InfiniBand Architecture Overview
Back To Basic

Mellanox® TECHNOLOGIES

CONFIDENTIAL

InfiniBand Technical Overview

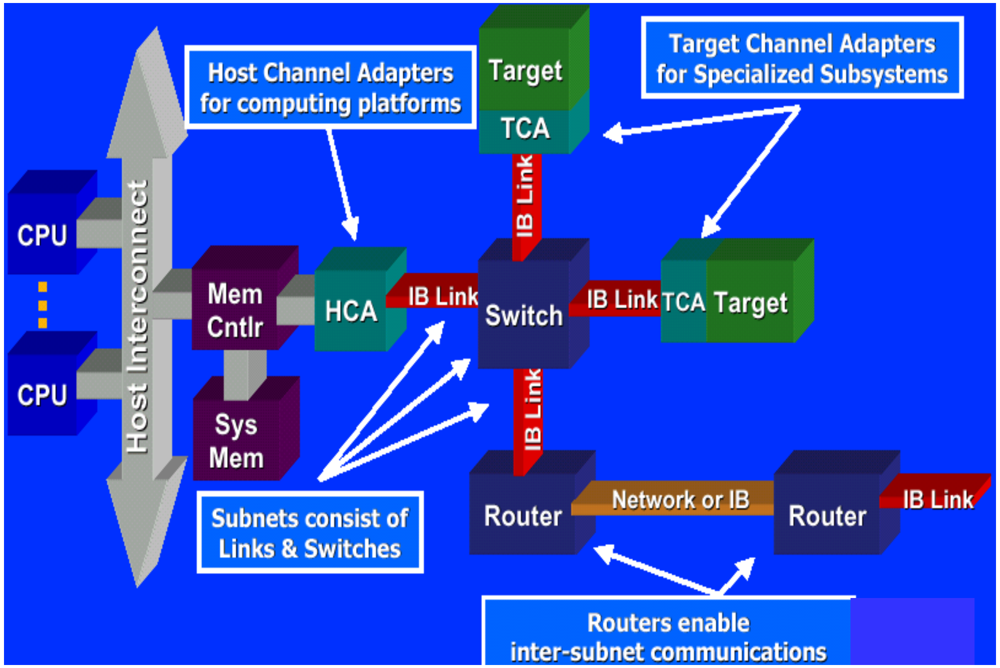
- What is InfiniBand?
  - InfiniBand is an open standard, interconnect protocol developed by the InfiniBand® Trade Association: http://www.infinibandta.org/home
  - First InfiniBand specification was released in 2000
- What does the specification includes?
  - The specification is very comprehensive
  - From physical to applications
- InfiniBand SW is developed under OpenFabrics Open source Alliance
  - http://www.openfabrics.org/index.html

Infiniband Feature Highlights

- Serial High Bandwidth Links
  - 10Gb/s to 40Gb/s HCA links
  - Up to 120Gb/s switch-switch
- Ultra low latency
  - Under 1 us
- Reliable, lossless, self-managing fabric
  - Link level flow control
  - Congestion control
- Full CPU Offload
  - Hardware Based Transport Protocol
  - Reliable Transport
  - Kernel Bypass
- Memory exposed to remote node
  - RDMA-read and RDMA-write
- Quality Of Service
  - I/O channels at the adapter level
  - Virtual Lanes at the link level
- Scalability/flexibility
  - Up to 48K nodes in subnet, up to 2^{128} in network

InfiniBand Components

- **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 (coming soon...)**
  - A device that transports packets between IBA subnets

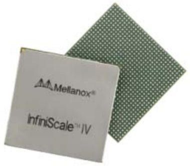


Host Channel Adapters for computing platforms

Target Channel Adapters for Specialized Subsystems

Subnets consist of Links & Switches

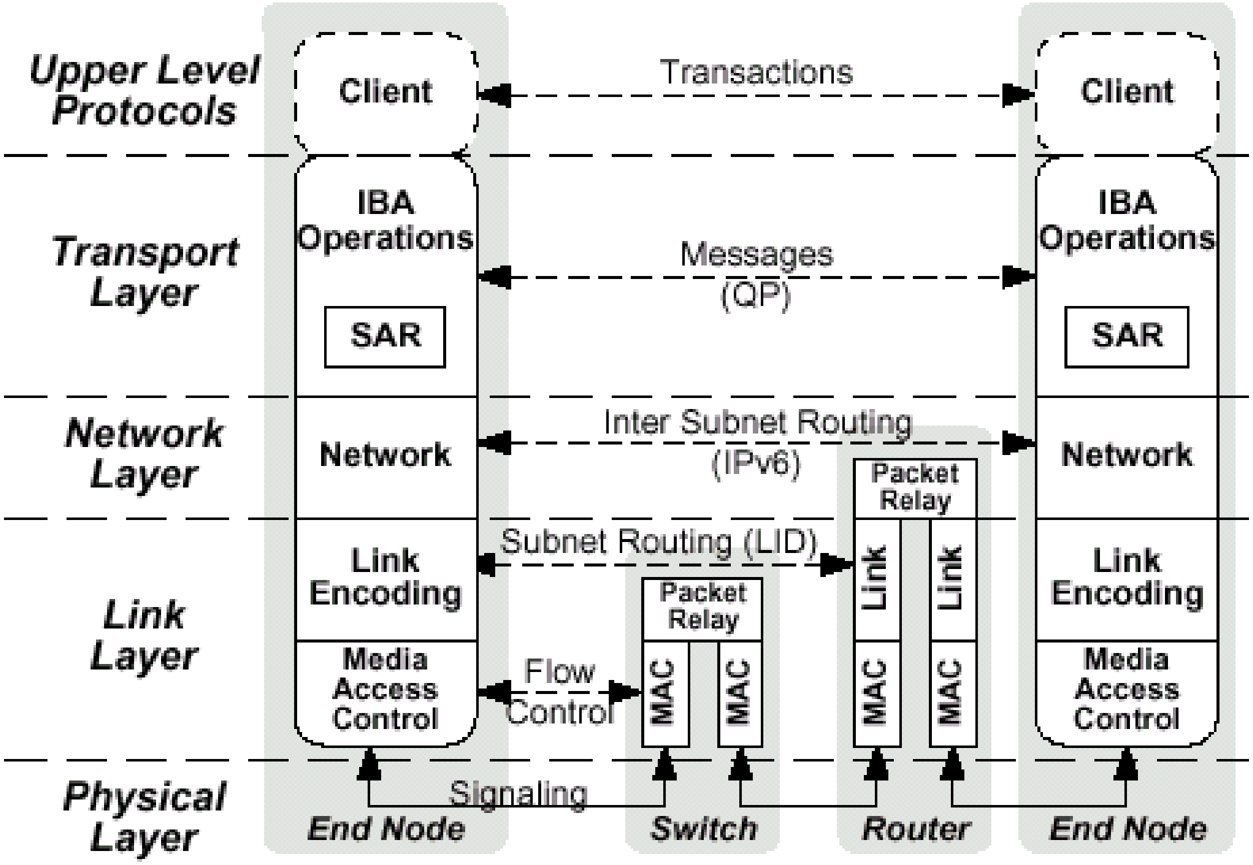
Routers enable inter-subnet communications



IB Architecture Layers

- Physical
  - Signal levels and Frequency; Media; Connectors
- Link
  - Symbols and framing; Flow control (credit-based); How packets are routed from Source to Destination
- Network:
  - How packets are routed between subnets
- Transport:
  - Delivers packets to the appropriate Queue Pair; Message Assembly/De-assembly, access rights, etc.
- Software Transport Verbs and Upper Layer Protocols
  - Interface between application programs and hardware.
  - Allows support of legacy protocols such as TCP/IP
  - Defines methodology for management functions

InfiniBand Layered Architecture



Physical Layer - Responsibilities

- The physical layer specifies how bits are placed on the wire to form symbols and defines the symbols used for framing (i.e., start of packet & end of packet), data symbols, and fill between packets (Idles). It specifies the signaling protocol as to what constitutes a validly formed packet

- InfiniBand is a lossless fabric. Maximum Bit Error Rate (BER) allowed by the IB spec is 10e-12. The physical layer should guaranty affective signaling to meet this BER requiermnet

Physical Layer – Link Rate

- InfiniBand uses serial stream of bits to transfer data
- Link width
  - 1x – One differential pair per Tx and per Rx
  - 4x – Four differential pairs per Tx and per Rx
  - 12x - Twelve differential pairs per Tx and per Rx
- Link Speed
  - Single Dada Rate (SDR) – 2.5 GHz signaling (2.5Gb/s for 1x)
  - Doable Data Rate (DDR) – 5 GHz signaling (5Gb/s for 1x)
  - Quad Data rate (QDR) - 10 GHz signaling (10Gb/s for 1x)
- Link rate
  - Multiplication of the link width and link speed
  - Most common 4x QDR (40Gb/s)

Physical Layer Cont'

- Media types
  - PCB: several inches
  - Copper: 20m SDR, 10m DDR, 7m QDR
  - Fiber: 300m SDR, 150m DDR, 100/300m QDR
  - CAT6 Twisted Pair in future.
- 8 to 10 bit encoding
- Industry standard components
  - Copper cables / Connectors
  - Optical cables
  - Backplane connectors

4X QSFP

4x QSFP Fiber

12X Cable

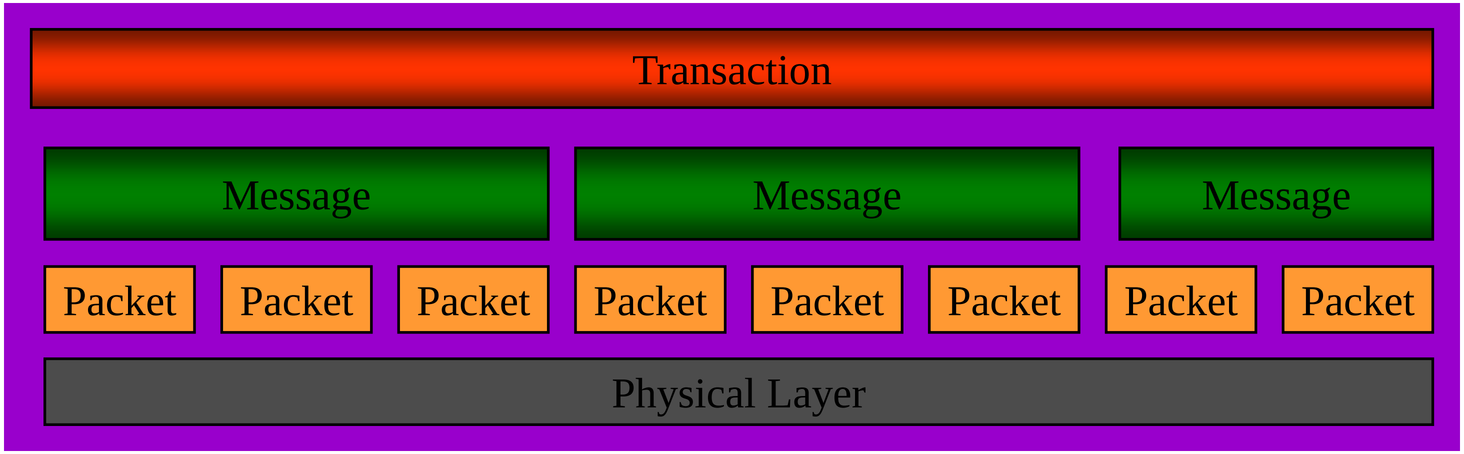
4x CX4 Fiber

4X CX4

FR4 PCB

Link Layer - Responsibilities

- The link layer describes the packet format and protocols for packet operation, e.g. flow control and how packets are routed within a subnet between the source and destination



Link Layer: Packets

- Packets are routable end-to-end fabric unit of transfer
  - Link management packets: train and maintain link operation
  - Data packets
    - Send
    - Read
    - Write
    - Acks

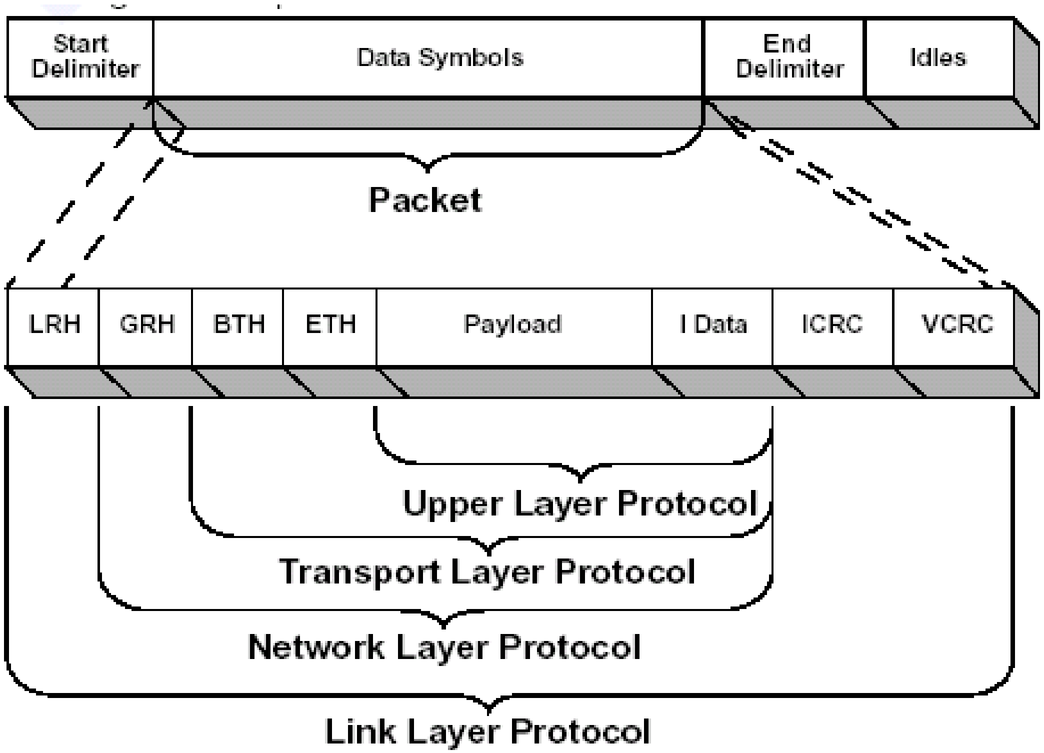


Figure 27 IBA Data Packet Format

Link Layer: Payload Size

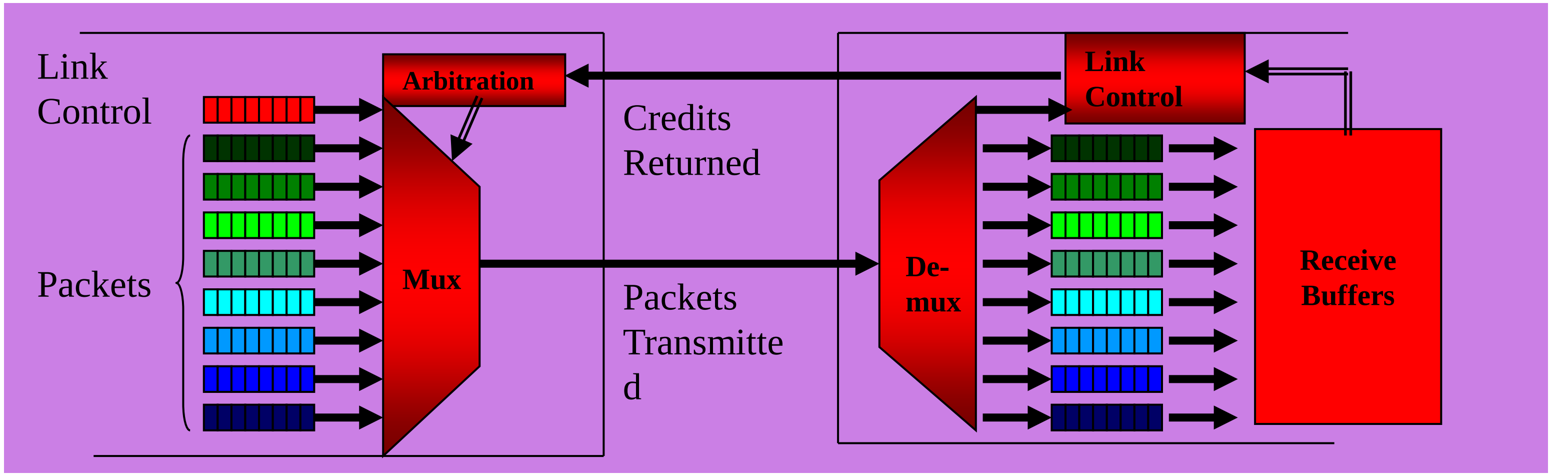
- Maximum Transfer Unit (MTU)
  - MTU allowed from 256 Bytes to 4K Bytes (Message sizes much larger).
  - Only packets smaller than or equal to the MTU are transmitted
  - Large MTU is more efficient (less overhead)
  - Small MTU gives less jitter
  - Small MTU preferable since segmentation/reassembly performed by hardware in the HCA.
  - Routing between end nodes utilizes the smallest MTU of any link in the path (Path MTU)

Link Layer: Virtual Lanes (Quality of Service)

- 16 Service Levels (SLs)
  - A field in the Local Routing Header (LRH) of an InfiniBand packet
  - Defines the requested QoS
- Virtual Lanes (VLs)
  - A mechanism for creating multiple channels within a single physical link.
  - Each VL:
    - Is associated with a set of Tx/Rx buffers in a port
    - Has separate flow-control
  - A configurable Arbiter control the Tx priority of each VL
  - Each SL is mapped to a VL
  - IB Spec allows a total of 16 VLs (15 for Data & 1 for Management)
    - Minimum of 1 Data and 1 Management required on all links
    - Switch ports and HCAs may each support a different number of VLs
  - VL 15 is a management VL and is not a subject for flow control

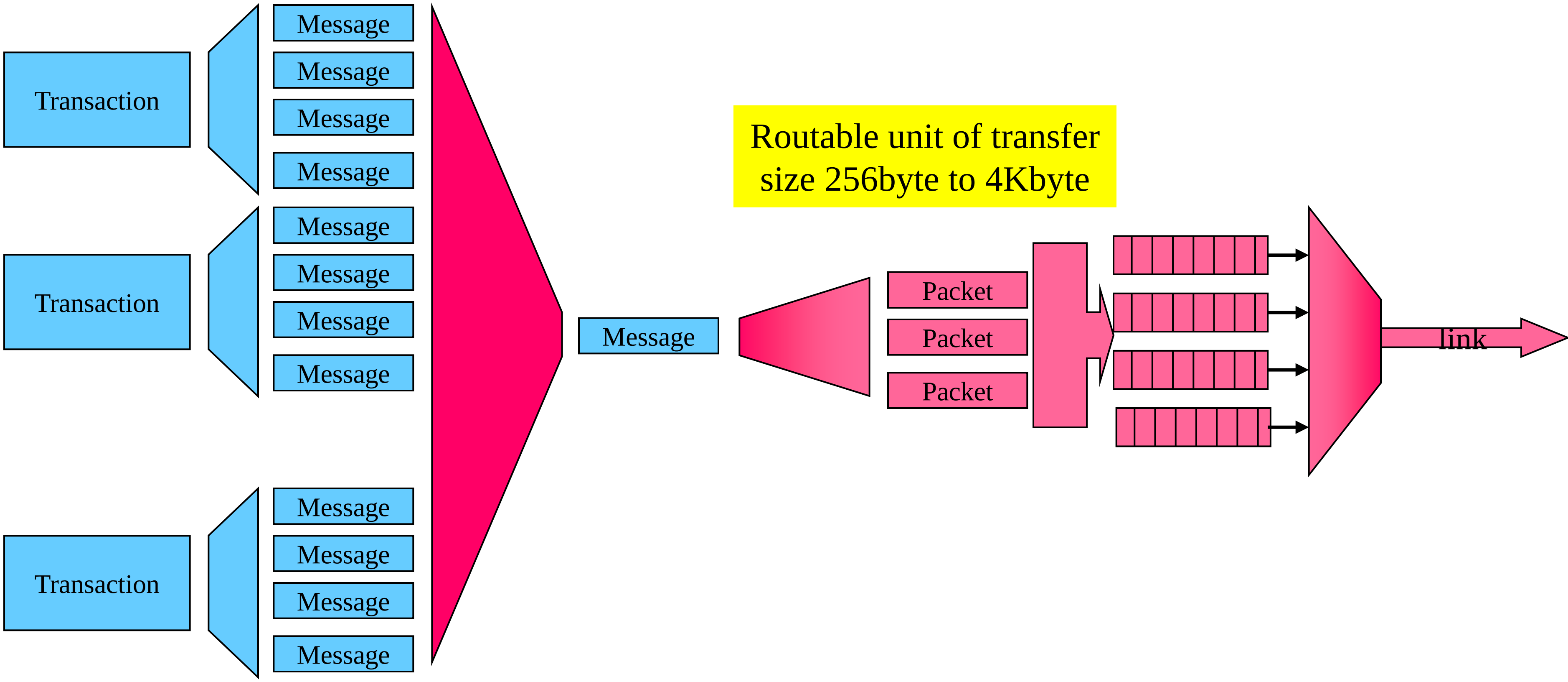
Link Layer: Flow Control

- **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



Link Layer: Example

Message size – up to 2Gbyte



Routable unit of transfer size 256byte to 4Kbyte

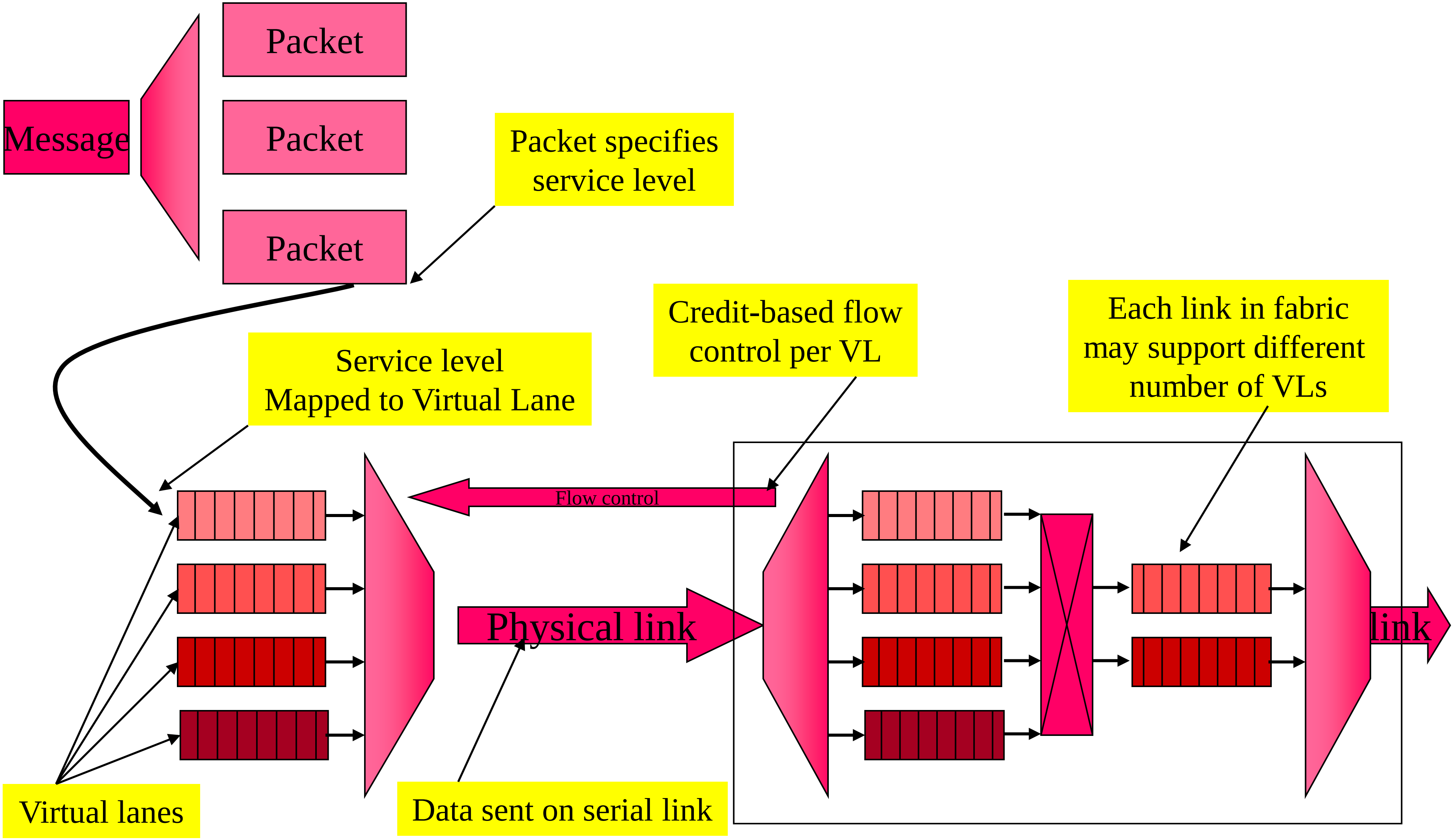
Application accesses HW to post message request

HW schedules execution

HW dis-assembles message to routable units of transfer

HW sends packets on serial link

Link Layer: Example



Packet specifies service level

Service level Mapped to Virtual Lane

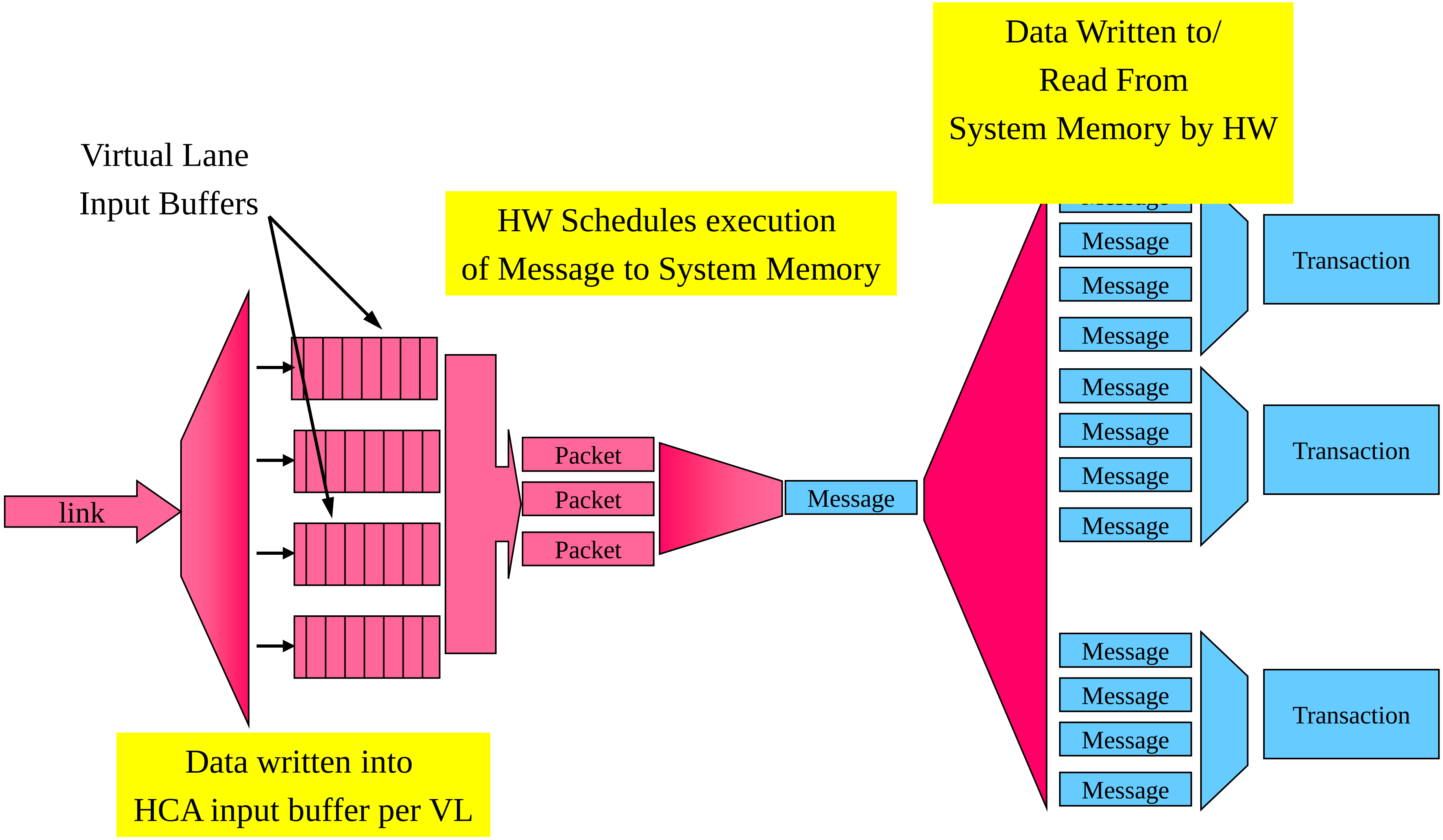
Credit-based flow control per VL

Each link in fabric may support different number of VLs

Virtual lanes

Data sent on serial link

Link Layer: Example



Data Written to/ Read From System Memory by HW

Virtual Lane Input Buffers

HW Schedules execution of Message to System Memory

Data written into HCA input buffer per VL

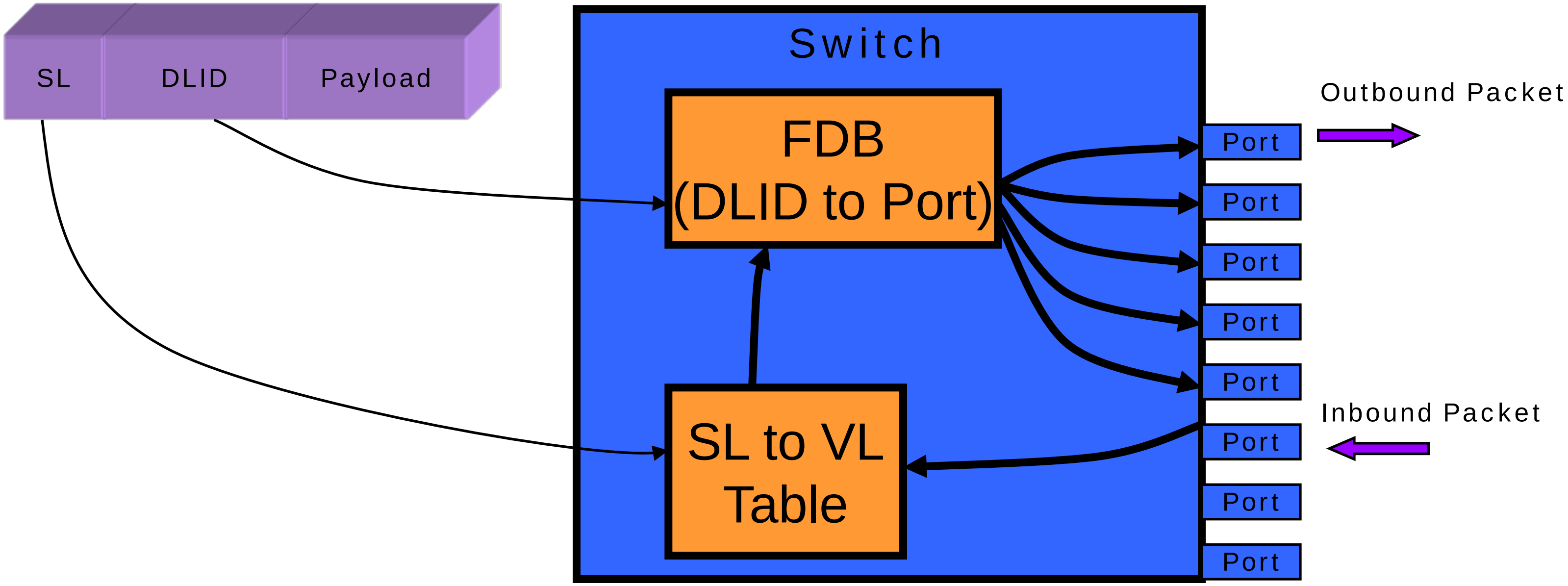
Link Layer: Addressing

- Local ID (LID)
  - 16 bit field in the Local Routing Header (LRH) of all IB packets
  - Used to rout packet in an InfiniBand subnet
  - Each subnet may contain up to:
    - 48K unicast addresses
    - 16K multicast addresses
- Assigned by Subnet Manager at initialization and topology changes

Layer 2 Forwarding

- **Switches use FDB (Forwarding Database)**
  - Based on DLID and SL a packet is sent to the correct output port.

Multicast Destinations supported!!



Network Layer

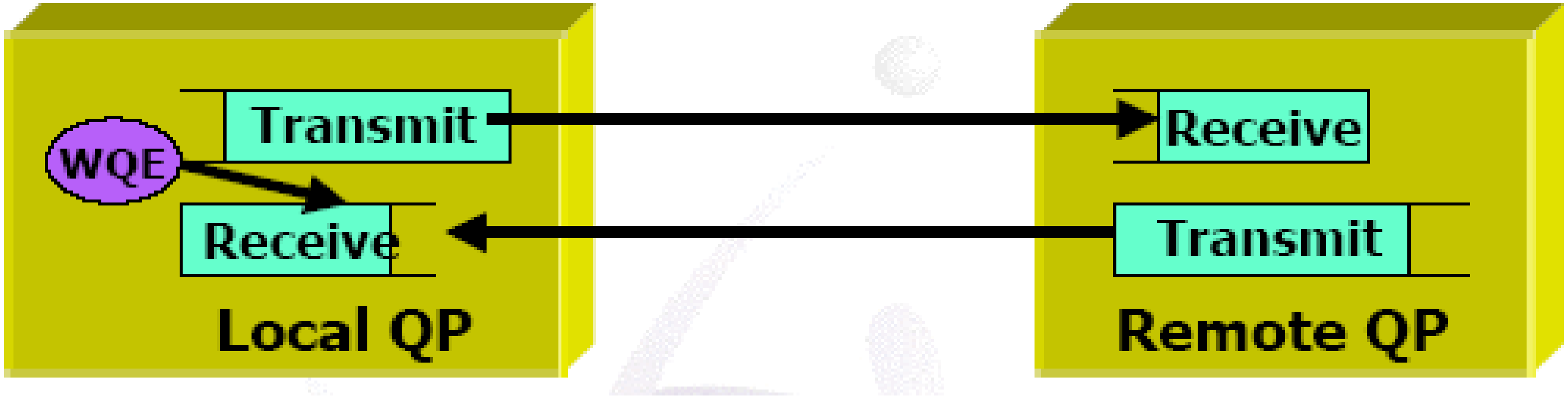
- Responsibility
  - The network layer describes the protocol for routing a packet between subnets
- Globally Unique ID (GUID)
  - A 64 bit field in the Global Routing Header (GRH) used to route packets between different IB subnets
  - Every node must have a GUID
  - IPv6 type header

Transport Layer – Responsibilities

- The network and link protocols deliver a packet to the desired destination. The transport portion of the packet delivers the packet to the proper QP and instructs the QP how to process the packet's data.

- The transport layer is responsible for segmenting an operation into multiple packets when the message's data payload is greater than the *maximum transfer unit (MTU) of the path. The QP on the receiving end* reassembles the data into the specified data buffer in its memory

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)

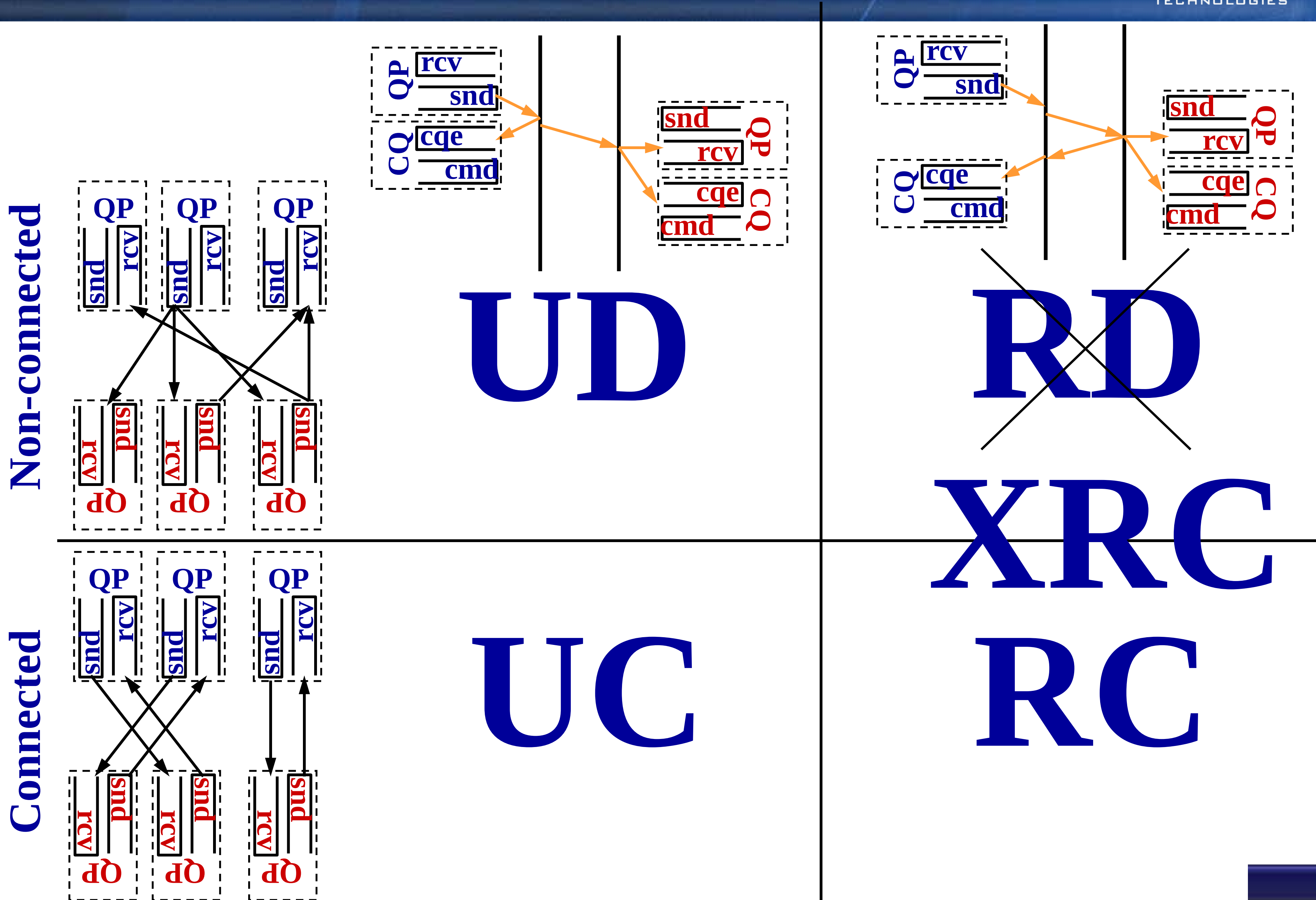
Transport Layer: Work Request Element

- Data transfer
  - Send work request
    - Local gather – remote write
    - Remote memory read
    - Atomic remote operation
  - Receive work request
    - Scatter received data to local buffer(s)
- Memory management operations
  - Bind memory window
    - Open part of local memory for remote access
  - Send & remote invalidate
    - Close remote window after operations' completion
- Control operations
  - Memory registration/mapping
  - Open/close connection (QP)

Transport Layer: Types Transfer Operations

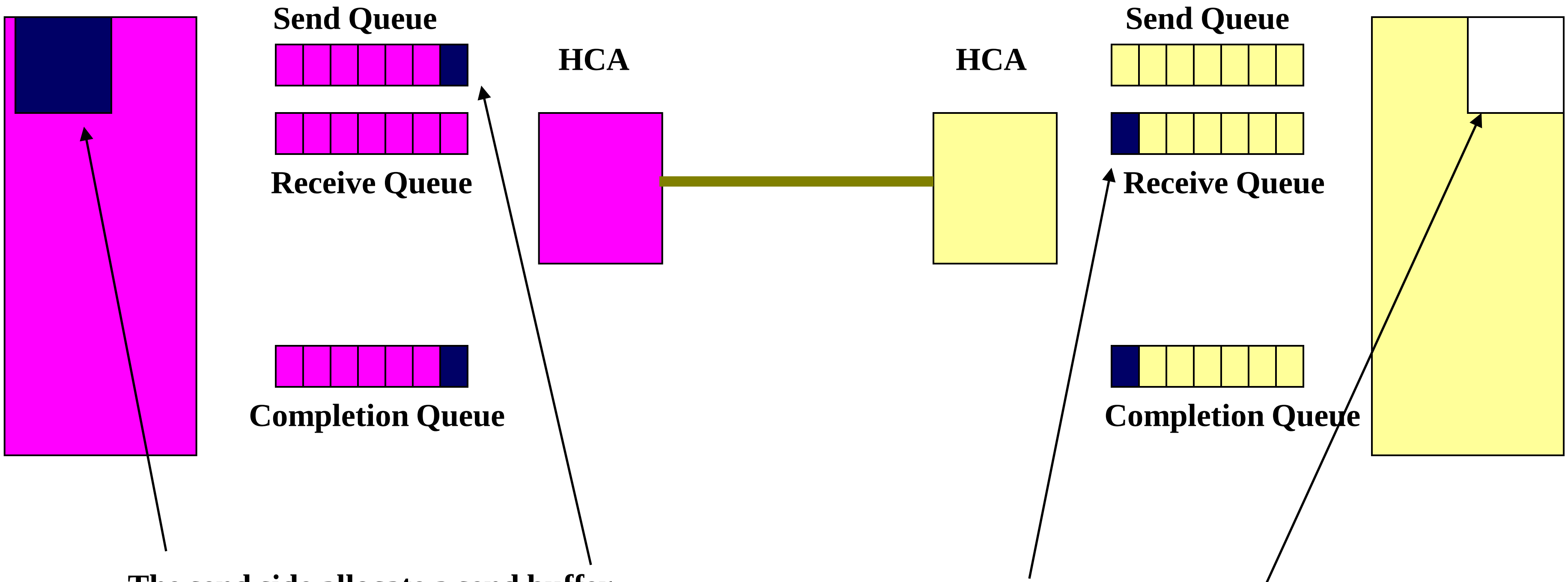
- SEND
  - Read message from HCA local system memory
  - Transfers data to Responder HCA Receive Queue logic
  - Does not specify where the data will be written in remote memory
  - Immediate Data option available
- RDMA Read
  - Responder HCA reads its local memory and returns it to the Requesting HCA
  - Requires remote memory access rights, memory start address, and message length
- RDMA Write
  - Requester HCA sends data to be written into the Responder HCA’s system memory
  - Requires remote memory access rights, memory start address, and message length

Transport Services



	Unreliable	Reliable
Non-connected	UD	~~RD~~ XRC
Connected	UC	RC

Transport Layer: Send operation example



HCA then consume the WQE,
read the buffer and send to remote side
send completion is generated

When the packet arrives to the HCA
It consumes a receive WQE, place
the buffer in the appropriate location
and generate a completion

Host A RAM

Host B RAM

The send side allocate a send buffer
register it with the HCA, place a send WQE

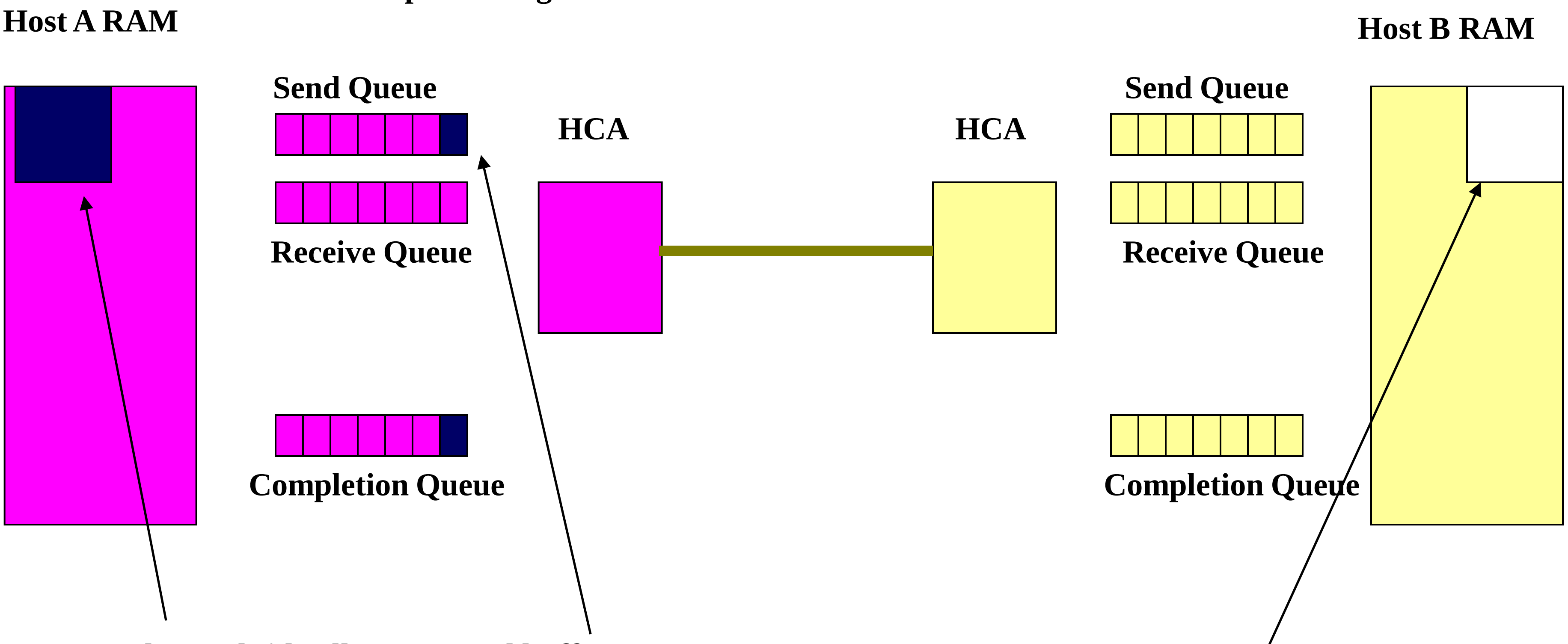
and ring a doorbell

Application allocate receive buffer
and place a receive WQE

Transport Layer: RDMA Write Example

HCA then consume the WQE,
read the buffer and send to remote side
send completion is generated

When the packet arrives to the HCA
It checks the address and memory
keys and write to memory directly



The send side allocate a send buffer
register it with the HCA, place a send WQE
with the remote side's virtual address

and ring a doorbell

Application allocate receive buffer
and pass address and keys to
remote side

Transport Layer: Retransmissions

- For reliable transport services (RC, XRC) QPs maintain the flow of packets and retransmit in case a packet was dropped
- Each packet has a Packet Serial Number (PSN) that is used by the receiver identify lost packets
- The receiver will send ACKs if packets arrive in order and NACKs otherwise
- The send QP maintain a timer to catch cases where packets did not arrive to the receive QP or ACK was lost
- Retransmission is considered a “bad flow” which reduce performance or may break a connection

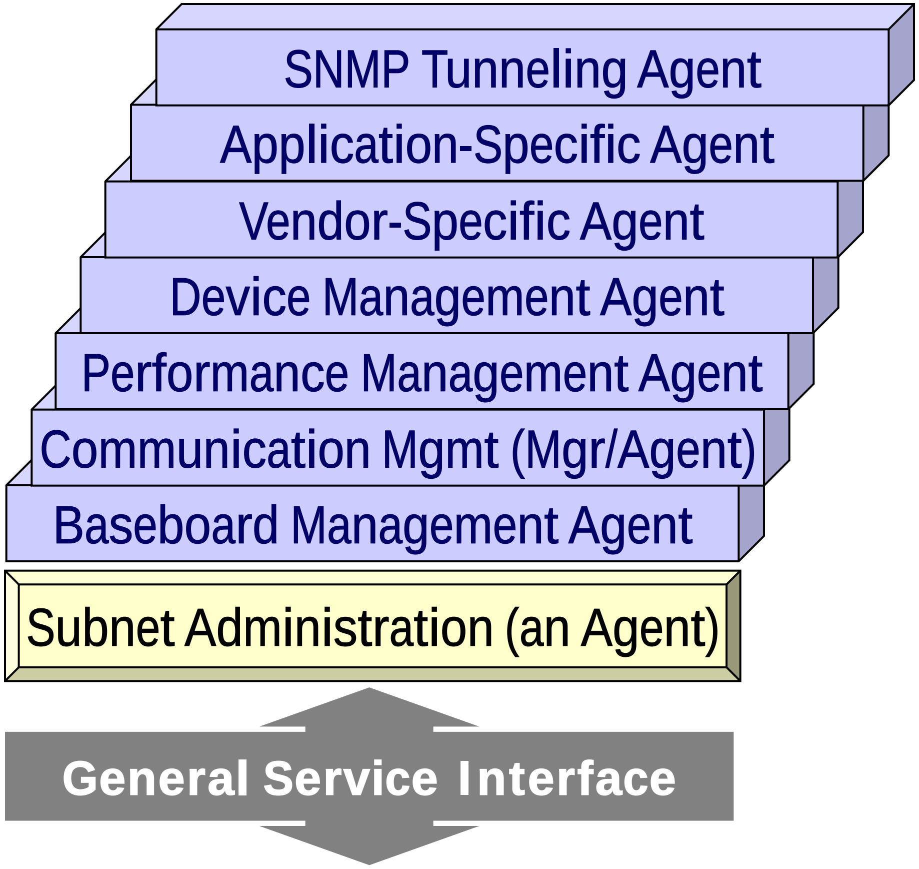
Verbs

- Verbs are the SW interface to the HCA and the IB fabric
- Verbs are not API but rather allow flexibility in the API implementation while defining the framework
- Some verbs for example
  - Open/Query/Close HCA
  - Create Queue Pair
  - Query Completion Queue
  - Post send Request
  - Post Receive Request
- Upper Layer Protocols (ULPs) are application writing over the verbs interface that bridge between standard interfaces like TCP/IP to IB to allow running legacy application intact

Management Model

- IBA management defines a common management infrastructure for
  - Subnet Management - provides methods for a subnet manager to discover and configure IBA devices and manage the fabric
  - General management services
    - Subnet administration - provides nodes with information gathered by the SM and provides a registrar for nodes to register general services they provide
    - Communication establishment & connection management between end nodes
    - Performance management - monitors and reports well-defined performance counters
    - And more…

Management Model



SNMP Tunneling Agent

Application-Specific Agent

Vendor-Specific Agent

Device Management Agent

Performance Management Agent

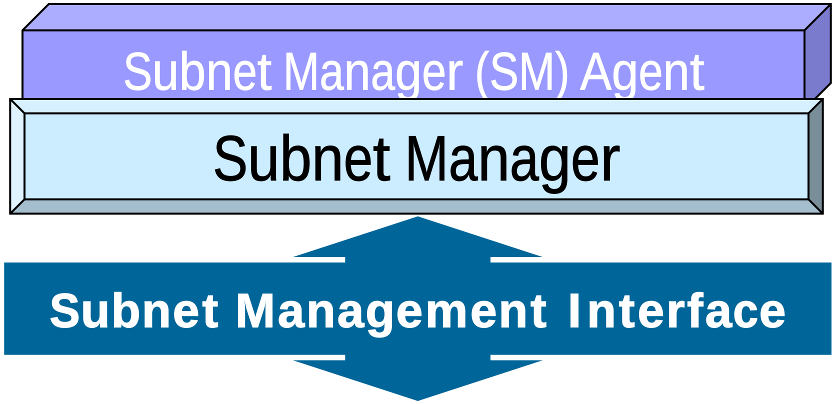
Communication Mgmt (Mgr/Agent)

Baseboard Management Agent

Subnet Administration (an Agent)

General Service Interface

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



Subnet Manager (SM) Agent

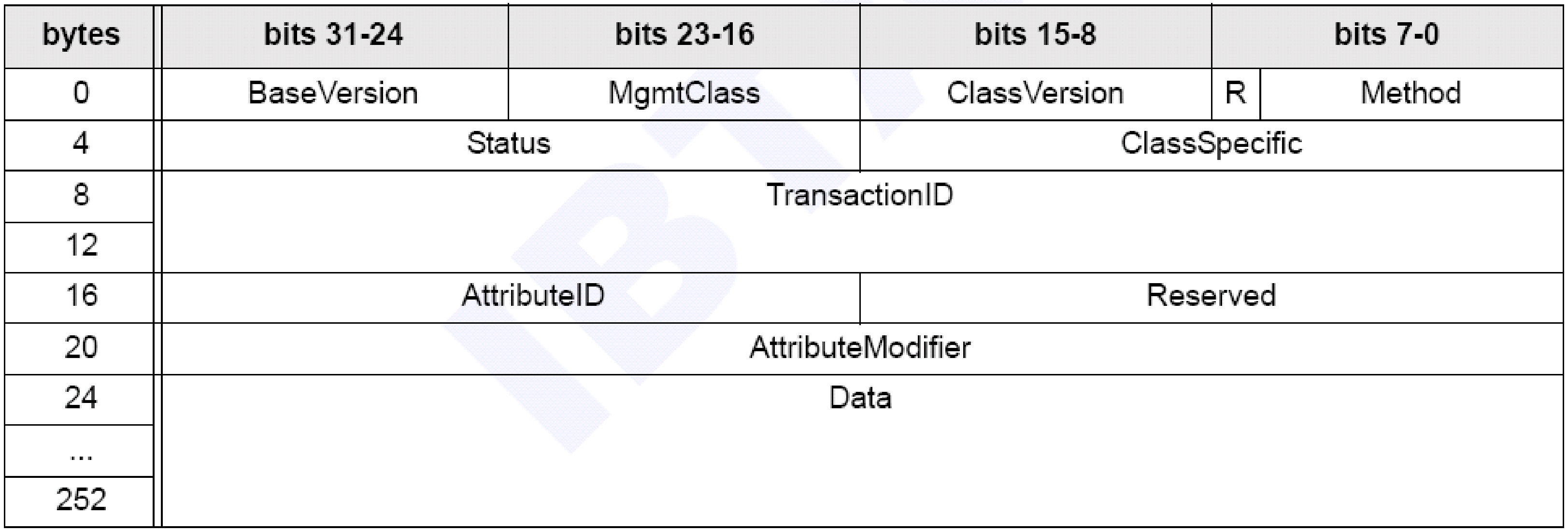
Subnet Manager

Subnet Management Interface

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

Management Model – Packets

- Management is done using Management Datagram (MAD) packets
  - SMP – Subnet Manager MADs
  - GMP – General Management MADs



bytes	bits 31-24	bits 23-16	bits 15-8	bits 7-0
0	BaseVersion	MgmtClass	ClassVersion	R Method
4	Status		ClassSpecific	
8	TransactionID			
12				
16	AttributeID		Reserved	
20	AttributeModifier			
24	Data			
...				
252				

Figure 145 MAD Base Format

Subnet Management

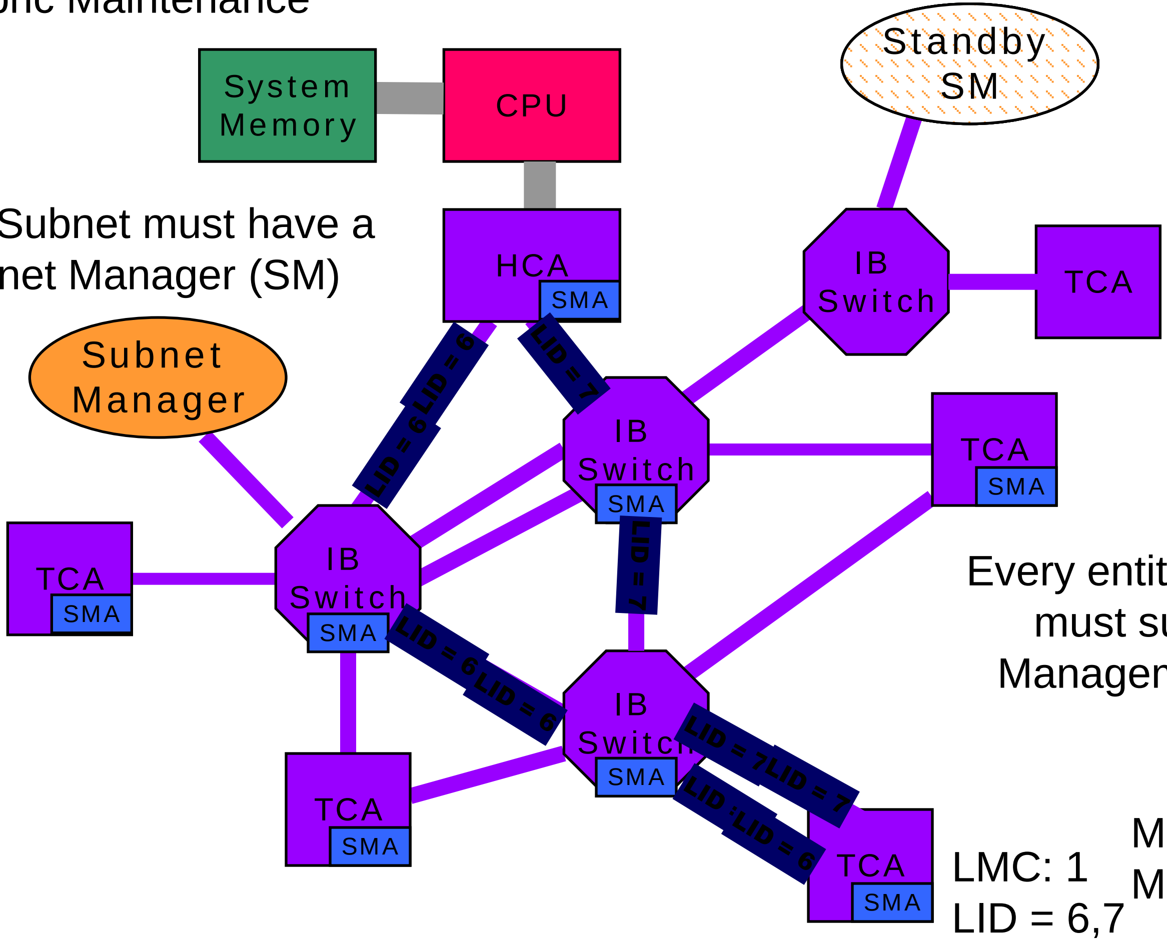
Topology Discovery
FDB Initialization
Fabric Maintenance

Initialization uses Directed Route MADs:

LID Route	Directed Route Vector	LID Route

MADs use unreliable datagrams

Each Subnet must have a Subnet Manager (SM)



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

LMC: 1
LID = 6,7

Multipathing: LMC Supports Multiple LIDS

Other management entities

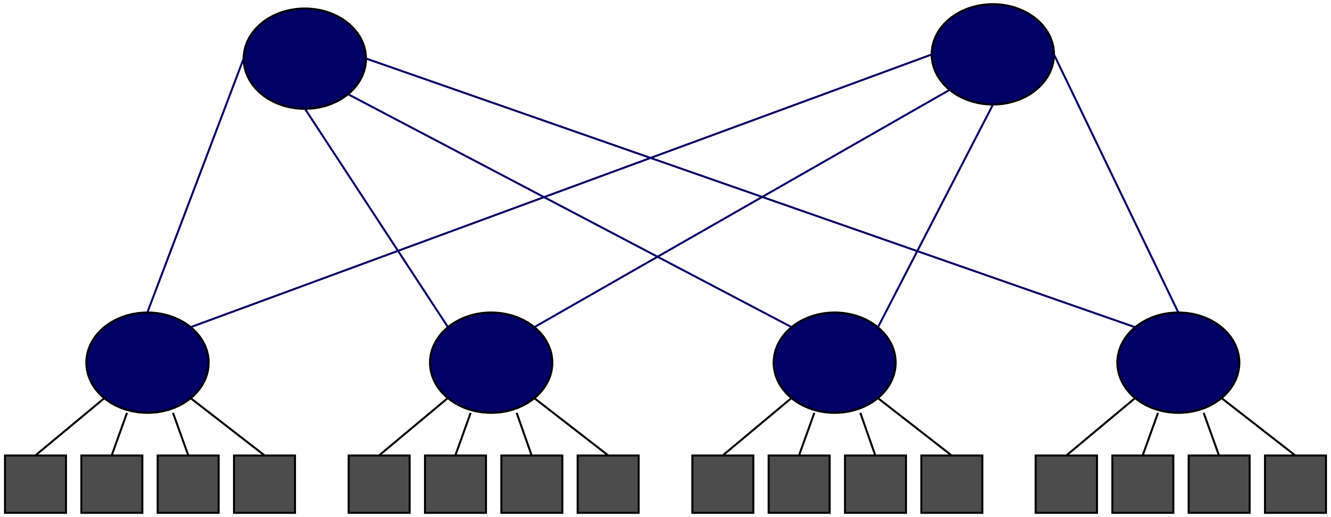
- Connection Manager (CM)
  - Establishes connection between end-nodes
- Performance Management (PM)
  - Performance Counters
    - Saturating counters
  - Sampling Mechanism
    - Counter works during programmed time period
- Baseboard Management (BSM)
  - Access Vital Product Data (VPD)
  - Bridge to/from IBML devices
    - Power Management
    - Hot plug in and removal of modules
    - Monitoring of environmental parameters

Topologies

- There are several common topologies for an IB fabric
  - Fat Tree – Most popular. A tree where the HCA are the leaf of the tree and that allow full bisectional Bandwidth (BW) between pair of nodes
  - Mash – each node is connected to 4 other nodes: positive and negative X and Y axis
  - 3D mash – Each node is connected to 6 other nodes: positive and negative X, Y and Z axis
  - 2D/3D torus – The ends of the 2D/3D mashes are connected

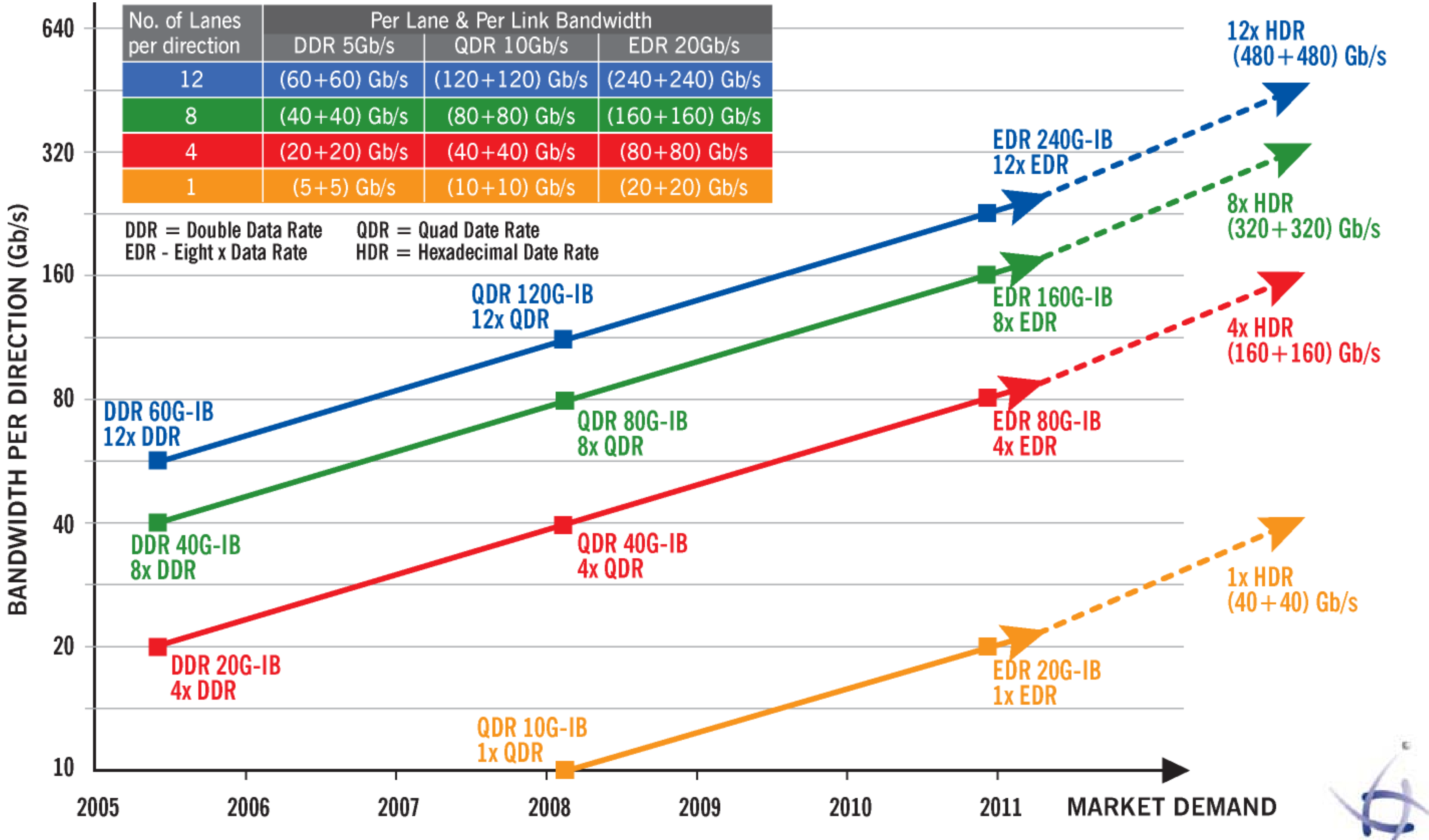
Topologies – Fat Tree Example

Full Fat Tree / Full CBB



Half Fat Tree / Half CBB

InfiniBand Link Speed Roadmap



No. of Lanes per direction	Per Lane & Per Link Bandwidth		
	DDR 5Gb/s	QDR 10Gb/s	EDR 20Gb/s
12	(60+60) Gb/s	(120+120) Gb/s	(240+240) Gb/s
8	(40+40) Gb/s	(80+80) Gb/s	(160+160) Gb/s
4	(20+20) Gb/s	(40+40) Gb/s	(80+80) Gb/s
1	(5+5) Gb/s	(10+10) Gb/s	(20+20) Gb/s

DDR = Double Data Rate QDR = Quad Date Rate
EDR - Eight x Data Rate HDR = Hexadecimal Date Rate

Thank You

www.mellanox.com

Mellanox® TECHNOLOGIES

CONFIDENTIAL