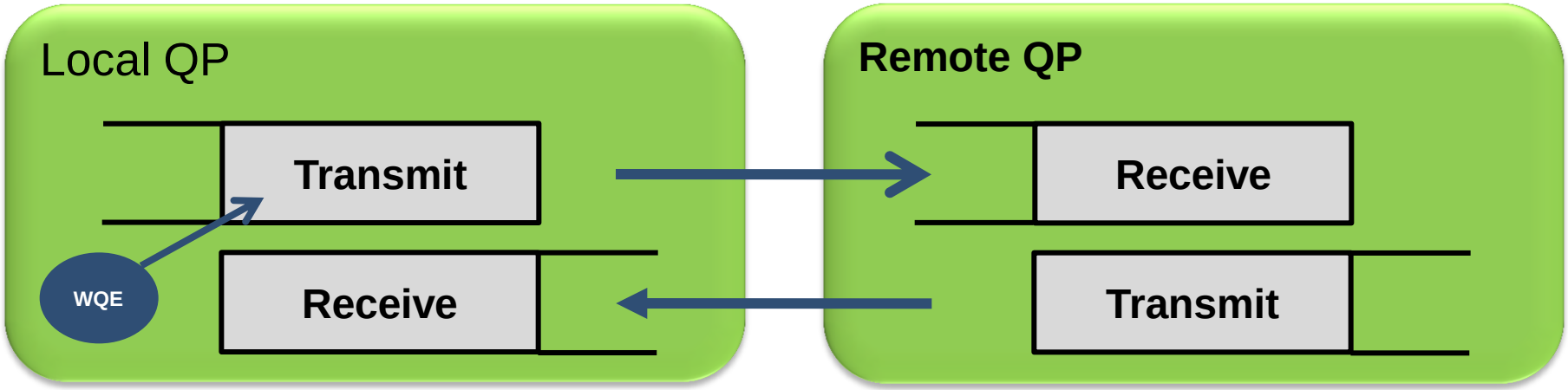
FR4 PCB

Link Layer – Flow Control

- **Credit-based link-level flow control**
  - Link Flow control assures no packet loss within fabric even in the presence of congestion
  - Link Receivers grant packet receive buffer space credits per Virtual Lane
  - Flow control credits are issued in 64 byte units
- **Separate flow control per Virtual Lanes provides:**
  - Alleviation of head-of-line blocking
  - Virtual Fabrics – Congestion and latency on one VL does not impact traffic with guaranteed QOS on another VL even though they share the same physical link



Transport Layer – Using 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 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, message length

- **RDMA Write**
  - Requester HCA sends data to be written into the Responder HCA's system memory
  - Requires remote memory access rights, memory start address, message length

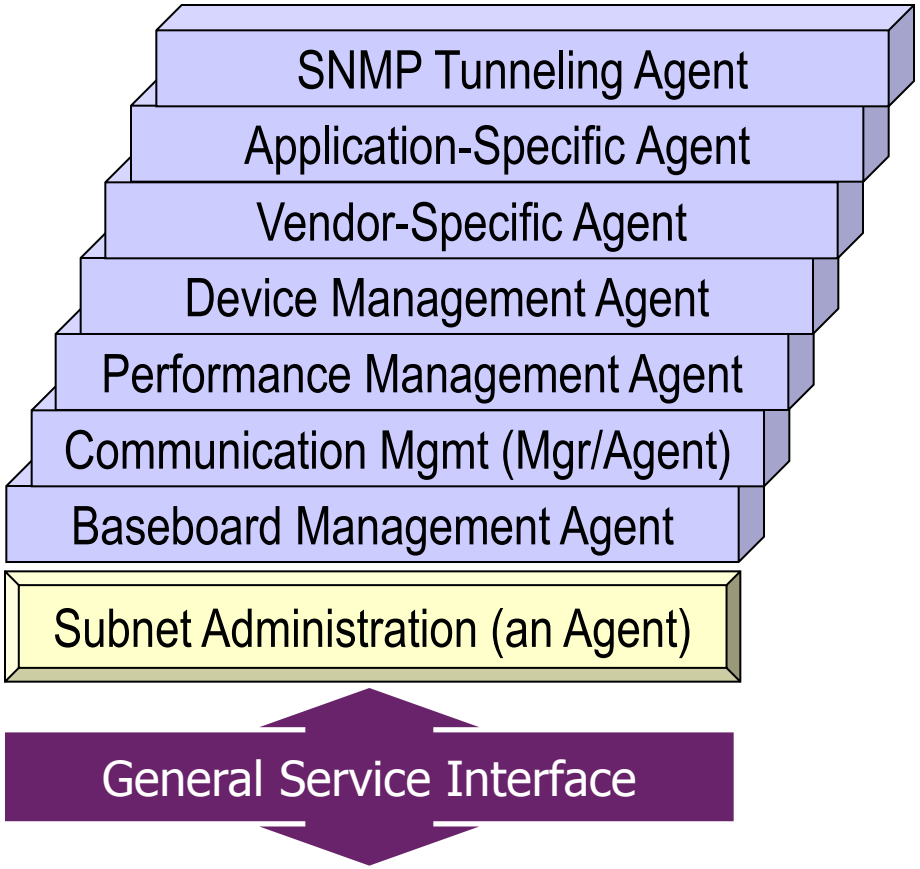
Management Model

- **IBA management defines a common management infrastructure**

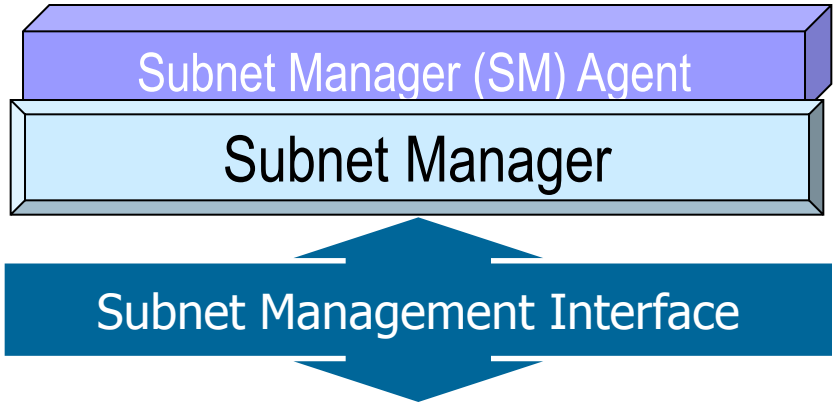
- **Subnet Management**
  - Provides methods for a subnet manager to discover and configure IBA devices
  - Manage the fabric

- **General management services**
  - Subnet administration - provides nodes with information gathered by the SM
  - Provides a registrar for nodes to register general services they provide
  - Communication establishment and connection management between end nodes
  - Performance management
    - Monitors and reports well-defined performance counters
  - And more…

Management Model



QP1 (virtualized per port)
Uses any VL except 15
MADs called GMPs - LID-Routed
Subject to Flow Control



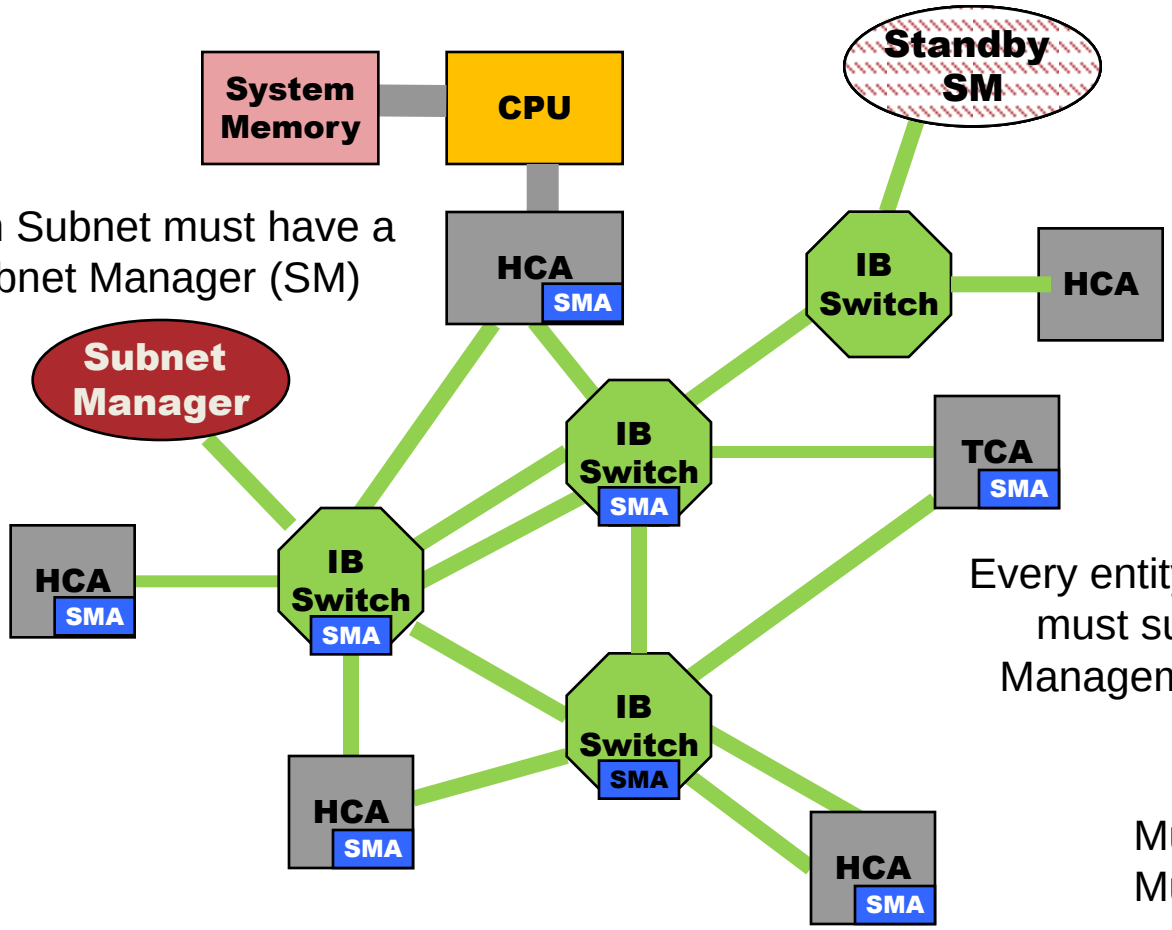
QP0 (virtualized per port)
Always uses VL15
MADs called SMPs – LID or Direct-Routed
No Flow Control

Subnet Management

Topology Discovery
Fabric Maintenance

Initialization uses
Directed Route packets:

Each Subnet must have a Subnet Manager (SM)

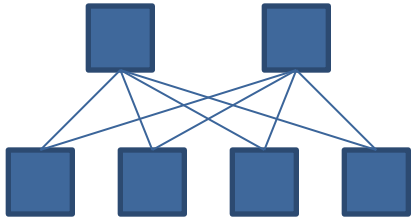


Every entity (CA, SW, Router) must support a Subnet Management Agent (SMA)

Multipathing: LMC Supports Multiple LIDS

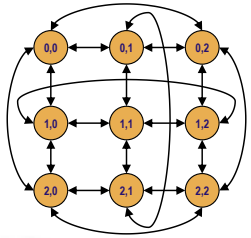
Cluster Topologies

- **Topologies that are mainly in use for large clusters**
  - Fat-Tree
  - 3D Torus
  - Mash



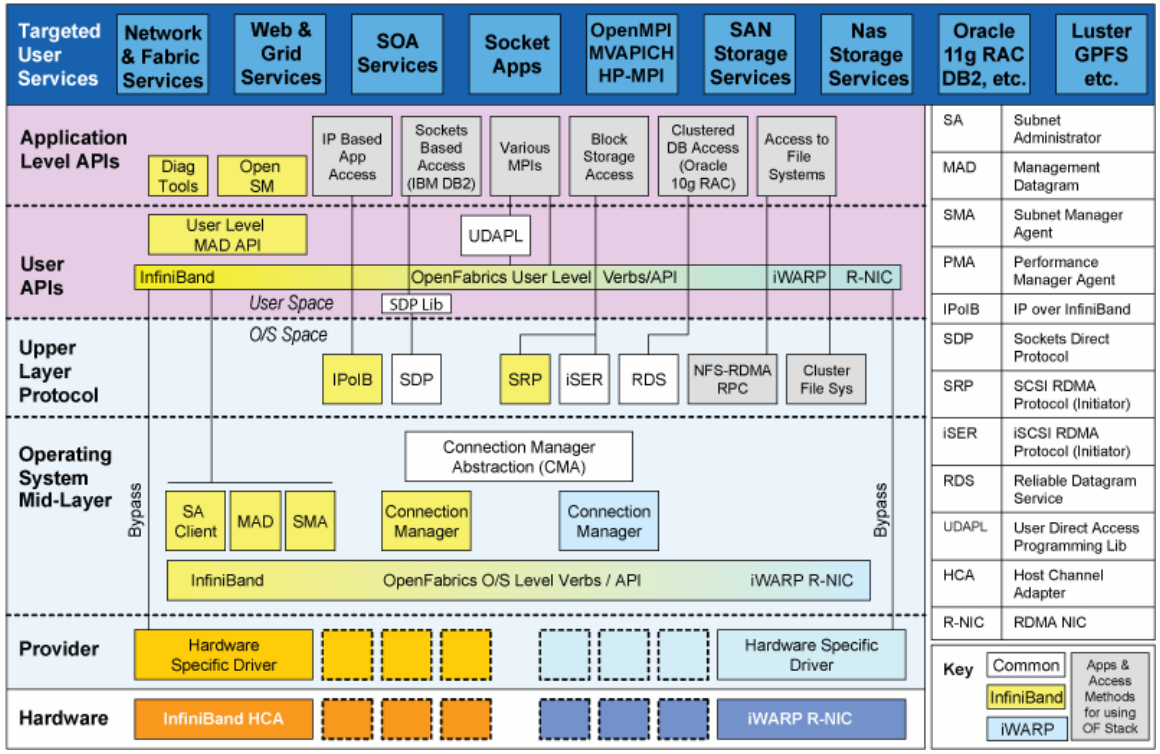
- **Fat-tree (also known as CBB)**
  - Flat network, can be set as oversubscribed network or not
    - In other words, blocking or non blocking
  - Typically the lowest latency network

- **3D Torus**
  - An oversubscribed network, easier to scale
  - Fit more applications with locality



Open Fabrics Linux/Windows Software Stack

- **Open Fabrics is an open-source software development organization**
- **Open Fabrics develops software stack for InfiniBand**
  - Linux and Windows
- **Contains low level drivers, core, Upper Layer Protocols (ULPs), Tools and documents**
- **Available on OpenFabrics.org web site**



Software Stack – Upper Layer protocols Examples

- IPoIB – IP over IB (TCP/UDP over InfiniBand)
- EoIB – Ethernet over IB
- RDS – Reliable Datagram Sockets
- MPI – Message Passing Interface
- iSER – iSCSI for InfiniBand
- SRP – SCSI RDMA Protocol
- uDAPL – User Direct Access Programming Library
- NetworkDirect (Windows only)

Thank You

www.hpcadvisorycouncil.com

info@hpcadvisorycouncil.com